# PolySpeech-100: A Large-Scale Benchmark for Speech Understanding Across 100+ Languages and Dialects

Sicheng Yang
Shulan Ruan
Shiwei Wu
yangsc25@mails.tsinghua.edu.cn
slruan@sz.tsinghua.edu.cn
davidwu16@sz.tsinghua.edu.cn
Shenzhen International Graduate School,
Tsinghua University
Shenzhen, China

Yu Liu*
liuyu_thu@mail.tsinghua.edu.cn
Department of Electronic Engineering,
Tsinghua University
Beijing, China

Lu Fan
fanlu@jd.com
JD AI Research
Beijing, China

Zhi Li
You He
zhilizl@sz.tsinghua.edu.cn
heyou@mail.tsinghua.edu.cn
Shenzhen International Graduate School,
Tsinghua University
Shenzhen, China

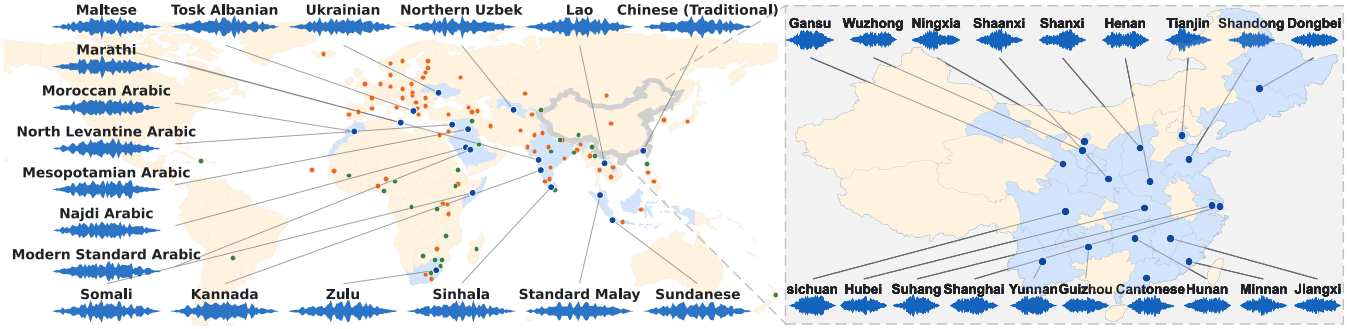



**Figure 1: Geographic distribution of PolySpeech-100. The points represent specific languages and dialects included in the benchmark. Orange shading indicates coverage of mainstream languages, while blue shading highlights regions covered by our newly added languages. The right panel highlights the common Chinese dialects included in the benchmark.**

## Abstract

While End-to-End (E2E) Speech-Large Language Models (Speech-LLMs) are rapidly evolving, their evaluation methodologies remain limited to the era of simple transcription. Existing benchmarks suffer from three critical limitations: a pronounced bias towards high-resource languages, a focus on low-level recognition (ASR) rather than semantic reasoning, and a neglect of regional dialects. To bridge this gap, we introduce PolySpeech-100, a massive-scale benchmark designed to assess 'native-level' speech comprehension across 110 linguistic variants. We employ a novel hybrid construction pipeline that augments gold-standard human recordings with instruction-driven synthetic speech, allowing us to cover 19 distinct Chinese dialects and over 80 low-resource languages. Extensive evaluation of 22 state-of-the-art models (including Gemini-3, GPT-Audio, and Qwen2.5-Omni) yields pivotal insights. First, we demonstrate that open-source E2E models outperform Cascade (ASR+LLM) systems on heavy dialects, proving that direct audio processing preserves critical paralinguistic cues and prosodic features (e.g., intonation, stress) that are often lost in standard transcription. Second, we reveal a significant performance gap: while commercial models maintain robustness, open-source models suffer catastrophic degradation on low-resource languages. Finally,

*Corresponding author.


