3MSPEECHBENCH: MEASURING BEHAVIORAL CONSISTENCY OF MULTIMODAL LLMS ACROSS SPEAKERS AND LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing benchmarks for multimodal large language models (MLLMs) with speech input primarily evaluate aggregate task performance, providing limited insight into whether model behavior remains stable when the same content is spoken by different speakers. We present **3MSPEECHBENCH**, a **m**ulti-speaker, **m**ultilingual, **m**ulti-task benchmark that holds linguistic content fixed while varying the speaker. Built from four established corpora, it covers spoken multiple-choice QA, spoken visual QA, article summarization, and dialogue summarization, with 50 speakers per language and up to 10 languages. We evaluate MLLMs using both task performance and within-instance consistency across speakers. Our results show that strong average performance does not necessarily imply stable behavior across speakers, highlighting the need to evaluate speech-facing MLLMs beyond aggregate task scores. A demo with data examples is available at https://anonymous.4open.science/w/3mspeechbench-demo-27A3/.

1 INTRODUCTION

Multimodal large language models (MLLMs) have progressed from text-only systems to end-to-end architectures that directly process spoken audio (Hurst et al., 2024; Google, 2025; Xu et al., 2025). Unlike cascaded pipelines that first transcribe speech (Zhang et al., 2023; Tang et al., 2024), these models encode the audio itself, so their outputs may depend not only on what is said but also on who says it. When speaker identity is irrelevant to the task, the same linguistic content should elicit stable task-relevant behavior across speakers.

Current evaluation rarely measures this stability. Spoken benchmarks typically use only one or a few voices, and even when multiple speakers are included, performance is usually aggregated across instances and speakers. These aggregate scores cannot reveal whether a model gives the same response to the same content when only the speaker changes. Two models with equal accuracy can differ substantially in how often their answers flip across speakers, while two summarizers with similar reference-based scores can state different facts depending on who reads the input.

We introduce **3MSPEECHBENCH**, a benchmark that holds linguistic content fixed while varying the speaker, enabling controlled within-instance comparisons across speakers. Our contributions are threefold:

- We render four established datasets, covering multiple-choice QA, visually grounded QA, article summarization, and dialogue summarization, in up to 10 languages with 50 speaker-cloned voices per language.
- We define within-instance consistency for both closed-form answers and open-ended summaries, extending consistency measurement beyond exact-match outputs.
- Evaluating nine open MLLMs, we show that consistency captures behavioral variation that aggregate performance alone does not reveal. Further analyses show that this instability is not explained by synthesis quality or a few atypical voices, reflects each model's existing response preferences, and persists when evaluated with human speech.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

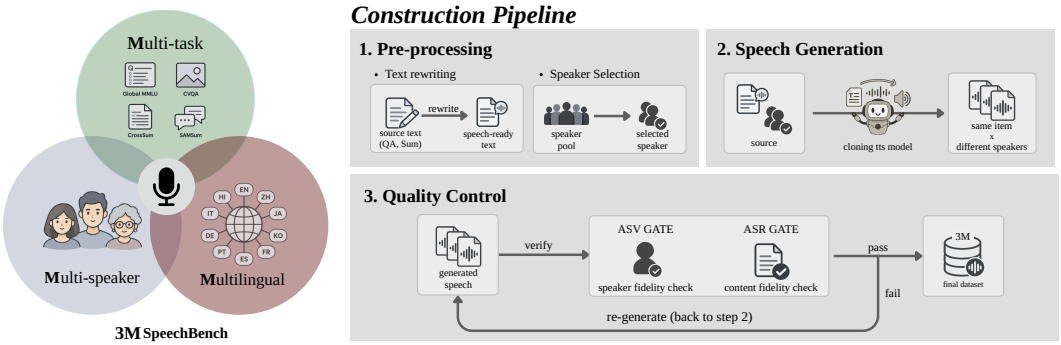

Figure 1: **Overview of 3MSPEECHBENCH.** The benchmark covers 10 languages, with 50 speakers per language, and four tasks. Its construction involves three stages, including text and speaker preprocessing, speech generation, and quality control to ensure speaker and content fidelity.

Task	Input	Output	Languages	Instances	Hours
Spoken Global MMLU	Exam question	Multiple choice	10	19,690	6,077
Spoken CVQA	Question + image	Multiple choice	7	7,118	801
Spoken CrossSum	News article	Summary	8	6,148	3,886
Spoken SAMSum	Two-speaker chat	Summary	1	13,386	1,332
Total				**46,342**	**12,096**

Table 1: **Statistics of 3MSpeechBench.** Input and output formats, language coverage, and overall task scale are summarized by the number of source instances and total duration of generated audio. Per-language statistics are provided in Appendix A.6.

2 DESIGN AND CONSTRUCTION OF 3MSPEECHBENCH

2.1 DESIGN OVERVIEW

Design Principle. The core design principle of 3MSPEECHBENCH is to *hold linguistic content fixed and vary only the speaker*. Each source instance is rendered by a pool of cloned voices, producing semantically identical but acoustically distinct speech. The evaluation framework in §3 thus scores each instance on both *performance* against its reference and *behavioral consistency* across its speakers. Figure 1 illustrates its three defining axes, and Table 1 summarizes its task settings and scale of 46.34k source instances and 12.10k hours of speech across 500 speakers and 10 languages.

Comparison with Prior Benchmarks. Table 2 compares 3MSPEECHBENCH with prior benchmarks that vary speaker characteristics. Prior work is largely English-only and limited to closed-form answers. It typically uses only a few synthetic voices and reports aggregate scores, making speaker effects visible mainly through differences between speaker groups. C3T (Kubis et al., 2025) is closest to ours, as it clones dozens of voices per instance and deems an instance fair only when all speakers receive the same answer, but it remains English-only and restricted to exact-match tasks. To our knowledge, 3MSPEECHBENCH is the first diagnostic benchmark to combine multi-speaker and multilingual coverage with open-ended generation and a graded instance-level measure.

2.2 SETUP AND NOTATION

Each source instance i consists of text x_i and a reference r_i, with an additional image v_i for the visually grounded task. The benchmark holds x_i fixed across speakers. Let $\mathcal{S}$ denote the speaker set for the language of x_i. Rendering x_i with each speaker yields

$$a_{i,s} = \mathrm{TTS}(x_i, s), \qquad s \in \mathcal{S}, \tag{1}$$

Benchmark	Voices per instance	#Lang	Voice source	MCQA	VQA	Open-ended	Consistency
VoiceBench (Chen et al., 2026)	16†	1	Google TTS, CosyVoice clones	✓	✗	✓	✗
VoxEval (Cui et al., 2025)	6	1	OpenAI TTS voices	✓	✗	✗	✗
C3T (Kubis et al., 2025)	43	1	Cloned from real speakers	✓	✗	✗	✓
BiasInEar (Wei et al., 2026)	2	3	Gemini TTS voices	✓	✗	✗	✓
3MSpeechBench (ours)	50	10	Cloned from real speakers	✓	✓	✓	✓

Table 2: **Comparison with prior speaker-varied benchmarks.** Benchmarks are compared by the number of voices per instance, language coverage, voice source, supported task types, and whether output consistency across speakers is measured. †VoiceBench reaches 16 voices per instance only in its speaker-variation analysis on a single subset, not across the full benchmark.

where each $a_{i,s}$ contains the same linguistic content but a different voice. Given each speech input, the same task prompt, and any fixed non-audio input, a model $\mathcal{M}$ produces

$$y_{i,s} = \mathcal{M}(a_{i,s}, v_i), \tag{2}$$

where v_i is omitted for tasks without an image. §2.3 specifies x_i and r_i for each task, §2.4 describes the construction of $\mathcal{S}$ and $a_{i,s}$, and §3 defines the evaluation framework over $\{y_{i,s}\}_{s\in\mathcal{S}}$.

2.3 Tasks and Languages

The four tasks are derived from and named after four source corpora. Their ten languages span four families: English, German, French, Italian, Spanish, Portuguese, and Hindi from Indo-European, Chinese from Sino-Tibetan, Japanese from Japonic, and Korean from Koreanic. Table 1 lists the languages retained for each task, while Appendix A.1 provides further details on the source corpora.

Spoken Global MMLU is derived from the 42-language extension of MMLU (Singh et al., 2025; Hendrycks et al., 2021). We select ten languages and retain its subject and cultural-sensitivity labels. Each x_i is a four-option question presented as a single audio stream, with the correct option as r_i.

Spoken CVQA is derived from the culturally grounded visual benchmark CVQA (Romero et al., 2024). Each x_i is a four-option question paired with its original image v_i, with the correct option as r_i. This setting examines whether visual grounding stabilizes predictions under speaker variation.

Spoken CrossSum is derived from the cross-lingual news corpus CrossSum (Bhattacharjee et al., 2023), which aligns XL-Sum articles (Hasan et al., 2021) across languages. Following Pernes et al. (2024), we regroup its pairwise alignments into multi-way parallel sets across eight languages. Each x_i is an article and r_i its same-language reference summary.

Spoken SAMSum is derived from English messenger-style dialogues (Gliwa et al., 2019). Each x_i is a dialogue and r_i its abstractive summary. We retain dialogues between one male and one female participant and assign each a distinct cloned voice, so s indexes a male–female voice pairing.

2.4 Construction Pipeline

As shown in Figure 1, we prepare source text and speaker pool, synthesize each instance with each cloned voice, and verify linguistic and speaker fidelity. Details are provided in Appendices A.2–A.4.

Text Rewriting. We make source text speech-ready using the rule-based filters in Table 5, which remove instances without a natural spoken form, including mathematical expressions, chemical formulas, code, tables, and other layout-dependent content, URLs, and personal contact details.

Speaker Selection. For each language, we select a gender-balanced set $\mathcal{S}$ of 50 real speakers from YouTube videos released under the CC BY 4.0 license. We retain only clear speech with at most mild background noise and no voice alteration. For each speaker, we extract a reference segment of at least 10 seconds in the target language for voice cloning. We also release each speaker's gender, coarse age group, acoustic scene, and self-stated nationality as metadata. Tables 6 and 7 report the per-language distribution of this metadata across 42 nationalities.

Speech Generation. We synthesize each speech input using the reference-conditioned Chatterbox Multilingual TTS model (Resemble AI, 2025). Each input is generated from the source text and

		Spoken Global MMLU						Spoken CVQA					
Language	**Metric**	**[0, 0.1)**	**[0.1, 0.2)**	**[0.2, 0.3)**	**[0.3, 0.4)**	**[0.4, 0.5)**	**$\geq$ 0.5**	**[0, 0.1)**	**[0.1, 0.2)**	**[0.2, 0.3)**	**[0.3, 0.4)**	**[0.4, 0.5)**	**$\geq$ 0.5**
English	WER	96.5	2.4	0.6	0.3	0.1	0.1	91.2	7.0	1.2	0.4	0.1	0.0
Chinese	CER	90.6	3.7	2.5	1.7	0.9	0.7	95.1	2.7	1.0	0.6	0.2	0.4
Korean	CER	93.4	3.2	1.6	0.9	0.4	0.6	91.0	6.2	1.8	0.7	0.3	0.1
Japanese	CER	91.8	4.8	1.8	0.9	0.3	0.3	90.4	6.7	2.0	0.6	0.4	0.0
Hindi	WER	91.7	4.7	1.6	0.9	0.5	0.6	94.9	3.8	0.7	0.1	0.0	0.5
Portuguese	WER	92.6	3.9	1.9	0.8	0.5	0.3	93.5	5.7	0.6	0.2	0.0	0.0
Spanish	WER	94.0	3.4	1.3	0.7	0.4	0.2	92.5	6.1	0.9	0.3	0.1	0.04
German	WER	92.4	3.0	2.4	1.2	0.5	0.5						
French	WER	91.2	4.1	2.8	1.2	0.5	0.3	*Language not covered by CVQA*					
Italian	WER	92.6	4.1	1.5	0.8	0.5	0.6						

Table 3: **Linguistic fidelity on the multiple-choice tasks.** Utterances are grouped by ASR error rate, and each cell gives the percentage of a language's utterances that fall in that bin.

the speaker's reference segment, preserving the timbre, accent, and recording conditions of the real speaker. For multiple-choice tasks, the question and options are synthesized separately with the same speaker, with option labels A–D also spoken. For summarization tasks, long passages are split at punctuation into segments of at most 100 characters and rejoined with 0.1 seconds of silence. For Spoken SAMSum, each turn uses the voice assigned to its participant, and every dialogue is rendered with all 25 male–female voice pairings. Appendix A.4 gives the synthesis configuration and the regeneration procedure.

Linguistic and Speaker Fidelity. We verify that the generated speech preserves the source content and matches the reference speaker, and regenerate samples that fail either check with a different synthesis configuration or reference segment from the same speaker.

For linguistic fidelity, we transcribe every utterance with two independent multilingual ASR systems, Omnilingual ASR (Omnilingual ASR team et al., 2025) and Qwen3-ASR (Shi et al., 2026). We compute word error rate against the source text, using character error rate for Chinese, Japanese, and Korean. Table 3 shows the error distribution for the multiple-choice tasks. Across all languages, more than 90% of utterances fall in the lowest error bin, and fewer than 1% exceed 0.5. Results for the summarization tasks are provided in Appendix A.5.1.

For speaker fidelity, we compute cosine similarity between the speaker embeddings of each cloned utterance and its reference audio using w2v-BERT 2.0 (Li et al., 2026). Cloned utterances achieve a mean similarity of 0.673 to their own reference, compared with 0.092 between different speakers of the same language, and the most similar reference is the speaker's own for 99.2% of utterances. This gap indicates that speaker identity is preserved while the voices within each language remain distinct. Per-language heatmaps are provided in Appendix A.5.2.

3 Evaluation Framework and Setup

We evaluate each model along two dimensions. **Performance** measures correctness against r_i, and **consistency** measures whether the output for an instance stays the same across its speakers. Let $y_{i,s}$ be the output for $a_{i,s}$. An output is valid when it contains a parseable final answer for question answering or a non-empty summary for summarization. Let $\mathcal{S}_i$ be the set of evaluation speakers (§3.3) with a valid output for instance i, with $m_i = |\mathcal{S}_i|$.

3.1 Performance Metrics

We score each $y_{i,s}$ against r_i with **accuracy** for question answering and with **BERTScore** F_1 (Zhang et al., 2020) and **ROUGE-1/2/L** (Lin, 2004) for summarization. BERTScore is the main metric, and ROUGE is a lexical complement. Configurations are provided in Appendices B.1 and B.2.

3.2 Consistency Metrics

Question Answering. Let n_{ik} be the number of speakers in $\mathcal{S}_i$ for which the model selects option $k \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$. Consistency is the probability that two speakers drawn without replacement from

$\mathcal{S}_i$ receive the same answer,

$$C_i^{\text{qa}} = \frac{\sum_k n_{ik}(n_{ik}-1)}{m_i(m_i-1)} \times 100\%. \tag{3}$$

Uniform guessing over four options gives 25% in expectation, and a model whose answer never depends on the speaker gives 100%.

Summarization. Because summaries are open ended, we compare them over information units rather than surface forms (Skjærholt, 2014). An LLM-based pipeline extracts named entities, temporal or numerical expressions, and atomic propositions from each $y_{i,s}$, then aligns coreferent mentions across the m_i summaries of the instance, giving a unit set $U_{i,s}$. Consistency is the mean Jaccard similarity over all pairs of valid speakers,

$$C_i^{\text{sum}} = \binom{m_i}{2}^{-1} \sum_{\{s,s'\}\subseteq\mathcal{S}_i} J(U_{i,s}, U_{i,s'}) \times 100\%, \qquad J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{4}$$

Both measures compare speaker pairs and thus require $m_i \geq 2$. Appendix B.3 details the pipeline.

3.3 Experimental Setup

Models. We evaluate nine open-source multimodal LLMs with native audio input support, including Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), Voxtral Small 24B and Mini 3B (Liu et al., 2025), Kimi Audio 7B Instruct (KimiTeam et al., 2025), Gemma 4 E4B and E2B (Team et al., 2026), Gemma 3n E4B and E2B (Team et al., 2025), and Phi-4 Multimodal (Abdin et al., 2024). We exclude Voxtral and Kimi from Spoken CVQA because they do not support visual input.

Instances and Prompting. We evaluate 1,000 instances per language on Global MMLU with 500 culturally sensitive and 500 culturally agnostic questions, the full prepared set per language on CVQA, up to 300 instances per language on CrossSum, and 300 dialogues on SAMSum. Question answering uses chain-of-thought prompting followed by a final answer, and summarization asks for a one- to two-sentence summary without intermediate reasoning. Appendix C gives the prompts, answer parsing, and handling of long audio.

Evaluation Speakers. Due to computational constraints, we evaluate 10 of 50 speakers per language across nine models and four tasks, with five male and five female speakers. We select those with the lowest mean ASR error across utterances to reduce instability caused by poor synthesis. Because this favors cleaner speech, our consistency estimates may be optimistic. Appendix C.1 lists the selected subset.

4 Performance Versus Consistency

4.1 Aggregate Level

Figure 2 compares task performance with within-instance consistency across models and languages, and Tables 21–24 in Appendix D report the per-language results for each task.

Question Answering. The top row of Figure 2 shows that accuracy and consistency are strongly correlated across model–language pairs, with $\rho = 0.97$ on Spoken Global MMLU. The correlation is weaker on Spoken CVQA at $\rho = 0.85$, where pairs with similar accuracy spread vertically. Among CVQA pairs with 45–50% accuracy, consistency ranges from 53.8% for Gemma 3n E2B in Korean to 75.7% for Phi-4-MM in English. Similar accuracy therefore does not imply similar stability.

Summarization. The bottom row of Figure 2 shows a weaker coupling. On Spoken CrossSum, ρ drops to 0.76, and 90% of model–language pairs fall within 87.5–89.5 BERTScore while their consistency ranges from 22.9% to 53.9%. Spoken SAMSum shows the same pattern, as Qwen3-Omni and Gemma 3n E2B differ by only 0.04 BERTScore points but by 11.2 points in consistency. BERTScore rewards semantic overlap with the reference and is insensitive to which acceptable details a summary retains, whereas consistency measures whether that selection changes with the speaker. Performance and consistency are thus related but non-redundant, especially in open-ended generation.

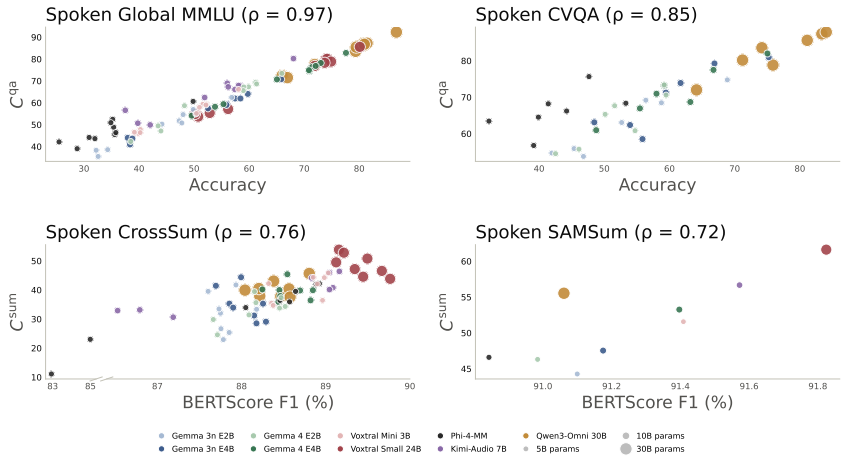

Figure 2: The top row plots accuracy against C^{qa} on the question-answering tasks, and the bottom row plots BERTScore F_1 against C^{sum} on the summarization tasks. Each point is one model and language, and marker size indicates the number of parameters.

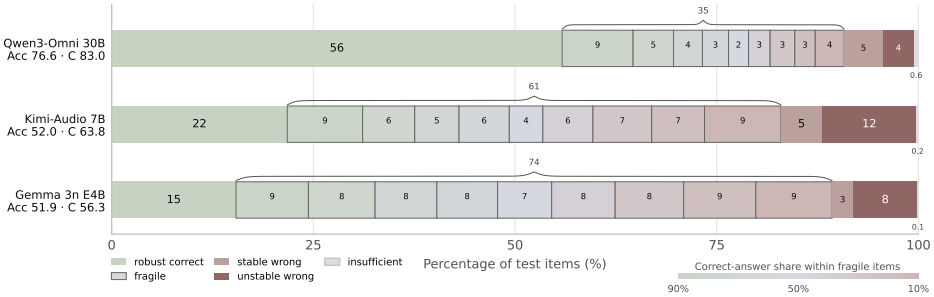

Figure 3: **Item-level outcomes on Spoken Global MMLU.** Each bar splits a model's items by outcomes across speakers. The bracket marks fragile items, shaded from green to red as fewer speakers answer correctly. Bin labels give item percentages, and Appendix E.1 covers all models.

4.2 Instance Level

Question Answering. Aggregate scores do not reveal how broadly speaker variation affects items or whether it changes correctness versus only the wrong answer selected. We therefore group items by their outcomes across speakers. An item is *robust correct* or *stable wrong* when all speakers produce the same correct or wrong answer, *fragile* when only some speakers answer correctly, and *unstable wrong* when speakers produce different wrong answers. Figure 3 shows that fragile items are widespread, accounting for 35% of Spoken Global MMLU even for Qwen3-Omni. Kimi-Audio and Gemma 3n E4B achieve similar accuracy at about 52%, yet Gemma 3n E4B has more fragile items, 74% versus 61%, and lower consistency, 56.3% versus 63.8%. Similar accuracy can therefore mask substantial differences in speaker dependence.

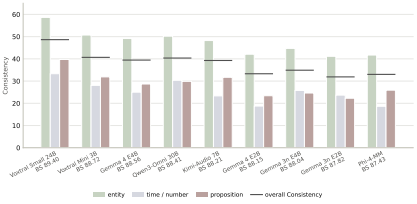

Figure 4: **Consistency by unit type on Spoken CrossSum.** Bars and marks give C^{sum} per unit type and overall, macro-averaged over eight languages.

Summarization. As shown in Section 4.1, similar BERTScore does not imply similar consistency. On Spoken CrossSum, all nine models lie within two BERTScore points, from 87.4 to 89.4, whereas C^{sum} ranges from 31.9 to 48.6. To identify which content changes across speakers, we recompute

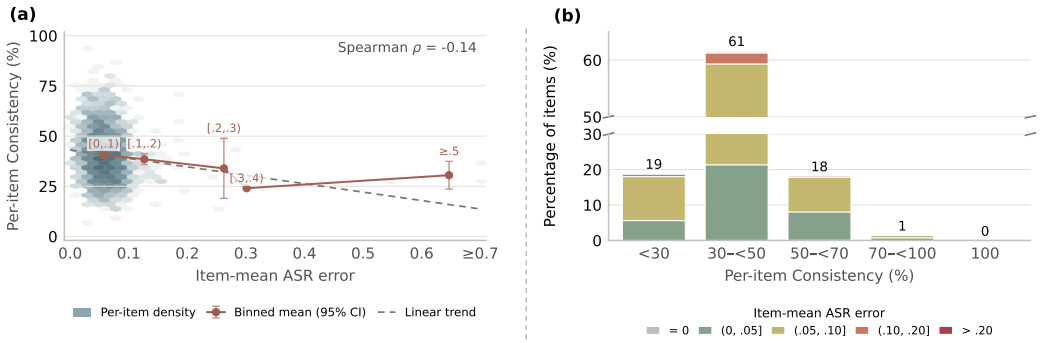

Figure 5: **Synthesis quality versus consistency for Qwen3-Omni on Spoken CrossSum.** (a) Consistency of each instance against its mean ASR error, with binned means and 95% confidence intervals. (b) Share of instances in each consistency range, stacked by ASR-error range.

C^{sum} for each unit type. Figure 4 shows that entities are the most consistent units for every model, exceeding propositions by 16 to 20 points, while temporal and numerical expressions are least consistent for six of nine models. Summaries from different speakers thus mention similar entities but state different facts about them, and Spoken SAMSum shows the same pattern in Appendix E.1.

5 Understanding Speaker-Induced Instability

5.1 Is Instability an Artifact of Synthetic Speech?

Because 3MSpeechBench uses cloned speech, speaker-dependent outputs could reflect synthesis artifacts rather than speaker sensitivity. We test two explanations, namely lexical errors introduced during synthesis and a gap between synthetic and human speech.

Synthesis Quality. If synthesis errors caused the instability, instances with poorly synthesized audio should be less consistent. We quantify synthesis quality by ASR error, averaged over the speaker renditions of each instance. Figure 5a plots this score against consistency for Qwen3-Omni on Spoken CrossSum. The two are only weakly correlated, with $\rho = -0.14$, and even instances with almost no ASR error vary widely in consistency. Figure 5b further shows that low-consistency instances remain common in the cleanest bins. Pooling all models gives similarly weak correlations on Spoken Global MMLU and Spoken CVQA, with $\rho = -0.14$ and $\rho = -0.19$, and the remaining model–task pairs in Appendix E.2 follow the same pattern. Synthesis errors detectable by ASR therefore explain little of the instability.

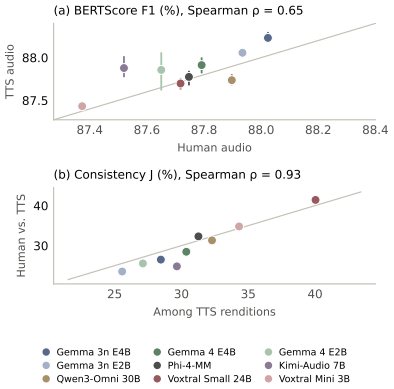

Figure 6: **Human and TTS speech on MCIF.** (a) BERTScore on human versus TTS audio. (b) Consistency among TTS renditions versus between human and TTS renditions.

Human Speech. We test whether our findings extend to human speech using MCIF (Papi et al., 2026), a benchmark of English scientific talks with human recordings. We use its summarization task on 94 talks, enabling a direct synthetic–human comparison. For each talk, six TTS systems re-synthesize the transcript in the speaker's own voice using a clip from the original recording for cloning. All nine models summarize both the human and TTS audio, as detailed in Appendix E.5. Figure 6 shows close agreement between the two settings. Median BERTScore differs by only +0.09 points between TTS and human audio.

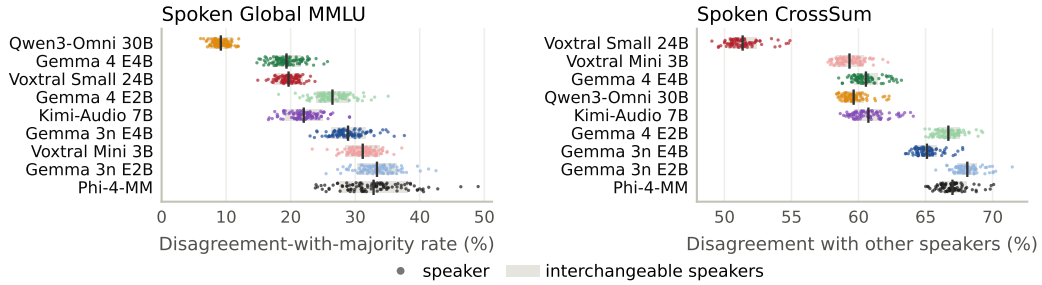

Figure 7: Speaker-level disagreement on Spoken Global MMLU (left) and Spoken CrossSum (right). Dots denote speakers and gray bands the interchangeable-speaker range.

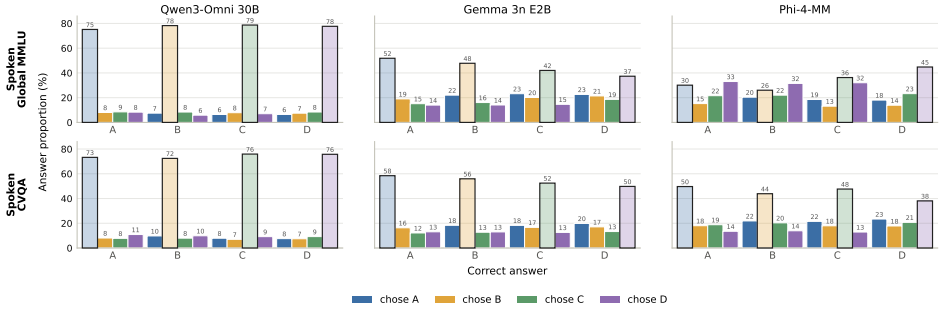

Figure 8: Answer distributions for three representative models on Spoken Global MMLU and Spoken CVQA. Questions are grouped by the correct option, bars show answer proportions across speakers, with the correct option outlined.

Model rankings by consistency across TTS renditions and between human and TTS renditions are also similar, with Spearman $\rho = 0.93$. The consistency patterns observed with synthetic speech therefore largely carry over to human recordings.

5.2 Is Instability Driven by Particular Speakers?

A second possibility is that a few atypical speakers drive most instability. We measure how often each speaker's output differs from others on the same instance and compare it with an interchangeable-speaker band from shuffled speaker identities, as defined in Appendix E.3.

Models Versus Speakers. Figure 7 shows that differences across models far exceed those across speakers. On Spoken Global MMLU, model means range from 9.2% for Qwen3-Omni to 33.4%, while speakers within the same model and language differ by a median of 7.3 points. On Spoken CrossSum, model means range from 51.4% to 68.1%, with a median speaker range of only 3.4 points. Instability is therefore largely model-driven rather than attributable to a few voices. It is also task-specific. Qwen3-Omni is the most stable model on Spoken Global MMLU but only mid-ranked on Spoken CrossSum, where Voxtral Small leads by about eight points.

Residual Speaker Effects. Speakers still matter somewhat. Across the two datasets, 21.4% and 31.1% of speakers fall outside the interchangeable band, compared with 5% expected by chance. Their deviations are moderately reproducible across random instance halves, with split-half ρ of 0.48 and 0.60. Certain voices can reduce model stability, but removing them would not eliminate the broader instability. Spoken CVQA and Spoken SAMSum show the same pattern in Appendix E.3.

5.3 How Is Speaker-Induced Inconsistency Structured?

Having shown that instability is neither a synthesis artifact nor driven by a few speakers, we next ask whether the changes induced by speaker variation are arbitrary or systematic. We find that they largely follow response preferences the model already has.

Figure 8 shows that models prefer particular answer options, even though correct answers are roughly balanced across A–D. These preferences also shape speaker-induced changes. When a speaker moves a model away from its majority answer, the new answer tends to be an option the model already favors. This relation is strong on both Spoken Global MMLU and Spoken CVQA, with ρ of 0.77 and 0.71, as detailed in Appendix E.4. Speaker variation therefore amplifies existing answer preferences rather than producing arbitrary changes.

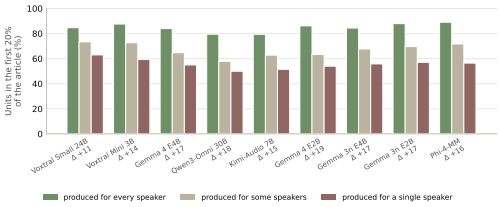

Figure 9: Source position of summary content units on Spoken CrossSum, grouped by how many speakers produce them. Bars show the share of units drawn from the first 20% of the article, and Δ gives the mean position shift from units produced for every speaker to those produced for a single speaker.

Open-ended generation shows a similar structure. Figure 9 shows that content units preserved across all speaker renditions concentrate near the beginning of the source article, whereas units produced by only some speakers or by a single speaker appear progressively later. Speaker variation thus leaves a stable, lead-focused core intact and mainly changes which secondary details are included.

6 Related Work

Speech-capable MLLMs. Recent multimodal large language models have shifted from cascaded ASR-then-LLM pipelines (Zhang et al., 2023; Tang et al., 2024; Chu et al., 2024) to end-to-end architectures that encode audio directly into a shared multimodal representation (Hurst et al., 2024; Google, 2025; Xu et al., 2025). This preserves prosodic and paralinguistic cues often lost in transcription, but also makes model predictions sensitive to speaker-specific acoustic variation that text-based evaluations cannot reveal.

Speech and Audio Benchmarks. Evaluation has expanded from multi-task speech leaderboards (Yang et al., 2021; Tsai et al., 2022; Shi et al., 2023; Huang et al., 2024) and audio understanding benchmarks (Wang et al., 2025; Yang et al., 2024; Sakshi et al., 2025) to multilingual resources (Conneau et al., 2023; Pratap et al., 2020; Di Gangi et al., 2019; Wang et al., 2021; Beyene et al., 2025), multi-speaker corpora (Ardila et al., 2020; Yu et al., 2022), and recent studies on robustness under spoken variation (Lin et al., 2024; Satish et al., 2025; 2026; Chowdhury et al., 2025; Du et al., 2026; Alam et al., 2025; Yang et al., 2025). However, no existing benchmark jointly provides broad multi-speaker and multilingual coverage, diverse task formats including open-ended generation, and explicit within-instance consistency measurement across speakers. Table 2 compares 3MSpeechBench with the speaker-varied benchmarks closest to ours.

7 Conclusion

We presented **3MSpeechBench**, a multilingual, multi-speaker benchmark for evaluating within-instance consistency across closed-form and open-ended tasks. Across nine open MLLMs, we find that consistency captures behavioral variation overlooked by aggregate performance: models can achieve similar task scores while responding differently when only the speaker changes. Further analyses show that this instability is not explained by synthesis quality or a few atypical voices, reflects existing response preferences, and also appears with human speech. These findings motivate reporting consistency alongside performance in spoken MLLM evaluation. Appendix F discusses further limitations.

AI USE STATEMENT

We used ChatGPT and Claude for grammar refinement and occasional assistance with code development. All language model outputs were used solely as references, and all final writing and code were carefully reviewed and produced by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Speaker data. The 500 reference voices are short segments cut from YouTube videos whose pages state a CC BY 4.0 license (Appendix A.3). This license permits reuse with attribution but is not an explicit consent to voice cloning. We therefore only synthesize the text of existing public benchmarks, never new content in a speaker's voice, and the release credits every speaker with a link to the source video. Speaker metadata are recorded only when stated by the speaker or the channel, and we do not infer sensitive attributes such as ethnicity or religion.

Misuse of cloned speech. Cloned voices could be used for impersonation. We release only benchmark audio, not the cloning pipeline or speaker embeddings, and the terms of use prohibit using the audio to impersonate the speakers.

Fairness. Our analyses of gender, age, and acoustic scene are associations measured over five speakers per group and should not be read as claims about demographic groups. Gender labels are binary, and the gender of Spoken SAMSum participants is inferred from their names, which is a simplification.

Licensing and human subjects. The audio follows the licenses of its source corpora (Table 4): it is released under CC BY-NC-SA 4.0, and Spoken SAMSum under CC BY-NC-ND 4.0 (Appendix A.6). No participants were recruited, and all annotations were made by the authors.

REPRODUCIBILITY STATEMENT

Appendix A describes corpus filtering, speaker sourcing and selection, the TTS configuration with its sampling parameters and regeneration procedure, and the quality checks. Appendix B specifies the metrics, including the model and verbatim prompts of the unit-extraction pipeline behind C^{sum}. Appendix C lists the evaluation speakers, model checkpoints, prompts, decoding settings, serving stack, and compute, and Appendix D reports per-language results for every model. Audio samples are available on the anonymous demo site linked in the abstract, and we will release the full audio, metadata, and code for synthesis, inference, and evaluation upon acceptance.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Firoj Alam, Md Arid Hasan, and Shammur Absar Chowdhury. SpokenNativQA: Multilingual everyday spoken queries for LLMs. In *Interspeech 2025*, pp. 2685–2689, 2025. doi: 10.21437/Interspeech.2025-2011.

Rosana Ardila, Megan Branson, Kelly Davis, et al. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

Luel Hagos Beyene, Vivek Verma, Min Ma, et al. mSTEB: Massively multilingual evaluation of LLMs on speech and text tasks. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025. doi: 10.1109/ASRU65441.2025.11434707.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, et al. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2541–2564, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.143. URL https://aclanthology.org/2023.acl-long.143/.

Yiming Chen, Xianghu Yue, Chen Zhang, et al. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026. doi: 10.1162/tacl.a.628.

Amartya Roy Chowdhury, Tonmoy Rajkhowa, and Sanjeev Sharma. Towards multilingual spoken visual question answering system using cross-attention. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9165–9175, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.615/.

Yunfei Chu, Jin Xu, Qian Yang, et al. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Alexis Conneau, Min Ma, Simran Khanuja, et al. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2023. doi: 10.1109/SLT54892.2023.10023141.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. VoxEval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16735–16753, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.818. URL https://aclanthology.org/2025.acl-long.818/.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, et al. MuST-C: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2012–2017, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202.

Yexing Du, Kaiyuan Liu, Youcheng Pan, et al. CCFQA: A benchmark for cross-lingual and cross-modal speech and text factuality evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30575–30583, 2026. doi: 10.1609/aaai.v40i36.40312.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409/.

Google. A new era of intelligence with gemini 3. Google Blog, November 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3/. Accessed 2026-02-13.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, et al. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413/.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, et al. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12136–12140, 2024. doi: 10.1109/ICASSP48485.2024.10448257.

Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Tahir Javed, Janki Nawale, Sakshi Joshi, et al. LAHAJA: A robust multi-accent benchmark for evaluating Hindi ASR systems. In *Interspeech 2024*, pp. 2320–2324, 2024. doi: 10.21437/Interspeech.2024-2376.

KimiTeam, Ding Ding, Zeqian Ju, et al. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Marek Kubis, Pawe Skórzewski, Iwona Christop, et al. Preservation of language understanding capabilities in speech-aware large language models, 2025. URL https://arxiv.org/abs/2509.12171.

Ze Li, Ming Cheng, and Ming Li. Enhancing speaker verification with w2v-BERT 2.0 and knowledge distillation guided structured pruning. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16462–16466, 2026. doi: 10.1109/ICASSP55912.2026.11461727.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yi-Cheng Lin, Wei-Chih Chen, and Hung yi Lee. Spoken StereoSet: On evaluating social bias toward speaker in speech large language models. In *Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 871–878. IEEE, 2024. doi: 10.1109/SLT61566.2024.10832259.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, et al. Voxtral, 2025. URL https://arxiv.org/abs/2507.13264.

Mistral AI. Ministral-3-14b-instruct-2512. https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512, 2025. Hugging Face model card.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, et al. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages, 2025. URL https://arxiv.org/abs/2511.09690.

Sara Papi, Maike Züfle, Marco Gaido, et al. MCIF: Multimodal crosslingual instruction-following benchmark from scientific talks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PtPYZYfa0h.

Diogo Pernes, Gonçalo M. Correia, and Afonso Mendes. Multi-target cross-lingual summarization: a novel task and a language-neutral approach. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12908–12924, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.755. URL https://aclanthology.org/2024.findings-emnlp.755/.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pp. 2757–2761, 2020. doi: 10.21437/Interspeech.2020-2826.

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

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, et al. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 11479–11505, 2024. doi: 10.52202/079017-0366.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, et al. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

Shree Harsha Bokkahalli Satish, Gustav Eje Henter, and Éva Székely. When voice matters: Evidence of gender disparity in positional bias of SpeechLLMs. In *Speech and Computer (SPECOM 2025)*. Springer, 2025. URL https://arxiv.org/abs/2510.02398.

Shree Harsha Bokkahalli Satish, Gustav Eje Henter, and Éva Székely. The voice behind the words: Quantifying intersectional bias in SpeechLLMs, 2026. URL https://arxiv.org/abs/2603.16941. Accepted to Interspeech 2026.

Jiatong Shi, Dan Berrebbi, William Chen, et al. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Interspeech 2023*, pp. 325–329, 2023. doi: 10.21437/Interspeech.2023-1396. URL https://www.isca-archive.org/interspeech_2023/shi23g_interspeech.html.

Xian Shi, Xiong Wang, Zhifang Guo, et al. Qwen3-asr technical report, 2026. URL https://arxiv.org/abs/2601.21337.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, et al. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Arne Skjærholt. A chance-corrected measure of inter-annotator agreement for syntax. In Kristina Toutanova and Hua Wu (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 934–944, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1088. URL https://aclanthology.org/P14-1088/.

Changli Tang, Wenyi Yu, Guangzhi Sun, et al. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, et al. SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8479–8492, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.580. URL https://aclanthology.org/2022.acl-long.580/.

Bin Wang, Xunlong Zou, Geyu Lin, et al. AudioBench: A universal benchmark for audio large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.218. URL https://aclanthology.org/2025.naacl-long.218/.

Changhan Wang, Morgane Riviere, Ann Lee, et al. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting*

of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80/.

Sheng-Lun Wei, Yu-Ling Liao, Yen-Hua Chang, et al. Bias in the ear of the listener: Assessing sensitivity in audio language models across linguistic, demographic, and positional variations. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1570–1589, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.80.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Qian Yang, Jin Xu, Wenrui Liu, et al. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, et al. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*, pp. 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. Speechr: A benchmark for speech reasoning in large audio-language models, 2025. URL https://arxiv.org/abs/2508.02018.

Fan Yu, Shiliang Zhang, Yihui Fu, et al. M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6167–6171, 2022. doi: 10.1109/ICASSP43922.2022.9746465. URL https://doi.org/10.1109/ICASSP43922.2022.9746465.

Dong Zhang, Shimin Li, Xin Zhang, et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

Tianyi Zhang, Varsha Kishore, Felix Wu, et al. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A Benchmark Construction

A.1 Source Corpora

This section complements §2.3 with corpus-level details. Table 4 lists the split, retained instances, and license of each corpus. The paragraphs below describe the corpus-specific restructuring, and Appendix A.2 describes the filtering behind the retained counts.

Global MMLU (Singh et al., 2025) extends MMLU (Hendrycks et al., 2021) to 42 languages. In its annotated subset of 2,850 questions, each question is reviewed by at least three annotators and labeled culturally sensitive (CS) if the majority judge that it requires cultural, geographic, or dialect knowledge, and culturally agnostic (CA) otherwise. We draw from this subset and retain 1,969 questions per language after filtering, of which 588 are CS and 1,381 are CA. We use ten languages that offer reliable TTS quality and overlap with the other corpora: English, Chinese, Japanese, Korean, German, Italian, Spanish, Portuguese, French, and Hindi.