*KDD '26, Jeju Island, Republic of Korea*

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817561

counter-intuitively, we observe that under standard zero-shot settings, Chain-of-Thought prompting frequently degrades speech understanding performance for most evaluated models, revealing a potential modality alignment gap in current architectures. PolySpeech-100 establishes a rigorous standard for the next generation of inclusive, omni-capable Speech-LLMs. The data, demo, and code are publicly available at https://github.com/YoungSeng/PolySpeech-100.

## CCS Concepts

• **Computing methodologies** → **Language resources**; **Neural networks**; *Speech recognition*; • **General and reference** → **Evaluation**; **Performance**.

## Keywords

Speech Large Language Models, Evaluation, Multilingual Benchmark, Dialect, Low-Resource Languages, Data Aggregation, Generalization

**ACM Reference Format:**

Sicheng Yang, Shulan Ruan, Shiwei Wu, Yu Liu, Lu Fan, Zhi Li, and You He. 2026. PolySpeech-100: A Large-Scale Benchmark for Speech Understanding Across 100+ Languages and Dialects. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3770855.3817561

## 1 Introduction

The field of Artificial Intelligence is undergoing a rapid transition from text-based Large Language Models (LLMs) to End-to-End (E2E) Speech-LLMs [41, 82]. Recent systems like ChatGPT [56] and Gemini [36] have demonstrated remarkable capabilities in perceiving and understanding directly with speech [97]. Unlike traditional cascaded systems, which first transcribe speech to text (ASR) and then process the text, these native multimodal models aim to capture paralinguistic cues, emotional tone, and dialectal nuances that are often lost in transcription [5, 16, 18].

Despite this progress, the evaluation infrastructure for Speech-LLMs lags significantly behind model development [24, 85]. Existing benchmarks suffer from three primary limitations. First, they are heavily **skewed towards high-resource languages**, predominantly English and Standard Mandarin, leaving the vast majority of the world's languages untested. Second, most benchmarks focus on **lower-level tasks** like Automatic Speech Recognition (ASR) or simple instruction following, rather than complex reasoning or semantic understanding. Third, and most critically, there is a lack of diverse **dialectal evaluation**. Current datasets often treat languages as monoliths (e.g., 'Chinese' or 'Arabic'), ignoring the rich regional variations—such as Cantonese, Sichuanese, or Maghrebi Arabic—that pose the greatest challenges to real-world deployment.

We introduce PolySpeech-100, a benchmark designed to shift speech evaluation from transcription ('hearing') to **content comprehension ('understanding') across 100+ languages and dialects**. Built upon the Belebele dataset [6, 22], this high-fidelity corpus challenges models to perform spoken question answering rather than relying on simple pattern matching. Notably, PolySpeech-100 emphasizes fine-grained dialectal diversity by incorporating 19 distinct Chinese regional variants (e.g., Minnan), allowing for a rigorous assessment of whether models possess true 'native-level' listening capabilities or merely overfit to standard broadcast speech.

To support this scale, we propose a hybrid data construction pipeline that effectively solves data scarcity issues for low-resource languages. Our human verification ($r = 0.83$ correlation) proves that high-quality synthesis is a valid proxy for evaluating dialectal robustness when human data is scarce. To the best of our knowledge, PolySpeech-100 represents the most linguistically diverse speech comprehension benchmark to date. It provides a comprehensive evaluation of current speech-centric models, establishing a new standard for the community to measure progress toward inclusive speech understanding. We evaluated 22 state-of-the-art models, including proprietary APIs (e.g., Gemini [36]) and open-source E2E models (e.g., MiMo-Audio [101], Step-Audio-2 [84]), as well as traditional ASR+LLM pipelines. Our experiments yield several critical insights: (1) For heavy dialects, open-source E2E models significantly outperform traditional ASR pipelines, acoustic features essential for semantic disambiguation that ASR systems typically filter out. (2) As for low-resource languages (e.g., Zulu, Lao), most open-source models degrade significantly while commercial models maintain robustness, highlighting a critical direction for future research. (3) In terms of reasoning strategies, we find that under our evaluated settings, many current Speech-LLMs struggle with Chain-of-Thought (CoT) in the audio modality, often performing better with direct answers than with intermediate reasoning steps.