Corpus	Split	Retained		License
Global MMLU	test	1,969	questions per language	Apache 2.0
CVQA	test	7,118	questions	Per item, free for research use
CrossSum	test	3,259	parallel article sets	CC BY-NC-SA 4.0
SAMSum	all	4,437	dialogues	CC BY-NC-ND 4.0

Table 4: **Source corpora.** Corpora are compared by the split we draw from, the number of retained instances, and the license.

CVQA (Romero et al., 2024) consists of questions written by native speakers about culturally situated images. Each item pairs the original-language question with an English translation, which we keep as metadata for cross-lingual alignment and error analysis. CVQA covers seven of our ten languages: English, Chinese, Japanese, Korean, Spanish, Portuguese, and Hindi.

CrossSum (Bhattacharjee et al., 2023) aligns XL-Sum news articles (Hasan et al., 2021) across languages through cross-lingual retrieval. CrossSum covers eight of our ten languages: English, Chinese, Japanese, Korean, Spanish, Portuguese, French, and Hindi. Its alignments are pairwise, whereas our evaluation requires multi-way parallel sets in which the same story appears in several languages. Following Pernes et al. (2024), we build an undirected graph whose nodes are articles and whose edges are CrossSum pairings, and take each connected component as one multilingual set. We use connected components rather than cliques because retrieval-based pairings can be incomplete, and requiring every pair to be linked would split a single story into fragments. We keep components that contain English and at least one other target language, yielding 3,259 parallel sets.

SAMSum (Gliwa et al., 2019) contains about 16k English messenger-style chats, each paired with one abstractive reference summary. We parse each dialogue in the canonical SAMSum format and merge consecutive utterances by the same participant into one turn. Each turn is stored as JSON with its dialogue ID, turn index, participant label, and text, so that voices can be reassigned to turns while the dialogue content stays fixed. We keep the 4,437 dialogues with exactly one male and one female participant, inferring gender from participant names with the `gender-guesser` package.[1] This two-speaker setting allows every dialogue to be rendered under systematic male–female voice pairings, as described in Appendix A.4. We keep Spoken SAMSum in English rather than machine-translating it, since translation artifacts could confound dialogue evaluation.

A.2 Text Rewriting

Text rewriting removes instances that rely on written-only conventions and cannot be read aloud naturally by a TTS model. Table 5 lists the rule-based patterns, each with an example and the corpora it affects. Filtering removes 881 of the 2,850 annotated Global MMLU questions, or 30.9%, of which 204 are CS and 677 are CA. It also removes 2,972 of the 16,368 SAMSum dialogues, or 18.2%, of which 2,966 contain attachment tags and 6 contain URLs. CVQA and CrossSum require no filtering. Figure 10 compares the subject composition of Global MMLU before and after filtering. The STEM share drops from 33.3% to 26.7% because mathematical expressions, chemical formulas, and code occur mostly in STEM questions, while Social Sciences rises from 21.1% to 26.5%. The other four categories change by at most 1.5 points, and every category remains represented.

A.3 Speaker Selection

Sourcing and Screening. Candidate speakers are found by searching YouTube with queries in each target language. A video is considered only if its page states the Creative Commons Attribution license (CC BY 4.0), and each candidate is then checked by listening against the criteria in §2.4: clear speech in the target language, at most mild background noise, and no voice alteration.

Reference Segments. The reference segment is cut manually from the recording in Audacity, selecting a stretch of speech by the target speaker in the target language. Over the 479 reference seg-

[1] https://pypi.org/project/gender-guesser/

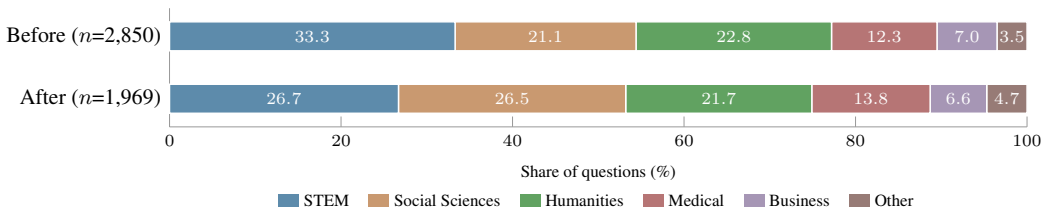

Figure 10: **Subject composition of Global MMLU before and after filtering.** Each bar splits the 2,850 annotated questions or the 1,969 retained questions into six subject categories, with the percentage of each category.

Pattern Category	Example	Affected Corpora
Mathematical expressions	$x^2 + y^2 = z^2$, $\int f(x)dx$	Global MMLU
Chemical formulas	H_2O, NaCl, $C_6H_{12}O_6$	Global MMLU
Code or programming syntax	`for i in range(n)`	Global MMLU
Layout-dependent content	Tables, matrices, aligned equations	Global MMLU
Confidential information	address, phone number	SAMSum
Attachment indicators	`<photo>`, `<file>`, `<link>`	SAMSum
URLs	`http://....com`	SAMSum
Emoji and emoticon codes	`:kisses:`, non-ASCII pictographs	SAMSum

Table 5: Oral-compatibility filtering patterns. We exclude instances containing these written-only constructs that cannot be naturally rendered by TTS. Emoji codes are removed from the dialogue text instead of excluding the dialogue.

ments of the ten languages, the median length is 12.7 s (mean 13.1 s, 5th–95th percentile 10.3–17.2 s, range 8.7–19.5 s). For the 100 evaluation speakers, the median is 12.6 s and the range 8.8–19.5 s. Ten segments, five of them from evaluation speakers, are shorter than 10 s.

Speaker Metadata and Annotation Protocol. Gender, coarse age group, and acoustic scene are annotated from each video by one of the authors, and nationality is recorded only when the speaker states it in the video or its description. Age group and acoustic scene are annotated only when reliably verifiable. As a single author annotates all metadata, no inter-annotator agreement is available. A value is filled in only when the speaker states it in the video or the channel description confirms it, otherwise the field is left empty rather than inferred.

Composition. Tables 6 and 7 summarize the pool by gender, age group, acoustic scene, and self-stated background, and report the per-language distribution of this metadata. The ten evaluation speakers per language are selected from this pool as described in Appendix C.1.

A.4 Speech Synthesis

TTS Configuration. We use Chatterbox version 0.1.4, with the English checkpoint for English and the multilingual checkpoint for all other languages. Every utterance is first synthesized with the default sampling settings of the release (temperature 0.8, classifier-free guidance weight 0.5, exaggeration 0.5, repetition penalty 1.2, min-p 0.05, and top-p 1.0), conditioned on the speaker's full reference segment. Input text is split into segments of at most 60 characters for the multiple-choice tasks and 100 characters for the summarization tasks. This initial synthesis does not fix a random seed.

Segmentation. Segment boundaries are placed preferentially at natural punctuation marks, such as commas and periods, to preserve prosodic coherence and avoid unnatural sentence breaks, and the segments are joined with a short silence to reduce boundary artifacts. Table 8 gives the segment length limit and silence of each task, together with the resulting segment statistics on the synthesized text. Most multiple-choice clips and dialogue turns are synthesized in one piece, whereas a Spoken

		Gender		Age				Acoustic Scene		
Language	**Speakers**	♂	♀	**Adult**	**Middle**	**Elderly**	**Teen**	**Indoor**	**Public**	**Hall**
English (en)	50	25	25	36	9	5	0	19	18	13
Chinese (zh)	50	25	25	37	10	1	2	28	13	9
Korean (ko)	50	25	25	39	9	2	0	27	15	8
Japanese (ja)	50	25	25	41	8	1	0	30	16	4
Hindi (hi)	50	25	25	33	13	4	0	16	27	7
Portuguese (pt)	50	25	25	44	5	1	0	34	14	2
Spanish (es)	50	25	25	34	13	3	0	28	14	8
German (de)	50	25	25	43	7	0	0	19	19	12
French (fr)	50	25	25	27	15	8	0	23	16	11
Italian (it)	50	25	25	35	8	7	0	18	22	10
Total	500	250	250	369	97	32	2	242	174	84

Table 6: **Speaker pool composition.** The 50 speakers of each language are counted by gender, coarse age group, and acoustic scene.

Language	**Nationalities**	**Speakers per Nationality**
English (en)	15	American 20, British 6, Indian 5, Chinese 4, French 3, German 2, Japanese 2; 1 each: Korean, Sri Lankan, Jamaican, Thai, Taiwanese, Nepali, Middle Eastern, Omani
Chinese (zh)	7	Chinese 40, Thai 2, Malaysian 2, American 2, Japanese 2; 1 each: German, Italian
Korean (ko)	8	Korean 42, Indonesian 2; 1 each: Canadian, American, Swedish, Japanese, Italian, Vietnamese
Japanese (ja)	5	Japanese 46; 1 each: Chinese, Korean, American, Russian
Hindi (hi)	1	Indian 50
Portuguese (pt)	3	Portuguese 25, Brazilian 24; 1 each: Angolan
Spanish (es)	10	Spanish 25, Peruvian 7, Argentine 5, Chilean 5, Mexican 3; 1 each: Dominican, Venezuelan, Cuban, Ukrainian, Costa Rican
German (de)	6	German 45; 1 each: Australian, Iraqi, Swiss, Russian, American
French (fr)	5	French 44, American 3; 1 each: Afghan, Iranian, Japanese
Italian (it)	6	Italian 44, American 2; 1 each: Croatian, Mexican, Slovak, Australian
Total	42	

Table 7: **Speaker pool nationalities.** The 50 speakers of each language are grouped by self-stated nationality, listed in descending order of speaker count.

CrossSum article is split into dozens of segments, from a median of 9.5 in Chinese, where each character carries more content, to 58 in Spanish. In Spoken SAMSum, each turn is also followed by 0.1 s of silence.

Task	**Clip**	**Max. chars**	**Silence**	**Single segment (%)**	**Segments per clip mean / median / p95**	**Chars per segment**
Spoken Global MMLU	Question/option	60	5 ms	71.7	1.7 / 1 / 4	33
Spoken CVQA	Question/option	60	5 ms	95.2	1.1 / 1 / 1	17
Spoken CrossSum	Article	100	0.1 s	0.0	43.7 / 37 / 96	69
Spoken SAMSum	Dialogue turn	100	0.1 s	95.6	1.1 / 1 / 1	36

Table 8: **Text segmentation for synthesis.** Max. chars is the segment length limit and Silence the pause inserted between the segments of a clip. The statistics apply the segmentation to the synthesized text of every clip; p95 is the 95th percentile.

Multiple-Choice Rendering. The question, each option, and five short cue phrases (“Question” and “Option A” to “Option D”) are synthesized separately with the same speaker. The cue phrases are in English for every language. The input is assembled as “Question”, the question, and then, for

each option in order, its cue phrase and the option text, with 0.5 s of silence after the question, after every cue phrase, and after every option. Option labels are therefore spoken, and the options always appear in their original order.

SAMSum Voice Pairing. For SAMSum, each dialogue is rendered with every pairing of the five male and five female English evaluation speakers, yielding 25 voice pairings. Within a pairing, the male voice reads every turn of the male participant and the female voice every turn of the female participant.

Audio Format. All audio is stored as single-channel 16 kHz WAV. More precisely, the synthesized questions, options, articles, and turns are kept at the 24 kHz output rate of Chatterbox, and the inputs passed to the models are assembled from them and resampled to single-channel 16 kHz WAV. No loudness normalization is applied.

Regeneration. Utterances whose question-level ASR error, taking for each segment the lower error of the two ASR systems, is at least 0.1 are re-synthesized. Candidates use two lower-temperature configurations, temperature 0.3 with guidance weight 0.5 and temperature 0.1 with guidance weight 0.3, each with a fixed seed derived from the utterance and configuration, so that every candidate can be reproduced. For part of Spoken CrossSum, two shorter reference segments cut from the same recording are tried as well. Before regeneration, numerals and symbols in the input text are written out as words, since the TTS model often misreads digits. For Chinese, Japanese, and Korean, we also restrict a stopping heuristic of the released decoder, which ends generation as soon as a speech token repeats, to the point where the whole text has been read, since it otherwise truncates clauses. A candidate can replace the original only if both ASR systems return a transcript, its error is below 0.1 and below that of the original, and the median speaker similarity of the speaker's candidates is at most 0.03 below that of the speaker's original audio. Among the candidates that pass, we keep the one with the lowest error, breaking ties by higher speaker similarity. Questions without a passing candidate keep their original audio.

A.5 Quality Diagnostics

A.5.1 Linguistic Fidelity

Scoring and Text Normalization. Before scoring, both the transcript and the source text are normalized by Unicode NFKC, lower-casing, and a symmetric rewrite of numerals, units and symbols, so that a digit in the source and its spoken form in the transcript do not count as an error. For languages written with spaces, punctuation becomes whitespace and the unit is the word. For Chinese, Japanese and Korean, whitespace and non-word characters are removed and the unit is the character. We added the normalization after finding that the TTS model mispronounces them. Every dataset-language records the exact scoring profile it was measured with, so numbers produced under different profiles are never mixed. Chinese characters are further compared through their tonal pinyin, so that a homophone written with a different character is not counted as an error while a wrong tone still is. Table 9 illustrates each step.

Step	Lang.	Source text	Transcript	Error
Case, punctuation	en	Is it true?	is it true	0.67 → 0
Numerals, symbols	en	About 50% of the 3 routes are closed.	about fifty percent of the three routes are closed	0.62 → 0
Numerals, symbols	ko	인구의 10%가 서울에 산다.	인구의 십 퍼센트가 서울에 산다	0.38 → 0
Numerals	ja	気温は25度です。	気温は二十五度です	0.44 → 0
Kana reading	ja	東京タワーは333メートルです。	とうきょうタワーは333メートルです	0.38 → 0
Pinyin, numerals	zh	她在2020年获得了第1名。	他在二零二零年获得了第一名	0.50 → 0

Table 9: **Examples of scoring normalization.** Each transcript is a correct reading of its source text, constructed for illustration. The error is WER for English and CER otherwise, computed with our scoring code on the raw strings (left of the arrow) and after normalization (right).

		Spoken CrossSum						Spoken SAMSum					
Language	**Metric**	**[0, 0.1)**	**[0.1, 0.2)**	**[0.2, 0.3)**	**[0.3, 0.4)**	**[0.4, 0.5)**	**$\geq$ 0.5**	**[0, 0.1)**	**[0.1, 0.2)**	**[0.2, 0.3)**	**[0.3, 0.4)**	**[0.4, 0.5)**	**$\geq$ 0.5**
English	WER	99.3	0.6	0.1	0.0	0.0	0.1	92.0	5.8	1.5	0.5	0.1	0.0
Chinese	CER	98.9	1.1	0.0	0.0	0.0	0.0	Not included in this dataset					
Korean	CER	97.3	2.7	0.0	0.0	0.0	0.0						
Japanese	CER	94.3	5.5	0.3	0.0	0.0	0.0						
Hindi	WER	96.9	1.2	1.6	0.1	0.1	0.0						
Portuguese	WER	97.6	0.9	0.6	0.7	0.1	0.0						
Spanish	WER	98.7	0.6	0.3	0.3	0.1	0.0						
French	WER	95.6	4.3	0.1	0.0	0.0	0.0						

Table 10: **Linguistic fidelity of synthesized speech on the summarization tasks.** Percentage of segments in each bin of ASR error rate on Spoken CrossSum and Spoken SAMSum, using WER, or CER for Chinese, Japanese, and Korean. Gray cells mark languages not included in that task.

Language-Specific ASR Mitigations. For Hindi, general-purpose multilingual ASR systems may exhibit elevated error rates due to multi-accent variation (Javed et al., 2024). We therefore additionally employ a Hindi fine-tuned checkpoint, `whisper-large-v3-vaani-hindi`,[2] to improve diagnostic reliability. For Japanese, the coexistence of kanji and kana scripts introduces orthographic mismatches that inflate CER without reflecting true synthesis errors. We normalize both hypotheses and references to kana using `pykakasi`[3] before computing CER.

Summarization Tasks. Table 10 reports the per-segment ASR error-rate distributions for Spoken CrossSum and Spoken SAMSum.

Flagging Criterion. A synthesized question, option set, article, or turn is flagged for regeneration when its ASR error is at least 0.1, where the error of every synthesized clip is the lower of the two systems' error rates. Taking the lower value means that an utterance is flagged only when both systems find errors, since a disagreement between the systems typically reflects a limitation of one recognizer rather than a synthesis failure. The threshold follows from the quality target that we set for every task and language.

Flagging Outcome. Regeneration is applied to the audio of the ten evaluation speakers, and flagged utterances are re-synthesized with the procedure of Appendix A.4, none are removed. Table 11 summarizes how many renditions were flagged and replaced by a candidate that passes both the ASR and the speaker-similarity gates. A flagged utterance without a passing candidate keeps its original audio, and its ASR error is released as metadata, so that users can exclude it. After regeneration, more than 90% of utterances have an error below 0.1 in every task and language, as shown in Tables 3 and 10.

Task	**Unit**	**Renditions**	**Flagged**	**Replaced**
Spoken Global MMLU	Question	196,900	36,517 (18.5%)	21,654
Spoken CVQA	Question	71,180	–	4,335
Spoken CrossSum	Article	61,480	–	10,610 (17.3%)
Spoken SAMSum	Turn	–	–	21,667

Table 11: **Regeneration outcome.** Renditions are counted over the ten evaluation speakers (question renditions include the four options). Replaced renditions passed both the ASR and the speaker-similarity gates. A dash marks a count that our regeneration logs do not record.

A.5.2 Speaker Fidelity

Speaker fidelity is assessed by computing the cosine similarity between w2v-BERT 2.0 (Li et al., 2026) embeddings extracted from the synthesized audio and the corresponding reference audio.

[2] https://huggingface.co/ARTPARK-IISc/whisper-large-v3-vaani-hindi
[3] https://pykakasi.readthedocs.io/

Across languages, the cloned utterances retain high similarity to their source speakers, indicating strong preservation of speaker identity under our synthesis setup.

Same-Speaker versus Cross-Speaker Similarity. Table 12 reports, per language, the mean cosine similarity between a cloned utterance and its own reference against the mean similarity to references of other speakers of the same language. Pooled over all languages, same-speaker similarity is 0.673 and different-speaker similarity 0.092, a per-utterance gap of 0.581. The gap is positive in every language, ranging from 0.469 in Chinese to 0.645 in English. Chinese also has the highest different-speaker similarity (0.175) and the lowest ten-way identification rate (91.8%), whereas identification exceeds 99.9% in every other language.

Language	Same speaker	Different speakers	Gap
English	$0.682_{\pm 0.103}$	$0.037_{\pm 0.022}$	$0.645_{\pm 0.105}$
Chinese	$0.644_{\pm 0.191}$	$0.175_{\pm 0.043}$	$0.469_{\pm 0.182}$
Korean	$0.747_{\pm 0.074}$	$0.130_{\pm 0.024}$	$0.618_{\pm 0.086}$
Japanese	$0.761_{\pm 0.051}$	$0.195_{\pm 0.030}$	$0.566_{\pm 0.066}$
Hindi	$0.672_{\pm 0.111}$	$0.083_{\pm 0.039}$	$0.589_{\pm 0.119}$
Portuguese	$0.658_{\pm 0.083}$	$0.083_{\pm 0.025}$	$0.575_{\pm 0.077}$
Spanish	$0.607_{\pm 0.079}$	$0.071_{\pm 0.031}$	$0.536_{\pm 0.074}$
German	$0.688_{\pm 0.044}$	$0.075_{\pm 0.039}$	$0.612_{\pm 0.068}$
French	$0.660_{\pm 0.132}$	$0.086_{\pm 0.026}$	$0.574_{\pm 0.131}$
Italian	$0.693_{\pm 0.067}$	$0.122_{\pm 0.037}$	$0.570_{\pm 0.089}$
All	$0.673_{\pm 0.113}$	$0.092_{\pm 0.059}$	$0.581_{\pm 0.118}$

Table 12: **Speaker fidelity.** Mean cosine similarity (w2v-BERT 2.0) between a cloned utterance and the reference recording of its own speaker or, on average, of the other nine evaluation speakers of the same language; gap is the per-utterance difference. $\text{Mean}_{\pm \text{sd}}$ over 192,646 utterances of Spoken Global MMLU, Spoken CVQA, and Spoken CrossSum.

Figure 11 breaks the similarity down by speaker: for every language, the mean similarity between the cloned utterances of each evaluation speaker and the reference recording of each evaluation speaker. The diagonal, where utterance and reference come from the same speaker, is clearly brighter than the off-diagonal cells in every language, and the weakest diagonal cells are in Chinese, where the identification rate is lowest.

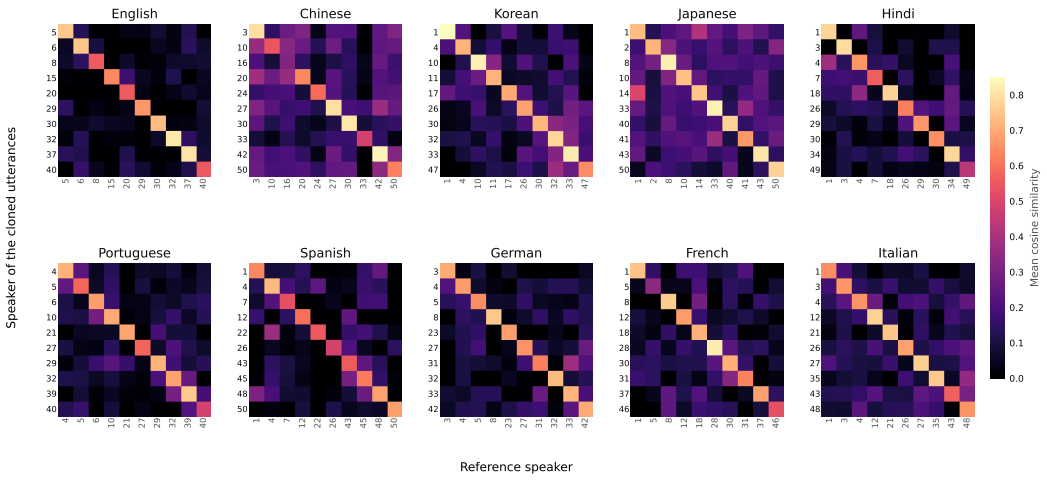

Figure 11: Speaker fidelity per language. Each cell is the mean cosine similarity between the w2v-BERT 2.0 embeddings of the cloned utterances of one evaluation speaker (row) and the reference recording of one evaluation speaker (column), pooled over Spoken Global MMLU, Spoken CVQA, and Spoken CrossSum.

A.6 Benchmark Statistics and Release

Per-Language Statistics. Table 13 lists the number of source instances and hours of speech for every task and language.

	Spoken Global MMLU		Spoken CVQA		Spoken CrossSum		Spoken SAMSum	
Language	Inst.	Hours	Inst.	Hours	Inst.	Hours	Inst.	Hours
English	1,969	643	3,559	512	3,259	2,024	13,386	1,332
Chinese	1,969	544	523	86	478	268	–	–
Korean	1,969	510	290	37	183	94	–	–
Japanese	1,969	577	203	29	310	188	–	–
Hindi	1,969	622	201	34	507	271	–	–
Portuguese	1,969	630	284	44	539	391	–	–
Spanish	1,969	629	2,058	60	681	528	–	–
German	1,969	630	–	–	–	–	–	–
French	1,969	595	–	–	191	123	–	–
Italian	1,969	698	–	–	–	–	–	–
Total	19,690	6,077	7,118	801	6,148	3,886	13,386	1,332

Table 13: Per-language statistics of 3MSpeechBench. Inst. is the number of source instances and Hours the total duration of speech over all 50 speakers, or over all 25 voice pairings for Spoken SAMSum.

Metadata. Each synthesized clip is stored with clip-level metadata that links it back to both the source speaker and the source instance. This includes the language, speaker ID, gender, nationality, age, acoustic-scene label, and item identifiers, together with the source text content used for synthesis, the ASR transcript obtained from the generated audio, and a discretized ASR error bin used for quality analysis. The speaker-related attributes are inherited from the source reference. These metadata support traceability and enable downstream analysis of variation across speakers, environments, and synthesis quality.

Release Format and License. We will release 3MSpeechBench on the Hugging Face Hub. The release contains the synthesized speech of every task and language, organized by task, language, speaker, and source instance, together with the clip-level metadata described above. Utterances that remain flagged after regeneration (Appendix A.5.1) are kept and can be identified through their ASR error. Each speaker is credited with a link to the source YouTube video of the reference recording. The audio is released under CC BY-NC-SA 4.0, as required by the share-alike terms of CrossSum, except Spoken SAMSum, which keeps the CC BY-NC-ND 4.0 terms of its source corpus; the code is released under Apache 2.0. The terms of use prohibit using the audio to impersonate the speakers.

B Evaluation Framework Details

B.1 BERTScore

Formulation. Let h_j be the contextual embedding of the j-th token in predicted summary $\hat{s}$ and r_k be that of the k-th token in reference summary s. BERTScore precision and recall are computed via maximum matching as

$$\mathrm{P}_{\mathrm{BERT}} = \frac{1}{|\hat{s}|} \sum_{j} \max_{k} \ \mathrm{sim}(h_j, r_k), \tag{5}$$

$$\mathrm{R}_{\mathrm{BERT}} = \frac{1}{|s|} \sum_{k} \max_{j} \ \mathrm{sim}(h_j, r_k), \tag{6}$$

where $\mathrm{sim}(\mathbf{h}, \mathbf{r}) = \frac{\mathbf{h}^\top \mathbf{r}}{\|\mathbf{h}\|_2 \|\mathbf{r}\|_2}$ is cosine similarity. The final score is the harmonic mean

	Spoken CrossSum			Spoken SAMSum		
Model	R-1	R-2	R-L	R-1	R-2	R-L
Qwen3-Omni 30B	29.3	10.8	19.7	38.8	12.9	29.4
Voxtral Small 24B	**33.2**	**13.7**	**24.6**	**42.1**	**15.7**	**33.4**
Gemma 4 E4B	29.5	10.3	20.1	39.9	14.0	30.7
Gemma 4 E2B	27.1	8.5	18.2	36.8	11.6	28.2
Gemma 3n E4B	26.8	9.2	18.0	39.9	13.9	29.9
Gemma 3n E2B	25.7	8.4	17.2	39.3	13.6	29.4
Kimi-Audio 7B	25.3	9.1	18.5	40.7	14.6	32.0
Voxtral Mini 3B	30.5	10.9	21.5	40.2	14.1	30.7
Phi-4-MM	23.3	8.1	16.6	36.7	10.8	27.6

Table 14: ROUGE-1, ROUGE-2 and ROUGE-L F_1 (%) on the two summarization tasks, averaged over all speaker realizations of every instance. Tokenization is Unicode-aware: Chinese, Japanese and Korean are tokenized per character, so that ROUGE does not collapse to zero as it would under the default ASCII tokenizer. Scores are computed on the same instances as the BERTScore of Table 23. Best per column in bold.

$$\mathrm{F1}_{\mathrm{BERT}} = \frac{2\,\mathrm{P}_{\mathrm{BERT}}\,\mathrm{R}_{\mathrm{BERT}}}{\mathrm{P}_{\mathrm{BERT}} + \mathrm{R}_{\mathrm{BERT}}}. \tag{7}$$

Model Checkpoint. We use `xlm-roberta-large`[4] as the encoder for all BERTScore computations because it provides coverage for all 10 languages in our benchmark. We disable importance weighting (IDF rescaling) to maintain comparability across languages with different tokenization granularities. BERTScore F1 is particularly suitable for this multilingual setting because it is more robust to paraphrastic and morphological variation than surface-form overlap metrics.

B.2 ROUGE

As a lexical-overlap complement to BERTScore, we compute ROUGE-1, ROUGE-2, and ROUGE-L F_1 (Lin, 2004) on the same outputs. We use a Unicode-aware tokenizer that splits Chinese, Japanese, and Korean text into characters, because the default ASCII tokenizer yields near-zero scores for these languages. Table 14 reports the scores averaged over all speaker realizations of every instance. Voxtral Small 24B obtains the highest score in all six columns.

B.3 Summarization Consistency Pipeline

Overview. The summarization consistency metric C_i^{sum} is computed with an LLM-based extraction-and-alignment pipeline that maps free-form summaries to structured unit sets. We describe the pipeline below.

Unit types. From each summary, we extract three categories of information units.

- **Named entities**, including persons, organizations, locations, and other proper nouns.
- **Temporal and numerical expressions**, including dates, times, quantities, percentages, and monetary values.
- **Atomic propositions**, defined as minimal factual claims that can be independently verified, such as "Company X acquired Company Y" or "the policy was reversed."

Extraction and canonicalization. We use Ministral-3-14B-Instruct-2512 (Mistral AI, 2025) ,served with vLLM. to extract these units from each summary and return them in structured JSON format. We then perform canonicalization in a second pass within each instance. Specifically, the model is given the union of units extracted from all m_i summaries of the same source text and asked

[4] https://huggingface.co/FacebookAI/xlm-roberta-large

to merge semantically equivalent mentions into shared canonical groups. For example, “Joe Biden,” “the president,” and “Biden” are merged when they refer to the same individual in context.

Prompt template. The prompts below are copied from the code used for the reported results, line breaks are added for width. The extraction prompt is as follows.

Unit extraction (system and user message; three summaries per call)

```
[system] You are a strict information extraction engine. Return ONLY
valid JSON. Do not include any extra text.

[user] You will extract information for multiple summaries in one pass.
INPUT_JSON:
{"items": [{"id": "...", "language": "...", "summary": "..."}, ...]}

Return ONLY valid JSON with this exact root schema:
{
  "items": [
    {
      "id": "...",
      "entities": [{"text": "...", "type": "PER|ORG|GPE|LOC|MISC"}],
      "time_numbers": [{"text": "...", "kind": "TIME|NUM"}],
      "propositions": [{"predicate": "...", "arg0": "...",
                        "arg1": "...", "negated": true|false}]
    }
  ]
}

Rules:
1) Output MUST be valid JSON only.
2) Keep one output object for each input id.
3) Every extracted field value MUST be exact substring copied from that
   id's summary, or empty string "".
4) Limits per id: entities<=10, time_numbers<=10, propositions<=8.
5) If nothing found for a field, return an empty list.
```

The canonicalization prompt is as follows, where `{label}` is `entity`, `time_number`, or `proposition`, `{n}` is the number of units, and each unit is listed in its serialized form, such as `ENT::PER::Joe Biden`.

Unit canonicalization (system and user message; one call per instance and unit type)

```
[system] You are a semantic alignment assistant. Identify which text
fragments from paraphrased summaries of the same news article refer to
the same real-world concept.
CRITICAL: Your response MUST be a single valid JSON object and nothing
else. Do NOT include any explanation, commentary, markdown fences, or
extra text. Output ONLY the raw JSON object, starting with '{' and
ending with '}'.

[user] The following are '{label}' units extracted from multiple
summaries of the same news article (same language).
Some units may refer to the same concept using different surface forms
(e.g. abbreviations, partial names, paraphrases).

Units:
  1. ...
  2. ...

Instructions:
1. Group units that are semantically equivalent.
2. Choose the most complete/canonical surface form as the group
   representative.
3. Return ONLY a JSON object: each original unit string (key) ->
   canonical unit string (value).
   Every unit must appear as a key. Non-equivalent units map to
   themselves.
4. Do NOT alter the structured prefix (e.g. 'ENT::PERSON::') - only
   normalise the text portion at the end.

Output format rules (STRICTLY FOLLOW):
- Your entire response must be a single JSON object.
- Start your response with '{' and end with '}'.
- Do NOT wrap the JSON in markdown fences (no ```json or ```).
- Do NOT include any explanation, preamble, or text outside the JSON.
- All {n} input units must appear as keys in the JSON.

Example output shape (do not copy values, use actual mappings):
{
  "<original unit 1>": "<canonical unit>",
  "<original unit 2>": "<canonical unit>"
}

Respond with the JSON object now:
```

Worked example. Table 15 presents an illustrative CrossSum example showing how summaries from different speaker realizations are mapped to aligned unit sets before pairwise Jaccard similarity is computed.

B.4 Consistency Aggregation and Validity Handling

For both task families, consistency is first computed at the instance level and then macro-averaged across instances. This design ensures that each source item contributes equally, regardless of the number of valid outputs available for that item.

For multiple-choice QA, let m_i denote the number of valid parsed answers for instance i. We compute consistency only when $m_i \geq 2$, since pairwise agreement is undefined otherwise. If a model output does not contain a parseable final answer in the required format, that output is excluded from the instance-level computation. The final dataset-level score is the arithmetic mean of all valid instance-level consistency values.

Component	Speaker 5	Speaker 6	Speaker 29
Summary excerpt	Erdogan defended Ozil's decision to leave the German national team and linked it to racist attacks. The summary also mentions Erdogan's upcoming visit to Germany and strained bilateral relations.	Erdogan defended Ozil's decision to leave the German national team and linked it to racist attacks. This version additionally mentions that Ozil retired in July and said he felt scapegoated and disrespected by the German football federation.	Erdogan defended Ozil's decision to leave the German national team and linked it to racist attacks and insults. The summary also mentions Erdogan's upcoming visit to Germany and years of political tension.
Aligned core units	Erdogan; Ozil; Germany; defend(Erdogan, Ozil's decision to leave the German national team); face(Ozil, racist attacks after the photo event)		
Additional aligned units	visit to Germany; strained bilateral ties; political tension	retired in July; felt scapegoated; felt disrespected by the German football federation	insults; visit to Germany; strained bilateral ties; political tension

Table 15: Illustrative example of unit extraction and alignment for summarization consistency. Different speaker realizations lead to summaries that preserve the same core event while differing in secondary contextual details.

For summarization, let P_i denote the set of valid summary pairs for instance i after excluding empty or malformed generations. We compute the mean pairwise Jaccard similarity over aligned unit sets for all pairs in P_i, and report the macro-average across instances. Instances with fewer than two valid summaries are excluded from consistency scoring.

We report standard deviations over instance-level consistency scores rather than over speaker-level outputs. This choice reflects the dispersion of stability across source instances, which is the level at which speaker variation is controlled while semantic content is fixed.

C Experimental Setup Details

C.1 Evaluation Speakers

Table 16 lists the ten evaluation speakers of each language, five male and five female. The same speakers are used for every task in that language, and on Spoken SAMSum the five male and five female English speakers form the 25 voice pairings.

Language	Male speakers	Female speakers
English	5, 6, 8, 15, 20	29, 30, 32, 37, 40
Chinese	3, 10, 16, 20, 24	27, 30, 33, 42, 50
Korean	1, 4, 10, 11, 17	26, 30, 32, 33, 47
Japanese	1, 2, 8, 10, 14	33, 40, 41, 43, 50
Hindi	1, 3, 4, 7, 18	26, 29, 30, 34, 49
Portuguese	4, 5, 6, 10, 21	27, 29, 32, 39, 40
Spanish	1, 4, 7, 12, 22	26, 43, 45, 48, 50
German	3, 4, 5, 8, 23	27, 31, 32, 33, 42
French	1, 5, 8, 12, 18	28, 30, 31, 37, 46
Italian	1, 3, 4, 12, 21	26, 27, 35, 43, 48

Table 16: **Evaluation speakers.** Speaker IDs of the ten evaluation speakers per language. Within each language, IDs 1–25 are male and 26–50 female.

The ten evaluation speakers of each language are chosen after a pilot run: about 500 utterances are synthesized with the candidate speakers and transcribed with ASR, and the five male and five female speakers with the best synthesis quality are kept. Compared with the full pool of 500 speakers, the

Model	Checkpoint	Params	Arch.	Audio	Vision	Max audio input
Qwen3-Omni 30B	`Qwen/Qwen3-Omni-30B-A3B-Instruct`	30B / 3B act.	MoE	✓	✓	not stated
Voxtral Small 24B	`mistralai/Voxtral-Small-24B-2507`	24B	dense	✓	✗	40 min (32K ctx)
Voxtral Mini 3B	`mistralai/Voxtral-Mini-3B-2507`	4.7B (3B LM)	dense	✓	✗	40 min (32K ctx)
Gemma 4 E4B	`google/gemma-4-E4B-it`	4.5B (8B w/emb)	dense	✓	✓	30 s
Gemma 4 E2B	`google/gemma-4-E2B-it`	2.3B (5.1B w/emb)	dense	✓	✓	30 s
Gemma 3n E4B	`google/gemma-3n-E4B-it`	4B (8B w/emb)	dense	✓	✓	30 s
Gemma 3n E2B	`google/gemma-3n-E2B-it`	2B (5B w/emb)	dense	✓	✓	30 s
Kimi-Audio 7B	`moonshotai/Kimi-Audio-7B-Instruct`	7B	dense	✓	✗	not stated
Phi-4-MM	`microsoft/Phi-4-multimodal-instruct`	5.6B	dense	✓	✓	40 s / 30 min summ.

Table 17: **Evaluated models and capabilities.** Parameter counts and maximum audio-input lengths follow vendor specifications. Models without vision are evaluated on all tasks except Spoken CVQA.

evaluation subset has the same gender balance and a similar age distribution (adult 80% against 74%, middle-aged 16% against 19%, elderly 4% against 6%), but it contains more studio recordings (60% against 48%) and fewer recordings in public scenes (21% against 35%), with lecture halls at 19% against 17%. This is consistent with selecting speakers by low ASR error, and it is one reason why consistency on the evaluation subset may be optimistic for the full pool.

C.2 Evaluated Models

Table 17 lists the evaluated models. All models are run locally from the public checkpoints listed in the table, and Appendix C.4 describes the serving stack.

C.3 Prompt Templates and Output Parsing

For multiple-choice QA, we use a task instruction (CoT) that asks the model to reason over the spoken input and select exactly one option among `A/B/C/D`. To ensure parseability across models, we require the final line of the response to follow the format

```
Answer: A
```

with `A` replaced by one of `B`, `C`, or `D` as appropriate. Outputs that do not contain a valid final-line answer are treated as unparseable for accuracy and consistency computation.

For summarization, we prompt the model to generate a concise one- to two-sentence summary of the spoken article or dialog, without requiring intermediate reasoning or chain-of-thought output. We keep the instruction fixed across languages and task instances, except for task-specific framing (article summarization versus dialog summarization).

The prompts are given below. The multiple-choice prompt is shared by Spoken Global MMLU and Spoken CVQA, where the image is passed as an additional input; `{language}` is the language of the article, and `{names}` lists the two participants of the dialogue.

Multiple-choice QA (Spoken Global MMLU and Spoken CVQA)

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

Article summarization (Spoken CrossSum)

```
You are a news summarization system.

Based on the spoken news content provided, write a summary with the
following requirements:
1. Output Language: {language}
2. Length: 1-2 sentences (aim for a short lead-style news intro, similar
   to professional BBC-style summaries)
3. Output only the summary text. Do NOT include a title, bullet points,
   explanations, or any extra commentary.
4. Do NOT add information that is not present in the input. If something
   is uncertain, phrase it conservatively.
```

Dialogue summarization (Spoken SAMSum)

```
You are a dialogue summarization system.

The speakers in this conversation are: {names}.

Based on the spoken conversation provided, write a summary with the
following requirements:
1. Output Language: English
2. Length: 1-2 sentences
3. Output only the summary text. Do NOT include a title, bullet points,
   explanations, or any extra commentary.
4. Do NOT add information that is not present in the conversation.
```

Table 18 reports the share of unparseable outputs. Unparseable outputs are frequent only for Phi-4-MM on Spoken Global MMLU, while all other model–task pairs stay at or below 6.1%, and no summary is empty.

Model	Spoken Global MMLU	Spoken CVQA	Spoken CrossSum	Spoken SAMSum
Gemma 3n E2B	4.4	0.0	0.0	0.0
Gemma 3n E4B	2.0	0.0	0.0	0.0
Gemma 4 E2B	0.5	0.0	0.0	0.0
Gemma 4 E4B	0.6	0.1	0.0	0.0
Phi-4-MM	22.1	0.0	0.0	0.0
Kimi-Audio 7B	6.1	–	0.0	0.0
Qwen3-Omni 30B	4.5	0.1	0.0	0.0
Voxtral Mini 3B	0.7	–	0.0	0.0
Voxtral Small 24B	1.0	–	0.0	0.0
Evaluated outputs	99,736	71,180	21,730	7,500

Table 18: Share of unparseable outputs (%) per model and task. MCQA: outputs without a recognisable A–D answer (no answer or unrecognised text). Summarization: empty summaries. Denominator: all evaluated outputs of the model on that task (last row; CVQA covers the six image-capable models).

C.4 Inference Details

For evaluation-time generation, we set the decoding temperature to 0 whenever this option is exposed by the serving API, in order to reduce variability unrelated to speaker perturbation. This makes the reported consistency less sensitive to stochastic generation and better isolates changes induced by the spoken input.

Audio Chunking and Image Input. For CVQA, images are provided in their original JPEG format without resizing or normalization, thereby preserving realistic input conditions. When a spoken input exceeds a model's maximum audio context window, we partition it into sequential, non-overlapping chunks at the nearest synthesis-defined segment boundary. The chunks are then processed in order while preserving the full prompt and all preceding context across chunks. For models with stricter input-length constraints, such as Phi-4 and the Gemma 3n family, we apply fixed 30-second audio chunking. This deterministic procedure ensures reproducible evaluation across speakers for the same instance.

Qwen3-Omni, Voxtral, Gemma 3n, and Phi-4-MM are served with vLLM, where Phi-4-MM uses its speech adapter and, on Spoken CVQA, its vision adapter. The Gemma 4 models run with Hugging Face Transformers, and Kimi-Audio with its official inference code. Decoding is greedy (temperature 0), with at most 8,192 output tokens for question answering and 256 for summarization.