## 2 Related Work

### 2.1 Traditional ASR and Translation Datasets

Early research focused on Automatic Speech Recognition (ASR) and Speech-to-Text Translation (S2TT). For English and multilingual ASR, LibriSpeech [58], GigaSpeech [12], and Multilingual LibriSpeech (MLS) [61] remain the standard. For Chinese, datasets like AISHELL [9, 31], WenetSpeech [100], and KeSpeech [73] cover Mandarin in diverse scenarios. Code-switching scenarios are specifically addressed in benchmarks like Mandarin-English Code-Switching Challenge dataset [68]. To evaluate multilingual capabilities, researchers use Common Voice [4], which includes many languages but focuses on short sentence reading. Fleurs [21] provides n-way parallel speech data for 102 languages, used for both ASR and translation tasks. Similarly, CoVoST 2 [79] offers large-scale speech translation benchmarks. Voxpopuli [78] provides European language data from political speeches. However, these datasets mainly evaluate transcription quality (WER or BLEU scores). They do not measure the semantic understanding or reasoning capabilities required by modern Large Language Models (LLMs).

### 2.2 Speech Understanding and Paralinguistics

Beyond transcription, speech contains rich information regarding emotion, environment, and speaker intent. Datasets like MELD [59] and IEMOCAP [10] focus on emotion recognition in dialogue. VocalSound [35] evaluates the classification of non-speech human sounds (e.g., laughter, sighing). In the general audio-text domain, WavCaps [54], Clotho [30], AudioCaps [42] and MusicCaps [29] test captioning capabilities for environmental sounds and music. More complex reasoning and environmental understanding are evaluated in MMAR [52], OmniBench [50], WorldSense [37], and DailyOmni

[105]. For dialogue understanding, the Fisher [20] dataset provides telephone conversation data, which is often used to test reference resolution and topic modeling. SpokenWOZ [70], MultiChallenge Audio [27] and SLURP [8] are designed for dialogue state tracking and spoken language understanding (SLU), providing multi-turn conversation data. While valuable, these tasks are often specific classifications (e.g., 'happy' vs. 'sad') or short-context descriptions rather than chat or reasoning across diverse languages.

## 2.3 Benchmarks for Speech-LLM

With the rise of Speech-LLMs, comprehensive benchmarks have emerged. AIR-Bench [95] is a comprehensive benchmark covering speech, sound, and music. It uses datasets like Fisher [20] and SpokenWOZ [70]. Dynamic-SUPERB [38] evaluates instruction-following abilities across various tasks. Recent suites like OpenAudioBench [1], UltraEval-Audio [67], URO-Bench [92], and Big Bench Audio [2] evaluate instruction following and reasoning. Instruction-tuning evaluations include InstructS2S [33], Moss [72], and AlpacaEval [49]. VoiceBench [17] focuses on safety, instruction following, and reasoning (QA). It adapts text benchmarks (e.g., AlpacaEval) into speech. MMAU [65] and MMSU [80] test multi-task understanding and reasoning, including math and logic in audio.

Many recent models propose their own internal evaluation sets. Qwen2-Audio [19], Qwen2.5/3-Omni [88, 89], and Qwen3.5-Omni [75] test on massive multi-task collections. LLaMA-Omni [33] and LLaMA-Omni 2 [34] focus on low-latency speech interaction, evaluating content and style. Mini-Omni [87] and Moshi [26] explore real-time streaming capabilities. SLAM-LLM [53] investigates speech encoders with datasets like GigaSpeech 2 [96]. Other notable models include Fun-Audio-Chat [76], MinMo [15], Step-Audio 2 [84], MiMo-Audio [101], Ming-Omni [1], Baichuan-Audio [48], Kimi-Audio [44] and X-Talk [51]. Most of these works evaluate English and Mandarin well. However, they lack comprehensive testing for low-resource languages and regional dialects (e.g., Sichuan, Cantonese) in a complex reasoning context.