Inference ran between 5 and 18 September 2026 on NVIDIA H100 GPUs, with part of the Qwen3-Omni runs on H200 GPUs. Table 19 lists the compute per task.

Task	**H100 GPU-h**	**H200 GPU-h**
Spoken Global MMLU	272	117
Spoken CVQA	222	–
Spoken CrossSum	83	5
Spoken SAMSum	7	–
Total (speech input)	584	122
Text input, Qwen3-Omni (Appendix E.6)	9	–

Table 19: GPU-hours per task, summed over all evaluated models and languages.

D Full Results

Table 20 summarizes performance and consistency for every model and task, macro-averaged over languages. The subsections below give per-language results.

	Spoken Global MMLU		Spoken CVQA		Spoken CrossSum		Spoken SAMSum	
Model	Acc	C^{qa}	Acc	C^{qa}	BERTScore	C^{sum}	BERTScore	C^{sum}
Qwen3-Omni 30B	77.4	83.0	76.3	82.2	88.4	40.4	91.1	55.5
Voxtral Small 24B	65.9	70.1			89.4	48.6	91.8	61.6
Voxtral Mini 3B	47.5	54.4			88.7	40.7	91.4	51.6
Gemma 4 E4B	66.0	70.6	60.9	71.5	88.6	39.4	91.4	53.3
Gemma 4 E2B	54.1	60.9	52.0	64.0	88.2	33.3	91.0	46.3
Gemma 3n E4B	52.1	56.3	60.2	69.9	88.0	34.9	91.2	47.5
Gemma 3n E2B	44.3	49.4	53.0	62.8	87.8	31.9	91.1	44.3
Kimi-Audio 7B	52.5	63.8			88.2	39.3	91.6	56.7
Phi-4-MM	34.4	47.4	42.6	66.2	87.4	33.0	90.8	46.6

Table 20: Overview of task performance and speaker consistency across all four spoken tasks. Performance is macro-averaged over the languages of each task: accuracy over valid parsed answers for the question-answering tasks, and BERTScore F_1 (`xlm-roberta-large`) over the eight Spoken CrossSum languages; Spoken SAMSum is English only. Gray background indicates the model does not support the required input modality, so it is not evaluated on Spoken CVQA. C^{sum} is the within-instance consistency of the summaries (mean pairwise Jaccard of aligned content units, %), macro-averaged over languages in the same way.

D.1 Spoken Global MMLU

Table 21 reports per-language accuracy and within-instance consistency on Spoken Global MMLU for all nine models.

	Qwen3-Omni 30B		Voxtral Small 24B		Gemma 4 E4B		Gemma 4 E2B		Gemma 3n E4B	
	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}
en	$86.8_{\pm 0.7}$	$92.3_{\pm 17.1}$	$80.1_{\pm 1.1}$	$85.6_{\pm 22.4}$	$77.6_{\pm 0.5}$	$82.8_{\pm 23.1}$	$65.9_{\pm 1.1}$	$73.5_{\pm 26.3}$	$65.8_{\pm 0.9}$	$70.8_{\pm 26.1}$
zh	$79.3_{\pm 1.3}$	$83.5_{\pm 23.3}$	$52.8_{\pm 1.5}$	$55.4_{\pm 25.3}$	$49.6_{\pm 1.0}$	$54.1_{\pm 25.3}$	$38.5_{\pm 1.9}$	$42.3_{\pm 19.8}$	$38.0_{\pm 1.1}$	$44.0_{\pm 22.0}$
ko	$66.0_{\pm 1.9}$	$72.4_{\pm 27.3}$	$50.8_{\pm 1.3}$	$53.7_{\pm 25.3}$	$55.3_{\pm 2.3}$	$59.4_{\pm 27.0}$	$44.0_{\pm 1.3}$	$47.1_{\pm 21.5}$	$38.4_{\pm 1.7}$	$41.0_{\pm 20.4}$
ja	$67.0_{\pm 1.3}$	$71.5_{\pm 27.7}$	$50.2_{\pm 1.1}$	$54.9_{\pm 26.2}$	$53.8_{\pm 1.7}$	$58.2_{\pm 26.9}$	$43.5_{\pm 2.4}$	$49.5_{\pm 23.1}$	$38.8_{\pm 2.2}$	$43.7_{\pm 21.4}$
hi	$71.9_{\pm 1.5}$	$77.5_{\pm 25.7}$	$56.2_{\pm 1.6}$	$57.2_{\pm 26.9}$	$65.1_{\pm 1.3}$	$70.7_{\pm 26.0}$	$48.3_{\pm 1.9}$	$58.7_{\pm 24.7}$	$52.5_{\pm 1.7}$	$57.5_{\pm 25.9}$
de	$79.4_{\pm 1.0}$	$85.5_{\pm 22.6}$	$71.9_{\pm 0.9}$	$76.9_{\pm 25.4}$	$70.9_{\pm 1.4}$	$74.9_{\pm 25.3}$	$58.9_{\pm 2.4}$	$65.6_{\pm 26.0}$	$55.9_{\pm 2.7}$	$59.0_{\pm 25.9}$
pt	$80.7_{\pm 1.0}$	$86.2_{\pm 22.3}$	$74.2_{\pm 0.5}$	$80.1_{\pm 25.0}$	$71.4_{\pm 1.5}$	$76.3_{\pm 25.6}$	$59.9_{\pm 1.7}$	$67.3_{\pm 26.1}$	$57.5_{\pm 1.8}$	$62.0_{\pm 26.3}$
fr	$81.4_{\pm 0.9}$	$87.3_{\pm 21.7}$	$74.9_{\pm 1.1}$	$78.9_{\pm 25.5}$	$71.0_{\pm 1.5}$	$74.4_{\pm 26.5}$	$59.0_{\pm 1.2}$	$67.2_{\pm 26.1}$	$55.9_{\pm 2.8}$	$59.5_{\pm 26.4}$
it	$80.9_{\pm 1.0}$	$86.8_{\pm 21.5}$	$73.7_{\pm 1.1}$	$78.3_{\pm 25.7}$	$72.1_{\pm 1.4}$	$76.9_{\pm 25.5}$	$61.4_{\pm 1.8}$	$68.6_{\pm 26.2}$	$58.4_{\pm 3.0}$	$62.0_{\pm 26.1}$
es	$80.8_{\pm 1.1}$	$87.2_{\pm 21.6}$	$73.9_{\pm 1.6}$	$80.0_{\pm 24.5}$	$73.0_{\pm 1.5}$	$78.2_{\pm 24.6}$	$61.2_{\pm 1.5}$	$69.4_{\pm 26.2}$	$59.8_{\pm 1.8}$	$64.1_{\pm 26.9}$

	Gemma 3n E2B		Kimi-Audio 7B		Voxtral Mini 3B		Phi-4-MM	
	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}
en	$56.8_{\pm 0.7}$	$62.5_{\pm 26.1}$	$68.1_{\pm 0.8}$	$80.2_{\pm 24.8}$	$58.1_{\pm 1.2}$	$66.2_{\pm 25.9}$	$49.8_{\pm 1.7}$	$60.7_{\pm 27.5}$
zh	$32.2_{\pm 1.2}$	$38.3_{\pm 19.4}$	$56.0_{\pm 1.9}$	$68.9_{\pm 25.7}$	$39.9_{\pm 1.6}$	$46.0_{\pm 21.4}$	$32.0_{\pm 1.3}$	$43.6_{\pm 21.2}$
ko	$32.6_{\pm 1.2}$	$35.5_{\pm 16.4}$	$42.0_{\pm 1.2}$	$49.9_{\pm 24.4}$	$40.3_{\pm 1.3}$	$47.8_{\pm 22.1}$	$28.8_{\pm 1.8}$	$39.0_{\pm 20.6}$
ja	$34.3_{\pm 1.8}$	$38.6_{\pm 18.1}$	$39.8_{\pm 1.3}$	$50.7_{\pm 23.6}$	$39.2_{\pm 1.6}$	$46.5_{\pm 22.1}$	$30.9_{\pm 1.3}$	$44.2_{\pm 25.5}$
hi	$44.2_{\pm 2.0}$	$50.3_{\pm 24.5}$	$37.5_{\pm 1.0}$	$56.6_{\pm 24.3}$	$40.2_{\pm 1.1}$	$46.4_{\pm 21.4}$	$25.5_{\pm 0.9}$	$42.1_{\pm 22.9}$
de	$47.8_{\pm 2.9}$	$50.8_{\pm 23.9}$	$51.9_{\pm 1.0}$	$62.4_{\pm 25.9}$	$50.4_{\pm 1.7}$	$55.2_{\pm 24.7}$	$35.8_{\pm 2.7}$	$46.4_{\pm 23.0}$
pt	$48.0_{\pm 1.3}$	$54.6_{\pm 25.0}$	$58.1_{\pm 1.5}$	$67.9_{\pm 26.8}$	$51.6_{\pm 1.3}$	$59.4_{\pm 25.8}$	$35.6_{\pm 1.5}$	$45.5_{\pm 23.4}$
fr	$47.3_{\pm 2.7}$	$51.8_{\pm 24.2}$	$56.2_{\pm 1.6}$	$67.3_{\pm 26.5}$	$52.2_{\pm 1.5}$	$59.0_{\pm 26.3}$	$35.2_{\pm 1.6}$	$52.5_{\pm 25.4}$
it	$50.3_{\pm 1.9}$	$54.9_{\pm 24.8}$	$57.4_{\pm 1.2}$	$66.1_{\pm 27.4}$	$51.0_{\pm 2.0}$	$57.9_{\pm 25.4}$	$34.9_{\pm 1.5}$	$50.9_{\pm 24.5}$
es	$49.9_{\pm 1.9}$	$56.9_{\pm 25.7}$	$58.4_{\pm 1.0}$	$67.8_{\pm 27.0}$	$52.3_{\pm 0.8}$	$59.5_{\pm 25.6}$	$35.4_{\pm 1.8}$	$48.8_{\pm 24.0}$

Table 21: Accuracy and within-instance consistency (%) per language on Spoken Global MMLU, for the 9 models evaluated in the current analysis. Accuracy is reported over valid parsed answers and its subscript is the standard deviation across the 10 evaluation speakers; the consistency subscript is the standard deviation across instances.

D.2 Spoken CVQA

Table 22 reports per-language accuracy and within-instance consistency on Spoken CVQA. Since the Voxtral models and Kimi-Audio do not support visual input, six models are evaluated: Qwen3-Omni 30B, Gemma 4 E4B and E2B, Gemma 3n E4B and E2B, and Phi-4-MM.

	Qwen3-Omni 30B		Gemma 4 E4B		Gemma 4 E2B		Gemma 3n E4B		Gemma 3n E2B		Phi-4-MM	
	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}	Acc	C^{qa}
en	$74.2_{\pm 0.3}$	$83.6_{\pm 23.3}$	$59.2_{\pm 0.5}$	$73.2_{\pm 26.2}$	$51.6_{\pm 0.5}$	$67.7_{\pm 26.2}$	$61.7_{\pm 0.7}$	$73.9_{\pm 26.1}$	$56.4_{\pm 0.6}$	$69.2_{\pm 27.0}$	$47.7_{\pm 0.5}$	$75.7_{\pm 24.8}$
zh	$83.4_{\pm 1.0}$	$87.4_{\pm 20.5}$	$55.5_{\pm 1.2}$	$66.9_{\pm 26.2}$	$46.1_{\pm 2.0}$	$55.8_{\pm 25.5}$	$54.0_{\pm 2.1}$	$62.4_{\pm 25.9}$	$45.4_{\pm 2.2}$	$55.9_{\pm 25.7}$	$44.2_{\pm 1.7}$	$66.2_{\pm 25.4}$
ko	$75.9_{\pm 1.7}$	$78.8_{\pm 26.9}$	$63.2_{\pm 2.3}$	$68.7_{\pm 27.0}$	$54.7_{\pm 1.8}$	$60.8_{\pm 25.9}$	$55.9_{\pm 3.2}$	$58.5_{\pm 26.9}$	$46.8_{\pm 2.4}$	$53.8_{\pm 26.2}$	$39.2_{\pm 2.4}$	$56.8_{\pm 25.3}$
ja	$64.2_{\pm 1.9}$	$72.0_{\pm 28.0}$	$48.8_{\pm 1.9}$	$61.0_{\pm 27.1}$	$42.5_{\pm 1.4}$	$54.5_{\pm 25.0}$	$48.5_{\pm 1.7}$	$63.1_{\pm 27.7}$	$42.0_{\pm 2.9}$	$54.7_{\pm 27.1}$	$32.3_{\pm 1.5}$	$63.4_{\pm 25.4}$
hi	$84.0_{\pm 1.2}$	$87.9_{\pm 20.9}$	$75.0_{\pm 2.8}$	$82.0_{\pm 24.1}$	$59.5_{\pm 4.9}$	$70.6_{\pm 24.1}$	$75.2_{\pm 1.6}$	$81.0_{\pm 24.4}$	$68.9_{\pm 2.9}$	$74.8_{\pm 27.1}$	$39.9_{\pm 3.0}$	$64.5_{\pm 26.0}$
pt	$81.1_{\pm 1.5}$	$85.6_{\pm 21.9}$	$66.7_{\pm 1.2}$	$77.5_{\pm 26.3}$	$59.2_{\pm 1.6}$	$73.4_{\pm 27.1}$	$66.9_{\pm 1.3}$	$79.3_{\pm 25.1}$	$58.8_{\pm 1.7}$	$68.5_{\pm 27.2}$	$53.3_{\pm 1.4}$	$68.4_{\pm 27.6}$
es	$71.2_{\pm 0.6}$	$80.2_{\pm 25.2}$	$58.0_{\pm 0.6}$	$71.0_{\pm 26.7}$	$50.1_{\pm 1.0}$	$65.3_{\pm 25.9}$	$59.4_{\pm 1.0}$	$71.3_{\pm 26.8}$	$52.7_{\pm 1.5}$	$63.1_{\pm 26.6}$	$41.4_{\pm 1.1}$	$68.2_{\pm 25.6}$

Table 22: Accuracy and within-instance consistency (%) per language on Spoken CVQA, for the 6 models evaluated in the current analysis. Accuracy is reported over valid parsed answers and its subscript is the standard deviation across the 10 evaluation speakers; the consistency subscript is the standard deviation across instances. CVQA uses the v1 answer parser.

D.3 Spoken CrossSum

Table 23 reports per-language BERTScore and within-instance consistency on Spoken CrossSum for all nine models

	Qwen3-Omni 30B		Voxtral Small 24B		Gemma 4 E4B		Gemma 4 E2B		Gemma 3n E4B	
	Perf.	C^{sum}	Perf.	C^{sum}	Perf.	C^{sum}	Perf.	C^{sum}	Perf.	C^{sum}
en	$88.04_{\pm 0.04}$	$39.9_{\pm 10.6}$	$89.16_{\pm 0.06}$	$53.9_{\pm 15.1}$	$88.25_{\pm 0.03}$	$40.2_{\pm 10.8}$	$88.17_{\pm 0.04}$	$35.6_{\pm 10.6}$	$87.69_{\pm 0.03}$	$41.4_{\pm 10.6}$
zh	$88.80_{\pm 0.04}$	$45.7_{\pm 12.3}$	$89.67_{\pm 0.05}$	$46.6_{\pm 15.9}$	$88.82_{\pm 0.05}$	$36.4_{\pm 10.9}$	$87.71_{\pm 0.09}$	$24.6_{\pm 9.0}$	$88.29_{\pm 0.03}$	$29.1_{\pm 10.1}$
ko	$88.58_{\pm 0.05}$	$37.6_{\pm 10.1}$	$89.77_{\pm 0.06}$	$43.8_{\pm 15.3}$	$88.85_{\pm 0.04}$	$39.9_{\pm 11.3}$	$88.09_{\pm 0.06}$	$31.4_{\pm 10.3}$	$88.18_{\pm 0.07}$	$28.5_{\pm 8.4}$
ja	$88.47_{\pm 0.04}$	$37.9_{\pm 10.5}$	$89.44_{\pm 0.06}$	$44.6_{\pm 14.7}$	$88.44_{\pm 0.05}$	$35.9_{\pm 11.0}$	$87.66_{\pm 0.07}$	$29.8_{\pm 8.8}$	$88.15_{\pm 0.07}$	$31.2_{\pm 9.8}$
hi	$88.38_{\pm 0.05}$	$43.1_{\pm 11.2}$	$89.34_{\pm 0.06}$	$47.2_{\pm 15.0}$	$88.54_{\pm 0.04}$	$45.4_{\pm 12.4}$	$88.15_{\pm 0.07}$	$39.4_{\pm 11.8}$	$87.99_{\pm 0.06}$	$44.4_{\pm 11.7}$
pt	$88.22_{\pm 0.04}$	$37.9_{\pm 11.3}$	$89.12_{\pm 0.03}$	$49.5_{\pm 15.5}$	$88.46_{\pm 0.03}$	$37.9_{\pm 11.8}$	$88.45_{\pm 0.04}$	$33.8_{\pm 11.1}$	$87.90_{\pm 0.03}$	$33.9_{\pm 11.1}$
fr	$88.56_{\pm 0.04}$	$40.4_{\pm 12.4}$	$89.50_{\pm 0.04}$	$50.8_{\pm 16.6}$	$88.70_{\pm 0.04}$	$39.8_{\pm 11.8}$	$88.52_{\pm 0.05}$	$34.3_{\pm 10.5}$	$88.26_{\pm 0.06}$	$35.3_{\pm 11.7}$
es	$88.21_{\pm 0.04}$	$40.5_{\pm 12.4}$	$89.22_{\pm 0.04}$	$52.8_{\pm 16.3}$	$88.45_{\pm 0.03}$	$40.0_{\pm 12.3}$	$88.47_{\pm 0.03}$	$37.4_{\pm 11.8}$	$87.86_{\pm 0.04}$	$35.3_{\pm 10.2}$

	Gemma 3n E2B		Kimi-Audio 7B		Voxtral Mini 3B		Phi-4-MM	
	Perf.	C^{sum}	Perf.	C^{sum}	Perf.	C^{sum}	Perf.	C^{sum}
en	$87.60_{\pm 0.04}$	$39.5_{\pm 10.6}$	$89.16_{\pm 0.10}$	$46.4_{\pm 15.4}$	$88.85_{\pm 0.03}$	$44.4_{\pm 13.7}$	$88.64_{\pm 0.06}$	$39.5_{\pm 12.1}$
zh	$87.78_{\pm 0.09}$	$22.9_{\pm 9.0}$	$89.04_{\pm 0.07}$	$46.0_{\pm 13.7}$	$88.96_{\pm 0.08}$	$36.4_{\pm 13.2}$	$88.92_{\pm 0.08}$	$42.2_{\pm 16.4}$
ko	$87.85_{\pm 0.07}$	$25.4_{\pm 8.5}$	$86.53_{\pm 0.13}$	$32.9_{\pm 12.9}$	$88.38_{\pm 0.07}$	$34.7_{\pm 12.6}$	$84.99_{\pm 0.24}$	$23.0_{\pm 11.4}$
ja	$87.76_{\pm 0.05}$	$26.7_{\pm 8.2}$	$86.79_{\pm 0.09}$	$33.1_{\pm 12.5}$	$88.36_{\pm 0.07}$	$35.5_{\pm 11.7}$	$88.05_{\pm 0.12}$	$33.9_{\pm 12.0}$
hi	$87.93_{\pm 0.05}$	$41.7_{\pm 12.2}$	$87.19_{\pm 0.14}$	$30.6_{\pm 13.6}$	$88.32_{\pm 0.04}$	$42.2_{\pm 14.1}$	$82.93_{\pm 0.28}$	$11.1_{\pm 10.1}$
pt	$87.75_{\pm 0.05}$	$32.0_{\pm 10.3}$	$89.04_{\pm 0.06}$	$40.1_{\pm 16.0}$	$88.90_{\pm 0.05}$	$42.1_{\pm 14.7}$	$88.57_{\pm 0.06}$	$35.9_{\pm 16.5}$
fr	$88.18_{\pm 0.06}$	$33.4_{\pm 10.8}$	$89.09_{\pm 0.09}$	$40.8_{\pm 15.1}$	$89.03_{\pm 0.08}$	$45.9_{\pm 16.5}$	$88.89_{\pm 0.06}$	$42.0_{\pm 18.0}$
es	$87.73_{\pm 0.06}$	$33.5_{\pm 10.6}$	$88.83_{\pm 0.08}$	$44.1_{\pm 15.7}$	$88.98_{\pm 0.03}$	$44.2_{\pm 14.5}$	$88.46_{\pm 0.08}$	$36.4_{\pm 15.6}$

Table 23: BERTScore F_1 (Perf.) and within-instance consistency C^{sum} (%) per language on Spoken CrossSum, for the 9 models evaluated in the current analysis. BERTScore uses `xlm-roberta-large` and its subscript is the standard deviation across the 10 evaluation speakers. the consistency subscript is the standard deviation across articles.

D.4 Spoken SAMSum

Table 24 reports BERTScore and within-instance consistency on Spoken SAMSum for all nine models.

	Qwen3-Omni 30B	Voxtral Small 24B	Voxtral Mini 3B	Gemma 4 E4B	Gemma 4 E2B
BERTScore F_1	$91.06_{\pm 0.04}$	$\mathbf{91.82}_{\pm 0.04}$	$91.41_{\pm 0.05}$	$91.40_{\pm 0.05}$	$90.98_{\pm 0.04}$
C^{sum}	$55.5_{\pm 16.1}$	$\mathbf{61.6}_{\pm 17.4}$	$51.6_{\pm 15.4}$	$53.3_{\pm 16.1}$	$46.3_{\pm 14.8}$
	Gemma 3n E4B	**Gemma 3n E2B**	**Kimi-Audio 7B**	**Phi-4-MM**	
BERTScore F_1	$91.18_{\pm 0.04}$	$91.10_{\pm 0.05}$	$91.57_{\pm 0.04}$	$90.84_{\pm 0.07}$	
C^{sum}	$47.5_{\pm 14.1}$	$44.3_{\pm 15.0}$	$56.7_{\pm 16.5}$	$46.6_{\pm 13.6}$	

Table 24: Results on Spoken SAMSum for the nine evaluated models, over 300 dialogues rendered with 25 male–female voice pairings each. BERTScore F_1 uses `xlm-roberta-large`; its subscript is the standard deviation across the 25 pairings. C^{sum} (%) is the within-instance consistency over the 25 pairings. Its subscript is the standard deviation across dialogues. Best value per row in bold.

E Additional Analysis

E.1 Consistency beyond Performance: Additional Results

Item-level taxonomy for all models. Figures 12 and 13 apply the partition of §4.2 to all models on Spoken Global MMLU and Spoken CVQA. On Spoken CVQA, as on Spoken Global MMLU, *fragile* items make up a large share for every model, from 35.4% for Qwen3-Omni to 61.1% for Gemma 3n E2B. The taxonomy also shows that high consistency does not imply correct behavior. Phi-4 and Gemma 3n E4B have similar consistency (71.1 and 71.8), but Phi-4 reaches it with far lower accuracy (44.8 versus 60.5) because 13.3% of its items are *stable wrong*, the largest share among the evaluated models. Its answers on these items do not depend on the speaker, yet they are consistently incorrect.

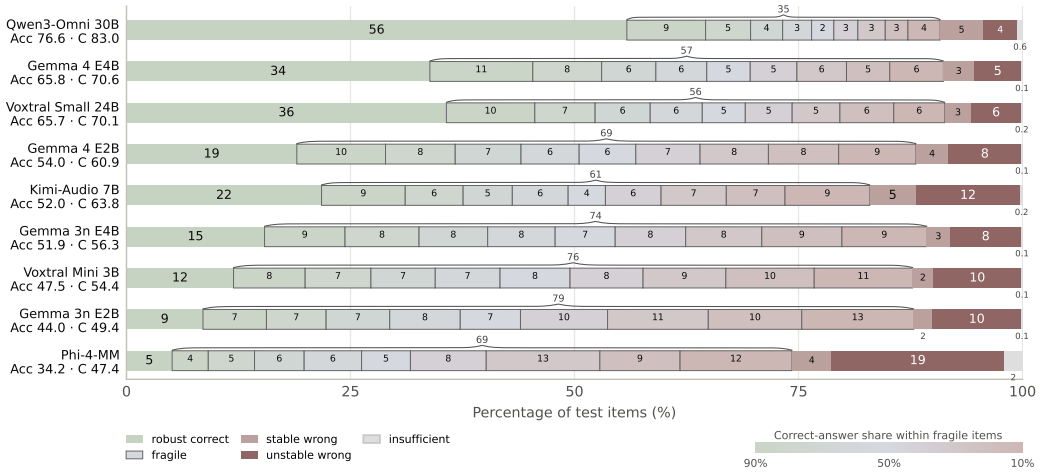

Figure 12: Item-level outcome taxonomy on Spoken Global MMLU. The bracket marks the total share of fragile items, which is further divided by the fraction of speakers answering correctly, from 90% (green) to 10% (red). Bin labels give percentages of all test items.

Consistency and Correctness. Across model–language pairs, accuracy and consistency are strongly rank-correlated, with Spearman $\rho = 0.97$ [0.95, 0.98] on Spoken Global MMLU ($n = 90$) and $\rho = 0.85$ [0.74, 0.92] on Spoken CVQA ($n = 42$), so less accurate models are also less consistent. Within a model, however, inconsistency is not confined to items the model cannot answer. Using the taxonomy of §4.2, fragile items, which are answered correctly for some speakers but not for others, make up between 35.0% (Qwen3-Omni) and 79.3% (Gemma 3n E2B) of the items on Spoken Global MMLU, whereas stable wrong items, on which every speaker leads to the same wrong answer, make up only 2.1% to 5.1%.

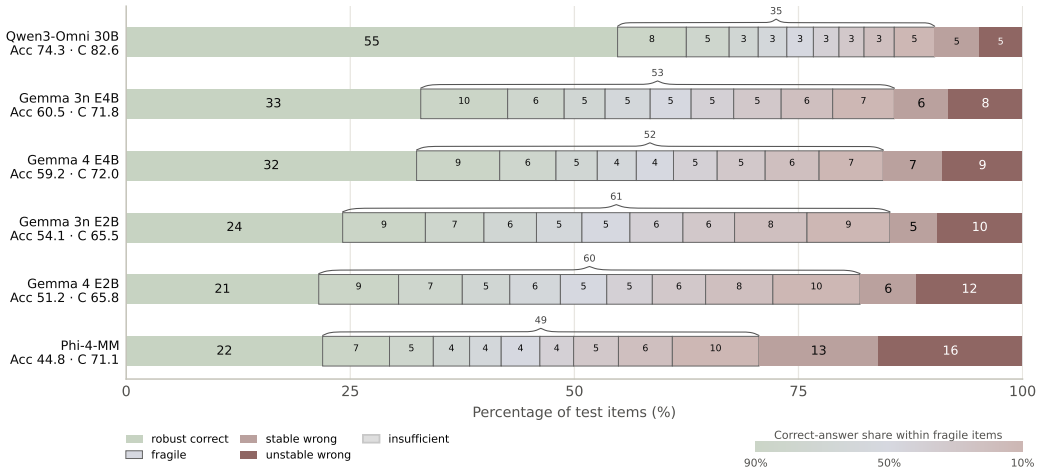

Figure 13: Item-level outcome taxonomy on Spoken CVQA, in the same format as Figure 12. Only the six models that accept image input are evaluated.

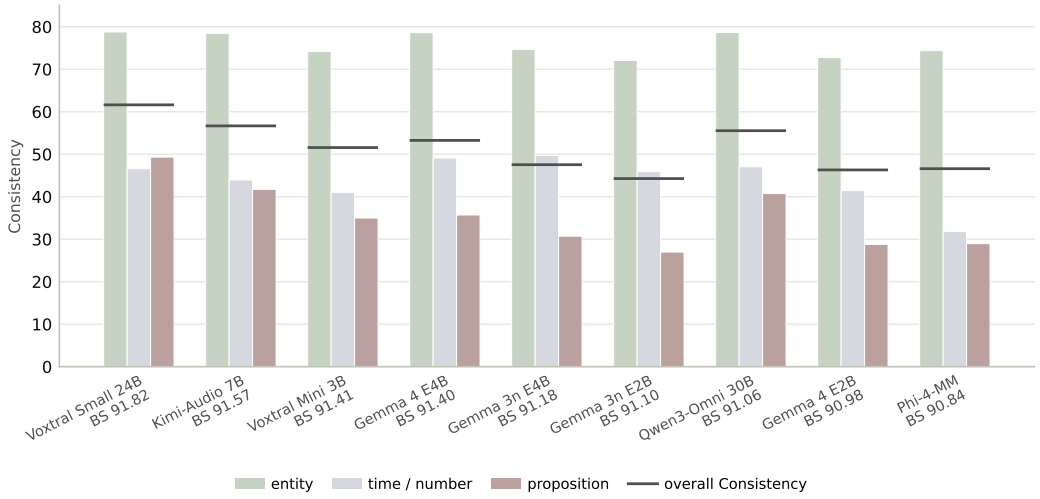

Figure 14: Summarization consistency by information-unit type on Spoken SAMSum, in the same format as Figure 4.

Unit-type decomposition on Spoken SAMSum. Figure 14 repeats the decomposition of Figure 4 on Spoken SAMSum, where consistency is computed over the 25 voice pairings of each dialogue. All nine models lie within one BERTScore point (90.8–91.8), whereas consistency ranges from 44.3 to 61.6. Entity consistency is high and nearly flat across models (72.1–78.8) because the participants of a dialogue are named in almost every summary. Proposition consistency varies from 27.0 to 49.3 and accounts for most of the difference between models. For example, Qwen3-Omni and Gemma 3n E2B have almost the same BERTScore (91.06 and 91.10), but their proposition consistency differs by 13.7 points (40.7 and 27.0). About 44% of dialogues yield no temporal or numerical unit in any summary, so that bar rests on fewer instances.

Same BERTScore, different content. Figure 15 lists all ten speaker renditions of one Spoken CrossSum article for Voxtral Mini and Voxtral Small, whose mean BERTScore is identical. Voxtral Small reports that 60% of wild coffee species are threatened for every speaker. Voxtral Mini reports 60% for seven speakers and 75 threatened species for five, and three of its summaries omit the 60% figure altogether. Both figures appear in the source article, so no summary is wrong, but what a

listener learns depends on who reads the article. BERTScore varies by less than one point across these renditions, whereas C^{sum} separates the two models (24.9 versus 79.7).

A multiple-choice example. Table 25 shows the answers of all nine models to one English Spoken Global MMLU question for each of the ten speakers. Only two models answer correctly for every speaker. The same two speakers (20 and 37) lead Qwen3-Omni, Voxtral Small, and Voxtral Mini to a wrong answer, yet the wrong option they choose differs between models, and the weaker models err for other speakers as well.

	5M	6M	8M	15M	20M	29F	30F	32F	37F	40F	**Correct**
Qwen3-Omni 30B	**C**	**C**	**C**	**C**	A	**C**	**C**	**C**	D	**C**	8/10
Voxtral Small 24B	**C**	**C**	**C**	**C**	B	**C**	**C**	**C**	D	**C**	8/10
Gemma 4 E4B	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	10/10
Gemma 4 E2B	**C**	**C**	D	**C**	**C**	**C**	**C**	**C**	**C**	**C**	9/10
Gemma 3n E4B	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	**C**	10/10
Gemma 3n E2B	B	**C**	**C**	**C**	**C**	B	**C**	B	**C**	A	6/10
Kimi-Audio 7B	**C**	**C**	**C**	B	**C**	**C**	B	**C**	**C**	**C**	8/10
Voxtral Mini 3B	**C**	**C**	**C**	**C**	A	**C**	**C**	**C**	A	**C**	8/10
Phi-4-MM	**C**	B	**C**	B	**C**	A	B	**C**	B	B	4/10

Table 25: **Case study on Spoken Global MMLU** (English, sociology, culturally sensitive). The correct answer is C. Each column is one evaluation speaker (ID and gender), each cell the option chosen by the model for that speaker, and bold marks correct answers.

Same BERTScore, different content. Spoken CrossSum (English), one article read by ten speakers and summarized by two models of the same family. *Reference:* "The first full assessment of risks to the world's coffee plants shows that **60%** of 124 known species are on the edge of extinction."

Speaker	BS	Summary	60%	75 spp.
Voxtral Mini 3B — BERTScore 91.6 (SD 0.9), C^{sum} = 24.9; states 60% in 7/10, 75 species in 5/10				
5 (M)	91.1	Wild coffee species, crucial for the global coffee crop, are facing a **60%** extinction risk, with **75 species** threatened and only **35** not at risk, according to a study published in Science Advances.	✓	✓
6 (M)	90.2	More than 100 types of coffee tree grow naturally in forests, with two used for the coffee we drink. Scientists warn that wild coffee is critical for sustaining the global coffee crop, with **75 wild coffee species** considered threatened with extinction.		✓
8 (M)	92.3	Wild coffee species are critical for sustaining the global coffee crop, with **60%** of wild coffee species threatened with extinction, according to a study published in Science Advances.	✓	
15 (M)	90.7	More than 100 types of coffee tree grow naturally in forests, with **75 wild species** threatened with extinction, and wild coffee is critical for sustaining the global coffee crop.		✓
20 (M) 32 (F)	92.5	Wild coffee species are critical for the global coffee crop, with **60%** of wild coffee species threatened with extinction, according to a study published in Science Advances. *(2 speakers, identical)*	✓	
29 (F)	91.3	More than 100 types of coffee tree grow naturally in forests, with **60%** of them threatened with extinction, according to a study published in Science Advances.	✓	
30 (F)	91.0	More than 100 types of coffee tree grow naturally in forests, with **75 wild species** threatened with extinction, according to a study published in Science Advances.		✓
37 (F)	92.6	Wild coffee species are critical for the global coffee crop, with **75 species** threatened with extinction, and **60%** of coffee species at risk of extinction, according to a study published in Science Advances.	✓	✓
40 (F)	91.4	More than 100 types of coffee tree grow naturally in forests, with about **60%** of wild coffee species threatened with extinction, according to a study published in the journal Science Advances.	✓	
Voxtral Small 24B — BERTScore 91.6 (SD 0.4), C^{sum} = 79.7; states 60% in 10/10, 75 species in 0/10				
5 (M)	91.1	More than 100 types of coffee tree grow naturally in forests, with scientists warning that **60%** of wild coffee species are threatened with extinction, which could impact global coffee production.	✓	
6, 15, 20 (M) 32, 40 (F)	91.4	More than 100 types of coffee tree grow naturally in forests, but scientists warn that **60%** of wild coffee species are threatened with extinction, which could impact global coffee production. *(5 speakers, identical)*	✓	
8 (M)	92.0	More than 100 types of coffee tree grow naturally in forests, with **60%** of wild coffee species threatened with extinction, according to scientists who warn that this poses a risk to the global coffee crop.	✓	
29 (F)	92.3	More than **60%** of wild coffee species are threatened with extinction, according to a new study, which highlights the need for better conservation measures to safeguard the future of the global coffee crop.	✓	
30 (F)	91.4	More than 100 types of coffee trees grow naturally in forests, with **60%** of wild coffee species threatened with extinction, according to scientists who warn that wild coffee is critical for sustaining the global coffee crop.	✓	
37 (F)	91.7	More than 100 types of coffee tree grow naturally in forests, but scientists warn that **60%** of wild coffee species are threatened with extinction, which could impact the global coffee crop.	✓	

Figure 15: Example of a consistency difference that BERTScore does not capture (Spoken CrossSum, English). BS = BERTScore F1 (%) of each summary against the reference; M/F = speaker gender; identical summaries are merged into one row; bold marks the key figures.

E.2 Synthesis Quality and Consistency for All Models

Figures 16 and 17 extend the ASR-proxy analysis in §5.1 across models and tasks. Table 26 reports the corresponding Spearman correlations. The association between mean ASR error and consistency is generally weak. Some model-task pairs show a modest decline at higher error rates, but no consistent monotonic relationship accounts for the observed instability.

The binned distributions lead to the same conclusion. Low-consistency instances remain present in the lowest-error region across models, where little lexical corruption is detected. ASR-observable lexical errors may therefore exacerbate instability, but they are not its primary explanation.

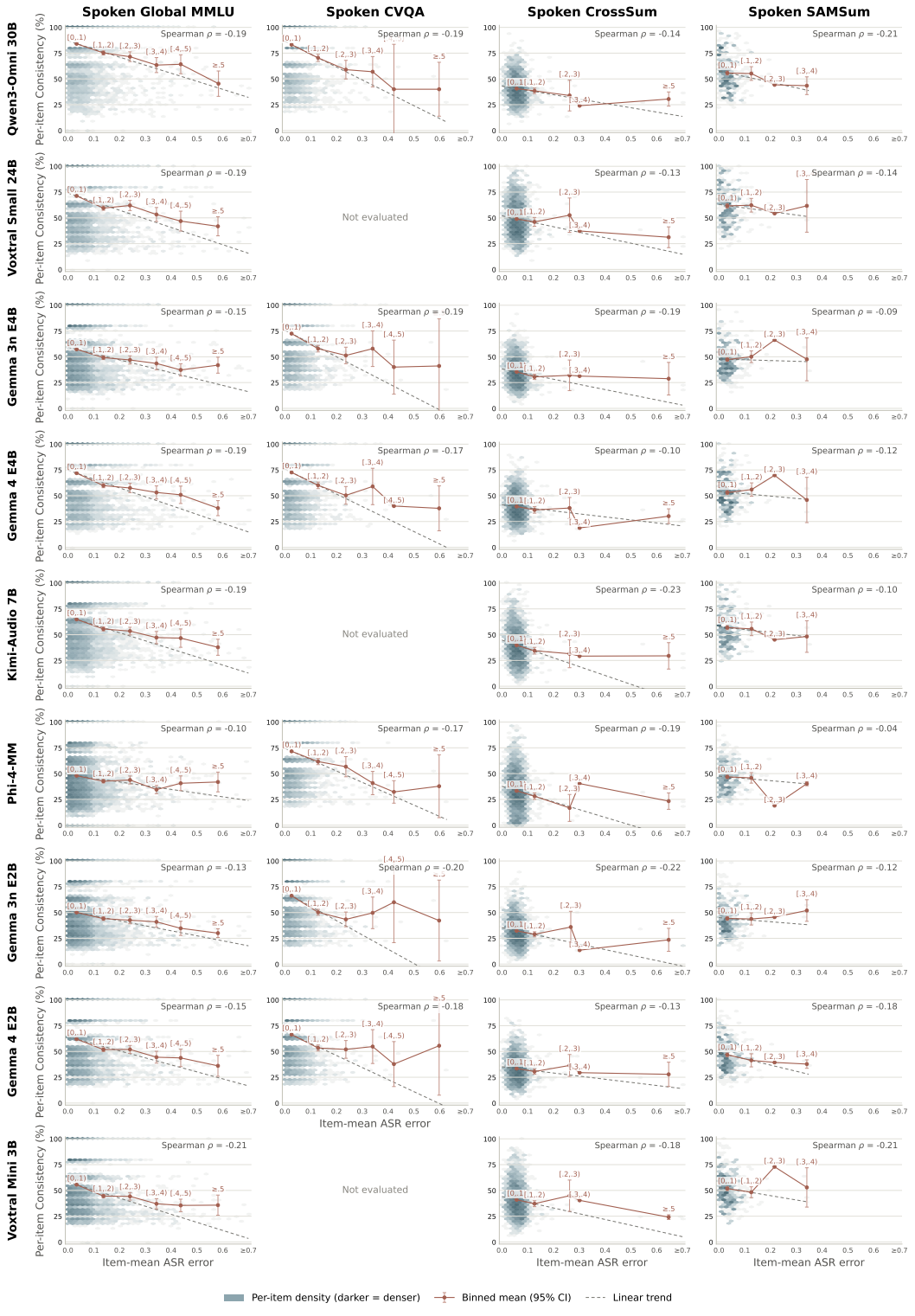

Figure 16: Instance-level relation between the ASR proxy and consistency across models and tasks. The ASR proxy is the mean WER, or CER for Chinese, Japanese, and Korean, over the speaker renditions of an instance.

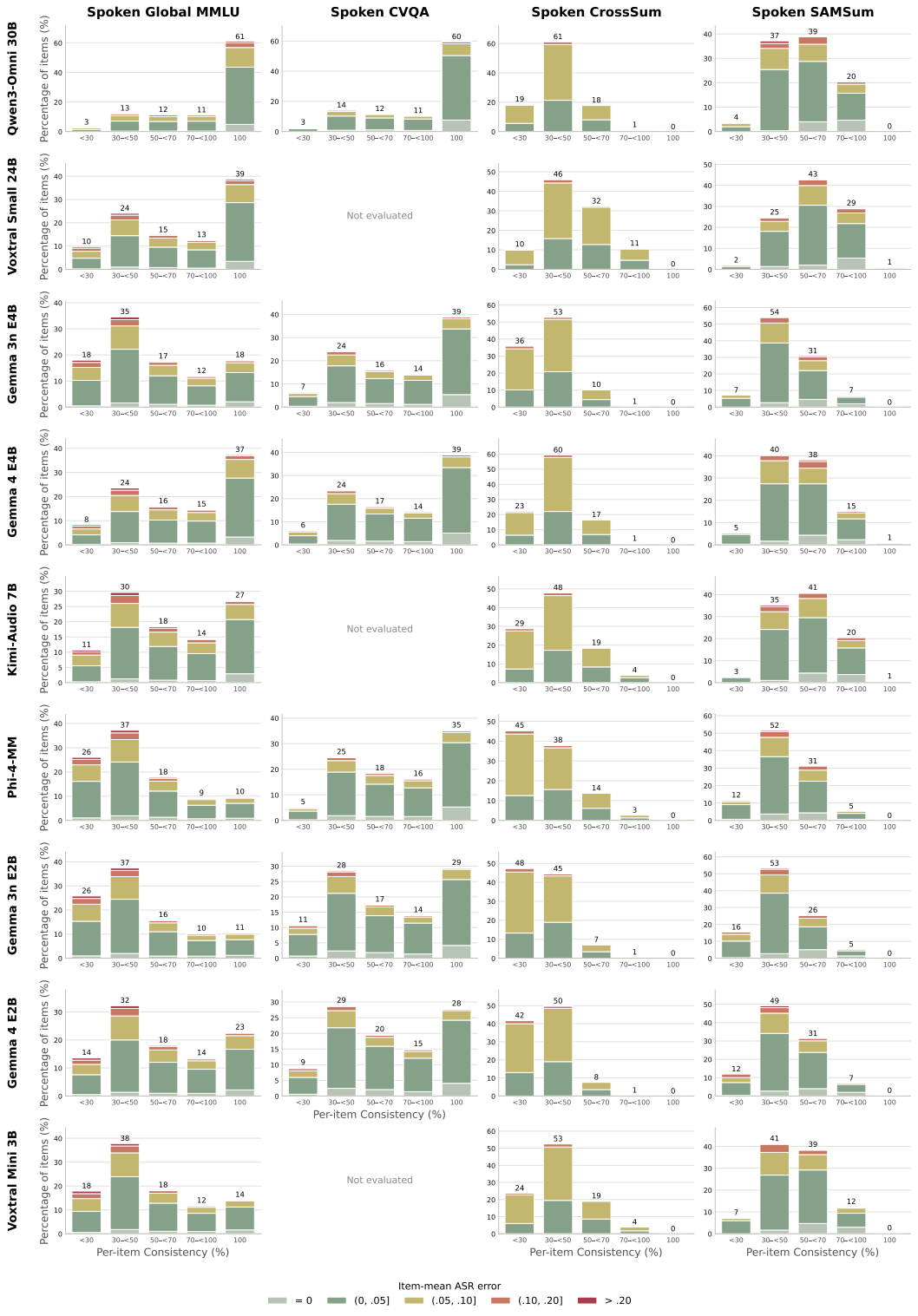

Figure 17: Distribution of instance-level consistency within ASR-error ranges across models and tasks. Inconsistent instances remain common in the lowest-error range.

Model	Spoken Global MMLU		Spoken CVQA	
	ρ(ASR, C)	ρ(ASR, Acc)	ρ(ASR, C)	ρ(ASR, Acc)
Qwen3-Omni 30B	−0.2	−0.1	−0.2	−0.1
Voxtral Small 24B	−0.2	−0.1		
Voxtral Mini 3B	−0.2	−0.1		
Gemma 4 E4B	−0.2	−0.1	−0.2	−0.1
Gemma 4 E2B	−0.1	−0.1	−0.2	−0.1
Gemma 3n E4B	−0.1	−0.1	−0.2	−0.1
Gemma 3n E2B	−0.1	−0.0	−0.2	−0.1
Kimi-Audio 7B	−0.1	−0.1		
Phi-4-MM	−0.1	−0.0	−0.1	−0.1
Pooled over all instances, cell-demeaned:	−0.141 / −0.078		−0.187 / −0.105	

Table 26: **Synthesis quality does not explain speaker-induced inconsistency.** Spearman correlations between the per-instance ASR proxy (mean WER, or CER for Chinese, Japanese and Korean, min-merged over the two recognizers) and per-instance consistency C and accuracy, averaged over languages. The pooled row gives the cell-demeaned correlation over all instances ($n = 89{,}629$ for Spoken Global MMLU, 42,708 for Spoken CVQA). Gray cells mark models without visual input.

E.3 Speaker-Level Instability: Additional Results

Definitions. For question answering, we use the items on which all ten speakers give a valid answer. The majority answer of an item is the option chosen most often, and a speaker's disagreement-with-majority rate is the share of items on which its answer differs from it. For Spoken CrossSum, a speaker's score on an item is 100 minus the mean Jaccard similarity between its unit set and those of the other nine speakers. Spoken SAMSum renders each dialogue with 25 pairings of five male and five female voices, and the pairing of voices is identical across dialogues. We therefore score each pairing against the other 24 and average over the five pairings in which a voice appears. Within each model and language, rates are centred on the mean of the speakers and shifted to the model mean. The interchangeable-speaker band is the 2.5–97.5% range of the same centred rates after shuffling speaker identities within each instance (200 permutations).

Spoken CVQA and Spoken SAMSum. Figure 18 repeats Figure 7 for the remaining datasets. On Spoken CVQA, model means range from 11.8% to 25.4%, whereas speakers within a model and language differ by a median of 8.8 points. On Spoken SAMSum, model means range from 38.4% to 55.7%, whereas the ten voices of a model differ by a median of only 1.1 points, and male and female voices are intermixed. Averaging each voice over five pairings reduces the SAMSum spread, so its magnitude is not directly comparable with the other datasets.

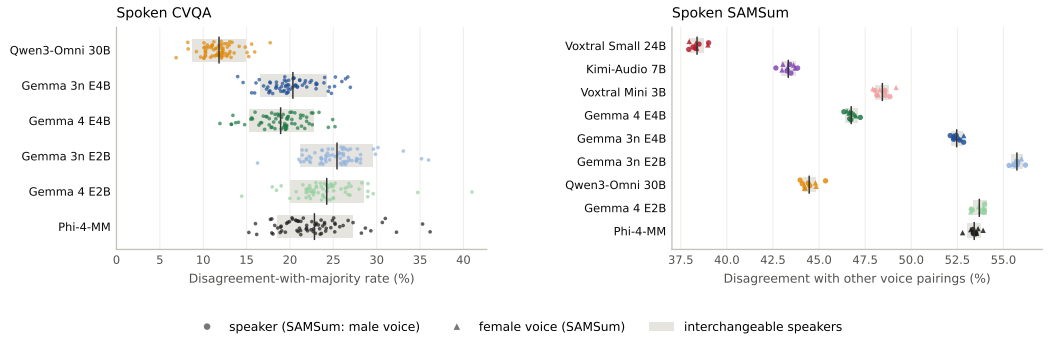

Figure 18: Speaker-level disagreement on Spoken CVQA (left) and Spoken SAMSum (right), in the same format as Figure 7. For Spoken SAMSum, each point is one voice (circles: male, triangles: female).

Stability of speaker effects. Table 27 summarizes, for every dataset, how large and how reproducible speaker differences are. Reproducibility is measured by splitting the instances of every model and language at random into two halves and computing the Spearman correlation between the centred speaker rates on the two halves, pooled over models and languages.

	Global MMLU	CVQA	CrossSum	SAMSum
Range of model means (%)	9.2–33.4	11.8–25.4	51.4–68.1	38.4–55.7
Median speaker range (pp)	7.3	8.8	3.4	1.1
Speakers outside the interchangeable band (%)	21.4	16.4	31.1	30.0
Split-half reliability (Spearman ρ)	0.48	0.40	0.60	0.24

Table 27: **Size and reproducibility of speaker effects**. The speaker range is the difference between the most and least deviating speakers of a model and language. About 5% of speakers would fall outside the interchangeable band by chance. Spoken SAMSum has one language and ten voices per model, so its reliability rests on 90 points.

Variance Decomposition. For each model and language of the two question-answering tasks, we decompose the variance of answer correctness over the grid of items by ten speakers, restricted to items with a valid answer from every speaker, into an item component, a speaker component, and a residual that combines the item-by-speaker interaction with generation noise. Averaged over models and languages, items account for 46% of the variance on Spoken Global MMLU and 56% on Spoken CVQA, speakers for 0.1% on both, and the residual for 54% and 44%. Speaker main effects are thus negligible, and speaker-induced variation arises almost entirely from how a speaker interacts with a particular item, consistent with §5.2 and Appendix E.3.

E.4 Structure of Speaker-Induced Disagreement: Additional Results

Answer distribution for all models. Figures 19 and 20 extend Figure 8 to all evaluated models. The correct answers are close to uniform over the four letters (22.9–26.6% on Spoken Global MMLU, 24.9–25.1% on Spoken CVQA). On Spoken Global MMLU, the strongest preference is Kimi-Audio's toward C: when C is wrong, it still receives 24–26% of the answers, and accuracy is 64.3% when the answer is C but 41.8% when it is A. On Spoken CVQA, letter preferences are weaker and most models spread their wrong answers nearly evenly; the clearest exception is Phi-4, which favors A and avoids D, with an accuracy of 49.7% when the answer is A and 38.1% when it is D.

Where deviating speakers go. Figure 21 gives the evidence behind the correlation reported in Section 5.3. For an item on which the speakers disagree, a deviating speaker is one whose answer differs from the item's majority answer. For every model and letter k, the vertical axis is the share of deviating answers that choose k, among items where k is not the majority answer, minus the uniform $1/3$. The horizontal axis is the model's option bias toward k among wrong answers: the observed share of k minus its exposure-normalized expectation $P(\text{key} \neq k)/3$.

	Human	Chatterbox	CosyVoice3	F5-TTS	IndexTTS2	Qwen3-TTS	XTTS-v2
Gemma 3n E4B	88.02_{1}	88.19_{1}	88.20_{1}	88.21_{1}	88.21_{1}	88.30_{1}	88.28_{1}
Gemma 3n E2B	87.93_{2}	88.05_{3}	88.03_{2}	88.05_{2}	88.05_{2}	88.05_{2}	88.11_{2}
Qwen3-Omni 30B	87.90_{3}	87.68_{8}	87.72_{6}	87.77_{6}	87.69_{7}	87.76_{7}	87.81_{5}
Gemma 4 E4B	87.79_{4}	87.93_{5}	87.82_{4}	87.88_{5}	88.00_{3}	87.92_{4}	87.93_{3}
Phi-4-MM	87.75_{5}	87.85_{6}	87.78_{5}	87.68_{8}	87.81_{5}	87.79_{6}	87.75_{6}
Voxtral Small 24B	87.72_{6}	87.68_{7}	87.71_{7}	87.76_{7}	87.67_{8}	87.74_{8}	87.63_{8}
Gemma 4 E2B	87.65_{7}	88.06_{2}	87.61_{8}	87.95_{4}	87.83_{4}	87.95_{3}	87.74_{7}
Kimi-Audio 7B	87.52_{8}	87.93_{4}	87.82_{3}	88.02_{3}	87.77_{6}	87.86_{5}	87.87_{4}
Voxtral Mini 3B	87.37_{9}	87.44_{9}	87.40_{9}	87.44_{9}	87.45_{9}	87.45_{9}	87.41_{9}

Table 28: **BERTScore F_1 on MCIF by input condition.** Mean over the 94 talks with outputs for every model and condition; the subscript is the rank of the model within each column.

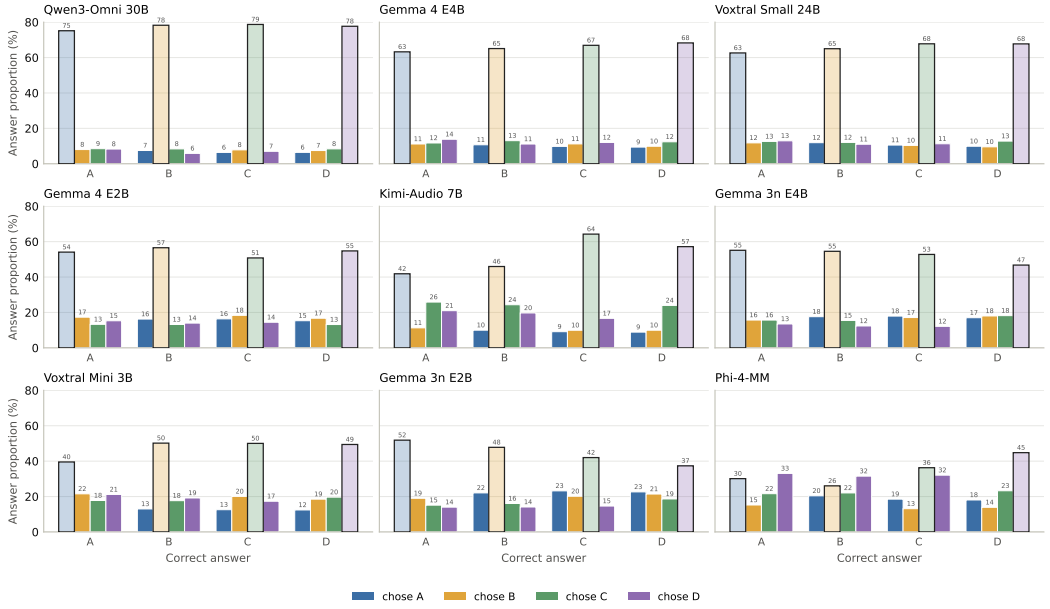

Figure 19: Answer distribution by correct answer for all nine models on Spoken Global MMLU, in the same format as Figure 8. Models are ordered by accuracy.

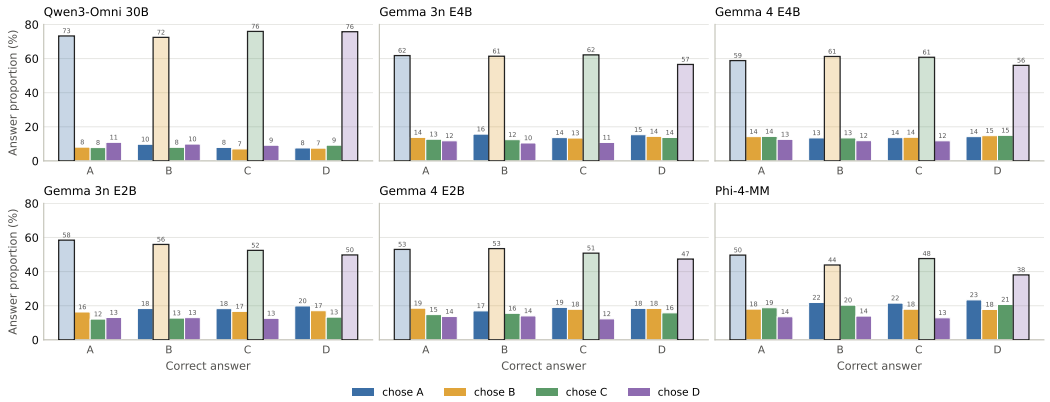

Figure 20: Answer distribution by correct answer on Spoken CVQA, in the same format as Figure 8. Only the six models that accept image input are evaluated; models are ordered by accuracy.

Reference drawn mainly from	Summary	First 20%	20–40%	40–100%
First 20% (584 articles)	Human reference	70	11	19
	Models (mean)	67	13	20
20–40% (134 articles)	Human reference	17	**61**	21
	Models (mean)	**61**	18	21
40–100% (152 articles)	Human reference	19	10	**72**
	Models (mean)	**59**	14	26

Table 29: Source position of summary content on Spoken CrossSum, with articles grouped by the region of the article that contributes most to the human reference summary. Each cell is the share (%) of summary shingles found in that region of the article; models are averaged over all nine models and all speaker renditions. Only articles whose reference has at least five shingles found in the article are included.

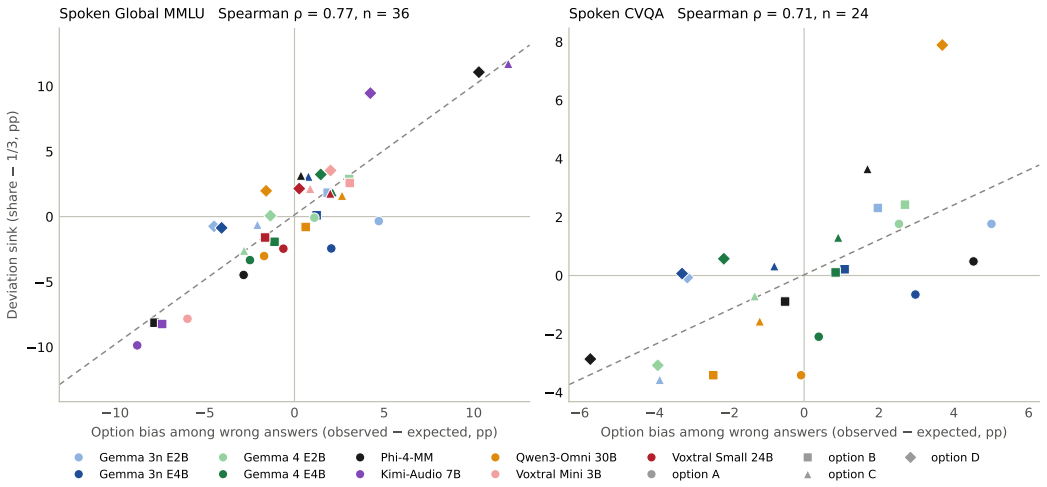

Figure 21: Letters chosen by deviating speakers against the model's option bias, for Spoken Global MMLU (left) and Spoken CVQA (right). Each point is one model and letter; the dashed line is a least-squares fit, and ρ is Spearman's correlation.

Position preference in summarization. Table 29 locates summary content in the Spoken Cross-Sum source article. Summaries and articles are cut into shingles (three words, or four characters for Chinese, Japanese, and Korean), and each summary shingle found in the article is placed at its first occurrence. Articles are grouped by the region that contributes most to the human reference summary. When the reference draws mainly on the first 20% of the article, the models draw on the same region to a similar extent. When the reference draws mainly on later parts, it takes only 17–19% of its content from the first 20%, whereas the models still take about 60% from it. The models' preference for the beginning of the article therefore persists regardless of where the reference locates the relevant content. Grouping articles by the reference's own content inflates the reference's share in its dominant region, but it does not raise the models' share in the first 20%. With stricter inclusion (at least 10 or 15 reference shingles found in the article), the models' share of content from the first 20% remains 57–62% in both later-region groups.

E.5 Synthetic versus Human Speech: Additional Results

Setup. MCIF provides 100 English scientific talks with human recordings and transcripts. For each talk, a reference clip is extracted manually from the talk's own recording, and the full transcript is synthesized by six TTS systems from the same text chunks and the same reference clip. All nine models summarize every talk under the seven conditions with the official MCIF summarization prompt, giving 6,246 summaries. Some TTS outputs are missing for a few talks (four for XTTS-v2 and one each for F5-TTS and Qwen3-TTS), so all paired comparisons use the 94 talks with outputs for every model and condition. Performance is BERTScore F_1 against the MCIF reference summary with the configuration of Appendix B.1. Consistency uses the pipeline of Appendix B.3, with the seven conditions of a talk taking the place of speakers. All intervals are 95% bootstrap intervals over talks (2,000 resamples).

Absolute scores. Table 28 reports BERTScore by model and condition. Synthetic minus human BERTScore has a median of $+0.09$ points over the 54 model–TTS pairs and ranges from -0.22 to $+0.50$. With an equivalence margin of ± 0.25 points, 30 of the 54 pairs are equivalent (two one-sided tests at the 90% level), and all 54 are equivalent at ± 1 point. We use the tighter margin because BERTScore separates models very little on MCIF: all 63 model–condition means lie within 0.93 points, and within 0.65 points on human speech. Models account for 84.6% of the variance of these means, conditions for 4.0%, and their interaction for 11.4%.

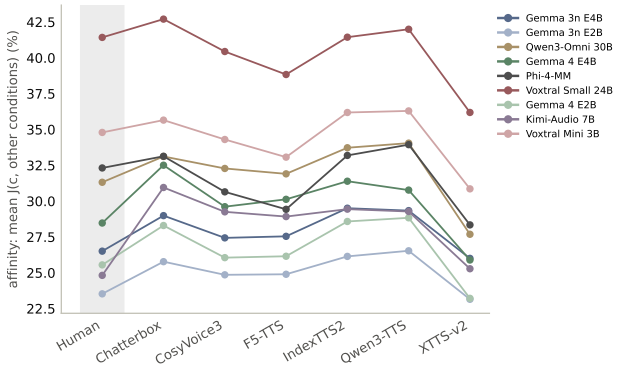

Figure 22: **Agreement of each input condition with the others on MCIF.** For every model, the mean Jaccard similarity between the summaries under one condition and those under the other six conditions of the same talk. Human speech is shaded.

Consistency. For each model, we compare the mean Jaccard similarity between human and TTS renditions of a talk with that between two TTS renditions. The difference is between -1.6 and -2.0 points for the four Gemma models, whose intervals exclude zero, -4.8 points for Kimi-Audio, and not distinguishable from zero for the other four models, while consistency itself ranges from 24 to 41 points across models. The ranking of models is nearly unchanged (Spearman $\rho = 0.93$ [0.78, 0.98], and 0.98 without short or truncated talks). Figure 22 gives, for every condition, the mean Jaccard similarity with the other six conditions. Human speech is not an outlier: averaged over models, its value of 29.9 lies between XTTS-v2 (27.4) and F5-TTS (30.1), and the other TTS systems reach 30.5 to 32.3.

E.6 Text versus Speech Input

Table 30 compares text and speech input for Qwen3-Omni, the only model for which text-input results are currently available. The text input contains the same question, options, article, or dialogue, and the prompt differs only in the words that refer to the modality. Speech input lowers accuracy on Spoken Global MMLU by 5.5 points [4.9, 6.1] on average, from 0.3 points in English to 13.1 points in Japanese and 10.5 points in Korean, whereas the gap on Spoken CVQA is 0.9 points [0.1, 1.7]. For summarization, the difference in BERTScore is at most 0.12 points on either task.

Dataset	Language	n	Text	Speech	min / max	Speech maj.	Text − Speech
Spoken Global MMLU	English	1,000	86.9	$86.8_{\pm 0.7}$	85.8 / 88.0	87.2	+0.3 [-1.1, +1.6]
	Chinese	1,000	84.8	$79.3_{\pm 1.3}$	77.4 / 81.5	82.0	+5.8 [+4.2, +7.4]
	Japanese	1,000	79.3	$67.0_{\pm 1.3}$	65.1 / 68.5	70.2	+13.1 [+10.9, +15.5]
	Korean	1,000	76.2	$66.0_{\pm 1.9}$	63.0 / 69.6	70.7	+10.5 [+8.2, +12.8]
	Spanish	1,000	85.0	$80.8_{\pm 1.1}$	79.1 / 82.3	82.9	+4.2 [+2.7, +5.7]
	French	1,000	84.3	$81.4_{\pm 0.9}$	79.3 / 82.5	83.5	+3.1 [+1.6, +4.6]
	German	1,000	82.9	$79.4_{\pm 1.0}$	78.6 / 81.3	81.8	+3.6 [+2.0, +5.2]
	Italian	1,000	84.9	$80.9_{\pm 1.0}$	79.6 / 82.7	83.6	+4.0 [+2.3, +5.7]
	Portuguese	1,000	85.5	$80.7_{\pm 1.0}$	79.3 / 82.4	83.2	+4.8 [+3.2, +6.4]
	Hindi	1,000	76.9	$71.9_{\pm 1.5}$	68.9 / 73.9	75.4	+5.5 [+3.3, +7.6]
	All	10,000	82.7	$77.2_{\pm 1.1}$		80.1	+5.5 [+4.9, +6.1]
Spoken CVQA	English	3,559	75.0	$74.2_{\pm 0.3}$	73.5 / 74.6	76.7	+0.6 [-0.5, +1.7]
	Chinese	523	86.0	$83.4_{\pm 1.0}$	81.8 / 85.5	85.8	+2.7 [+0.3, +5.2]
	Japanese	203	71.9	$64.1_{\pm 1.9}$	61.4 / 67.0	68.9	+6.4 [+1.1, +11.2]
	Korean	290	79.3	$75.9_{\pm 1.7}$	73.8 / 78.6	79.6	+3.2 [-0.8, +7.4]
	Spanish	2,058	72.1	$71.2_{\pm 0.6}$	70.1 / 72.0	73.9	+0.6 [-0.9, +2.2]
	Portuguese	284	80.4	$81.1_{\pm 1.5}$	77.5 / 82.7	84.3	-1.3 [-5.4, +2.3]
	Hindi	201	82.7	$84.0_{\pm 1.2}$	81.6 / 86.0	87.3	-1.7 [-5.5, +2.3]
	All	7,118	75.5	$74.6_{\pm 1.2}$		77.0	+0.9 [+0.1, +1.7]
Spoken CrossSum	English	300	88.24	$88.04_{\pm 0.04}$	87.99 / 88.10		+0.20 [+0.12, +0.28]
	Chinese	300	88.81	$88.80_{\pm 0.04}$	88.75 / 88.86		+0.01 [-0.07, +0.09]
	Japanese	299	88.79	$88.47_{\pm 0.04}$	88.40 / 88.51		+0.33 [+0.21, +0.43]
	Korean	183	88.61	$88.58_{\pm 0.05}$	88.51 / 88.65		+0.04 [-0.07, +0.15]
	Spanish	300	88.21	$88.21_{\pm 0.04}$	88.16 / 88.30		+0.00 [-0.08, +0.08]
	French	191	88.53	$88.56_{\pm 0.04}$	88.51 / 88.64		-0.03 [-0.14, +0.08]
	Portuguese	300	88.24	$88.22_{\pm 0.04}$	88.12 / 88.26		+0.03 [-0.07, +0.13]
	Hindi	300	88.68	$88.38_{\pm 0.05}$	88.32 / 88.47		+0.30 [+0.19, +0.40]
	All	2,173	88.51	$88.39_{\pm 0.04}$			+0.12 [+0.09, +0.15]
Spoken SAMSum	English	300	90.98	$91.06_{\pm 0.04}$	90.99 / 91.14		-0.08 [-0.17, +0.01]

Table 30: Per-language text-versus-speech comparison for Qwen3-Omni (accuracy for MMLU / CVQA, BERTScore F1 for CrossSum / SAMSum, all in %, BERTScore to two decimals because speaker-to-speaker variation is below 0.1). Speech = $\text{mean}_{\pm \text{sd}}$ over the 10 speakers (25 mixes for SAMSum) with the worst / best speaker. Speech maj. = majority-vote accuracy over speakers; Text − Speech = paired item-level difference with 95% bootstrap CI. *All* pools the items of all languages.

E.7 Speaker Attributes

Gender. Figure 23 reports, for every evaluated model and language, the absolute gap between the five female and the five male evaluation speakers in performance and in consistency, with consistency computed within each gender group. Performance gaps are small, with a median absolute gap of 0.65 points on Spoken Global MMLU, 0.80 points on Spoken CVQA, and 0.02 BERTScore points on Spoken CrossSum. Consistency gaps are larger, with medians of 1.32, 1.81, and 1.15 points, and their intervals exclude zero in 31 of 90, 17 of 42, and 32 of 72 model–language pairs, respectively. Their direction is not uniform: female speakers yield higher consistency in 37 of 90 pairs on Spoken Global MMLU and 31 of 72 on Spoken CrossSum, but in only 7 of 42 on Spoken CVQA. Because each group contains only five speakers, a gap can reflect individual speakers rather than gender (§5.2). Spoken SAMSum is excluded because every dialogue contains both a female and a male voice.

Age, Acoustic Scene, and Nationality. Table 31 relates the annotated age group and acoustic scene of the evaluation speakers to model behavior, pooled over all models. For each speaker, difficulty is measured within items, as the difference between the speaker's outcome and the mean over the ten speakers of the same item, so that item difficulty cancels. No level differs reliably from zero: all intervals include zero, and the largest point estimates are below one point, for example -0.47 for elderly speakers and -0.28 for lecture-hall recordings on Spoken Global MMLU. Consistency among speakers of the same level is also within 1.3 points of consistency among all ten speakers. Elderly speakers are rare (three to four per task), and nationality is not analyzed because it almost coincides with language in our speaker pool, so these results should be read as the absence of large effects rather than as evidence of equality.

Task	Attribute	Level	Spk.	Lang.	Speaker difficulty	C within level vs. all 10
Spoken CVQA	Age	adult	56	7	−0.11 [−0.43, +0.17]	69.2 vs 69.4 (42 pairs)
Spoken CVQA	Age	elderly	3	3	+0.93 [−0.52, +3.30]	–
Spoken CVQA	Age	middle-aged	11	6	+0.28 [−0.09, +0.68]	71.5 vs 70.3 (18 pairs)
Spoken CrossSum	Age	adult	63	8	−0.00 [−0.01, +0.01]	–
Spoken CrossSum	Age	elderly	4	4	+0.01 [−0.02, +0.04]	–
Spoken CrossSum	Age	middle-aged	13	7	+0.00 [−0.01, +0.01]	–
Spoken Global MMLU	Age	adult	80	10	+0.02 [−0.19, +0.21]	61.9 vs 61.8 (90 pairs)
Spoken Global MMLU	Age	elderly	4	4	−0.47 [−1.48, +0.18]	–
Spoken Global MMLU	Age	middle-aged	16	9	+0.03 [−0.48, +0.47]	64.5 vs 64.2 (45 pairs)
Spoken CVQA	Acoustic scene	indoor studio	42	7	+0.03 [−0.22, +0.27]	69.6 vs 69.4 (42 pairs)
Spoken CVQA	Acoustic scene	lecture hall	13	7	−0.27 [−1.39, +0.50]	68.7 vs 70.0 (36 pairs)
Spoken CVQA	Acoustic scene	public scene	15	7	+0.16 [−0.43, +0.82]	70.1 vs 69.4 (42 pairs)
Spoken CrossSum	Acoustic scene	indoor studio	48	8	−0.00 [−0.01, +0.00]	–
Spoken CrossSum	Acoustic scene	lecture hall	15	8	+0.01 [−0.01, +0.02]	–
Spoken CrossSum	Acoustic scene	public scene	17	8	+0.00 [−0.01, +0.01]	–
Spoken Global MMLU	Acoustic scene	indoor studio	60	10	+0.14 [−0.06, +0.34]	62.1 vs 61.8 (90 pairs)
Spoken Global MMLU	Acoustic scene	lecture hall	19	10	−0.28 [−0.77, +0.16]	62.3 vs 62.8 (81 pairs)
Spoken Global MMLU	Acoustic scene	public scene	21	10	−0.16 [−0.64, +0.24]	62.1 vs 61.8 (90 pairs)

Table 31: **Speaker attributes.** Speaker difficulty is the speaker's outcome minus the mean over the ten speakers of the same item, averaged over items and pooled over models and languages, in points (accuracy, or BERTScore for Spoken CrossSum, negative means harder), with a 95% bootstrap interval over speakers. The last column compares consistency among the speakers of a level with consistency among all ten speakers, over model–language pairs with at least two speakers of that level. Groups are unbalanced and the comparison is descriptive.

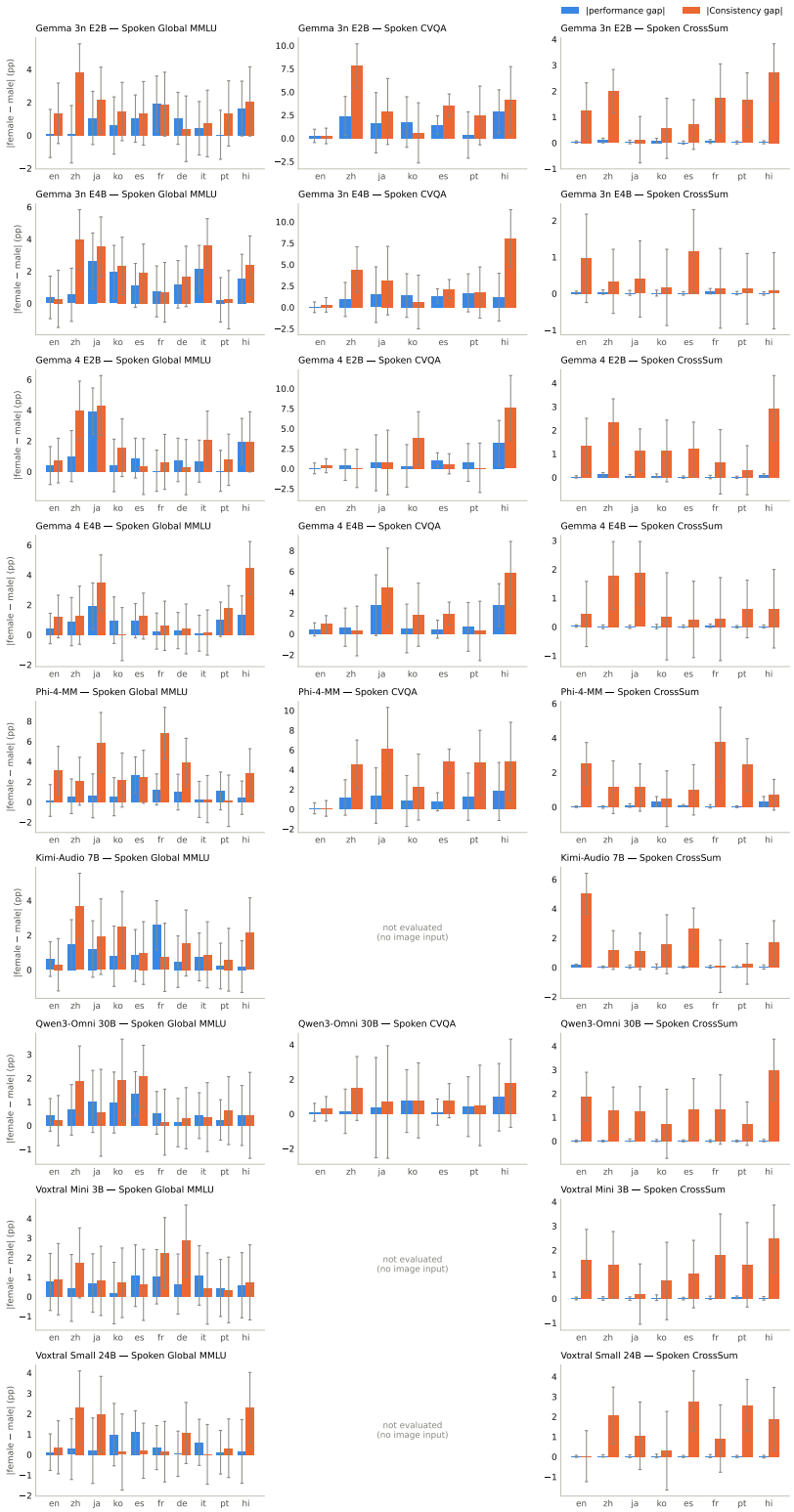

Figure 23: Absolute gender gaps in performance and consistency across languages for all evaluated models, one row per model. Columns are Spoken Global MMLU, Spoken CVQA, and Spoken CrossSum. The error bars are 95% bootstrap intervals.

E.8 Content Attributes

Culturally Sensitive versus Culturally Agnostic Questions. Table 32 compares culturally agnostic (CA) and culturally sensitive (CS) questions for every model and language of Spoken Global MMLU. The direction of the gap depends on the model. Qwen3-Omni answers CA questions more accurately and more consistently in every language, by 6.9 and 5.2 points on average, and both Gemma 4 models show smaller gaps in the same direction in every language except Chinese and Korean. Voxtral Mini shows the opposite pattern, with CS questions answered more accurately and more consistently in most languages, and the remaining models are close to zero on average. Over the 90 model–language pairs, 29 accuracy gaps and 35 consistency gaps have intervals that exclude zero, and the two gaps have the same sign in 77 pairs. A model that finds one question type harder therefore tends to be less consistent on it as well.

Model	En	Zh	Ja	Ko	Es	Fr	De	It	Pt	Hi	**Avg**
Accuracy gap (CA − CS, pp)											
Qwen3-Omni 30B	**5.1**	**6.9**	**5.8**	2.3	**7.8**	**6.1**	**9.7**	**8.8**	**9.9**	**6.8**	6.9
Voxtral Small 24B	1.1	**−7.9**	**−5.2**	−3.7	2.2	1.2	4.0	0.8	1.4	3.0	−0.3
Gemma 4 E4B	**6.8**	−3.1	2.1	−1.2	**5.4**	**6.6**	**6.9**	**6.7**	**6.1**	**6.4**	4.3
Gemma 4 E2B	**7.8**	−1.3	3.2	−1.7	**5.5**	**5.4**	**5.5**	**4.9**	**5.4**	**7.6**	4.2
Gemma 3n E4B	1.1	**−4.3**	−1.2	−2.3	1.5	1.4	2.8	−0.5	3.9	3.0	0.5
Gemma 3n E2B	−1.0	**−5.1**	−1.5	−1.1	0.1	0.6	0.4	−0.4	1.3	1.7	−0.5
Kimi-Audio 7B	0.2	2.6	0.0	−2.3	−0.8	−2.5	0.8	−1.1	−0.6	1.2	−0.3
Voxtral Mini 3B	−0.4	**−7.0**	−3.4	**−7.1**	−2.8	−3.8	−1.7	−2.2	−3.4	−2.6	−3.4
Phi-4-MM	3.0	2.9	3.7	−2.8	−0.7	−2.6	2.3	−1.1	−2.0	2.4	0.5
Consistency gap (CA − CS, pp)											
Qwen3-Omni 30B	**3.4**	**4.9**	0.9	1.7	**6.5**	**6.4**	**9.2**	**5.5**	**8.2**	**5.6**	5.2
Voxtral Small 24B	1.6	**−3.9**	−2.8	**−3.4**	**3.6**	1.4	**3.9**	1.9	2.6	**4.1**	0.9
Gemma 4 E4B	**3.4**	**−5.9**	−0.6	−1.7	**3.2**	**4.8**	**3.5**	**4.8**	**3.6**	**4.1**	1.9
Gemma 4 E2B	**4.8**	**−2.9**	−0.8	**−4.2**	**4.7**	2.8	**3.2**	**4.6**	**4.6**	2.2	1.9
Gemma 3n E4B	0.8	−0.6	−1.5	−0.3	−0.2	0.3	2.6	0.2	2.3	1.9	0.5
Gemma 3n E2B	3.0	−2.0	−0.8	1.2	1.4	0.4	1.9	1.2	2.2	2.0	1.0
Kimi-Audio 7B	0.0	−1.9	−0.8	**−4.6**	−0.2	−1.9	1.3	−3.1	−0.1	1.2	−1.0
Voxtral Mini 3B	0.0	−1.8	**−4.2**	**−4.0**	−2.2	**−5.2**	−0.7	**−5.3**	−2.7	−1.9	−2.8
Phi-4-MM	2.2	0.5	**7.0**	−0.8	−1.1	1.2	0.9	0.8	2.6	**−6.2**	0.7

Table 32: **Culturally agnostic versus culturally sensitive questions on Spoken Global MMLU.** Difference CA − CS in accuracy and consistency, in percentage points, positive values mean that culturally agnostic questions are answered more accurately or more consistently. Each language has 500 CA and 500 CS questions. Bold entries have a 95% bootstrap interval that excludes zero.

Subject Categories. Table 33 breaks Spoken Global MMLU down by subject category. Consistency ranges from 56.4% for *Other* to 64.4% for *Social Sciences*, but after accounting for item accuracy the residual differences are within ± 0.7 points, so the differences between categories mostly reflect question difficulty. Categories are also strongly confounded with cultural sensitivity, as only 2.7% of STEM questions are CS compared with 89.3% of Humanities questions. Among subjects with at least ten questions, residuals lie between -3.6 and $+2.4$ points, and the largest negative residuals occur in abstract algebra (-3.6), elementary mathematics (-3.6), and business ethics (-3.0).

Visual Grounding. Table 34 compares Spoken CVQA with Spoken Global MMLU on the seven languages covered by both tasks. Every vision-capable model is more consistent on Spoken CVQA, most strongly Phi-4-MM ($+19.9$ points), Gemma 3n E4B ($+15.2$), and Gemma 3n E2B ($+14.7$), and these three models are more consistent on Spoken CVQA in all seven languages. Their share of fragile items also drops, for example from 74.9% to 55.5% for Gemma 3n E4B. Qwen3-Omni, the most consistent model, changes by only $+0.7$ points. Because the two tasks differ in their questions and difficulty, this comparison is descriptive and does not isolate the effect of the image.

Category	CS share	Acc	C	C resid.	Fragile	Stable wrong
Other	84.5	48.2	56.4	−0.7	71.1	3.2
Humanities	89.3	51.9	60.1	0.2	67.0	3.9
STEM	2.7	56.4	62.0	−0.5	63.4	2.8
Business	50.0	56.4	62.9	−0.4	61.4	2.4
Medical	21.4	54.1	64.0	0.2	58.5	4.9
Social Sciences	73.5	57.3	64.4	0.7	61.6	3.7

Table 33: **Spoken Global MMLU by subject category**, pooled over models and languages (%). C resid. is consistency minus the consistency expected from item accuracy, estimated within each model over six accuracy bins. Fragile and stable-wrong shares follow the taxonomy of §4.2.

	Accuracy		Consistency				Fragile	
Model	MMLU	CVQA	MMLU	CVQA	Δ	Lang.	MMLU	CVQA
Qwen3-Omni 30B	76.1	76.3	81.5	82.2	0.7	4/7	37.5	36.2
Gemma 4 E4B	63.7	60.9	68.5	71.5	2.9	5/7	59.9	53.1
Gemma 4 E2B	51.6	52.0	58.3	64.0	5.8	5/7	71.4	63.7
Gemma 3n E4B	50.1	60.2	54.7	69.9	15.2	7/7	74.9	55.5
Gemma 3n E2B	42.6	53.0	48.1	62.8	14.7	7/7	79.2	64.0
Phi-4-MM	34.0	42.6	46.3	66.2	19.9	7/7	69.2	55.1

Table 34: **Spoken CVQA versus Spoken Global MMLU** on the seven languages covered by both tasks, averaged over languages (%). Δ is CVQA minus MMLU consistency, and Lang. counts the languages in which CVQA is more consistent.

F LIMITATION

Although 3MSPEECHBENCH covers multiple tasks, languages, and speakers, it remains a controlled benchmark rather than a complete reflection of real-world spoken interaction. In particular, the spoken inputs are synthesized through a voice-cloning pipeline, which allows us to isolate speaker-related variation more systematically, but may not capture the full range of phenomena found in naturally occurring speech, such as spontaneous disfluencies, turn-taking effects, or severe environmental distortions. In addition, our evaluation is limited to the selected datasets, languages, and model set considered in this study. The observed gaps in performance and consistency should therefore be interpreted as benchmark-based evidence of speaker-induced variability, rather than as a comprehensive characterization of spoken-language robustness in all deployment settings.