## 2.4 Synthetic Data for Evaluation

Collecting human recordings for rare dialects is expensive and time-consuming [28, 55]. A growing trend in recent model evaluations (e.g., for LLaMA-Omni2 [34], Mini-Omni [87], and MiMo-Audio [101]) is to convert text-based benchmarks [33, 72] [3,4] into audio using Text-to-Speech (TTS) systems. However, these TTS-based evaluations often lack real-world noise and linguistic variations ('clean' data issue). In contrast, 2M-BELEBELE [22] serves as a gold standard for multilingual reading comprehension with high-quality human recordings, though its coverage of specific dialectal nuances and code-switching remains limited. To address these gaps, we propose PolySpeech-100, a large-scale hybrid benchmark covering 100+ languages and dialects. By combining human-recorded gold standards with diverse synthesized data covering dialects and long-tail languages, we provide a robust assessment of Speech-LLMs' 'Babel' capabilities in complex, real-world linguistic scenarios.

[1]https://huggingface.co/datasets/baichuan-inc/OpenAudioBench
[2]https://huggingface.co/datasets/ArtificialAnalysis/big_bench_audio
[3]https://huggingface.co/datasets/XiaomiMiMo/SpeechMMLU
[4]https://huggingface.co/collections/mistralai/speech-evals

# 3 Dataset Creation

We introduce PolySpeech-100, a scalable and linguistically diverse speech understanding benchmark covering over 100 language variants. To address the scarcity of evaluation data for dialects and low-resource languages [7, 90], we propose a hybrid construction pipeline that augments human recordings with state-of-the-art synthetic generation. Our methodology follows a three-stage framework: (1) Multi-Source Data Aggregation [7, 45], (2) Generative Audio Synthesis, and (3) Multi-Level Quality Assurance.

## 3.1 Data Foundation: The Belebele Backbone

To ensure rigorous cross-lingual comparability, we utilize the Belebele benchmark [6] as our textual foundation. Belebele offers parallel reading comprehension passages and multiple-choice questions (MCQs) aligned across 122 languages. This parallel structure is ideal for speech understanding as it allows for controlled evaluation of acoustic modeling capabilities across diverse language families without semantic variation.

## 3.2 Multi-Source Audio Construction

Our dataset construction employs a stratified approach to audio acquisition, balancing authenticity with coverage. We categorize our data sources into three distinct tracks.

*3.2.1 Track 1: Human-Recorded Speech (High-Resource).* For the core set of 73 languages, we leverage the 2M-BELEBELE corpus [22]. This component provides professionally recorded human speech, serving as the gold standard for acoustic naturalness. We parse the source data to extract aligned tuples of (Passage, Question, Options), ensuring that high-resource languages (e.g., English, Spanish, French) are represented by authentic native speaker recordings.

*3.2.2 Track 2: Generative Dialect Adaptation.* A critical limitation of existing speech benchmarks is the neglect of regional dialects. To bridge this gap, we propose a high-fidelity generation pipeline that challenges the view that synthetic data is unsuitable for evaluation [22]. While prior studies relied on concatenation-based systems, we utilize CosyVoice 3.0 [32], a state-of-the-art generative speech model with zero-shot instruction following. Unlike traditional TTS [39, 60, 99] which often flattens prosody, CosyVoice enables precise control over accent and intonation via natural language prompts.

We implement a two-stage Rewrite-then-Synthesize strategy to bridge the 'long-tail dialect gap': (1) *Lexical Adaptation:* We employ a LLM (Qwen3-Instruct [93]) to structurally rewrite standard text into dialectal vernacular (e.g., converting Mandarin vocabulary to distinct Cantonese or Northeastern lexical forms) while preserving semantic meaning. (2) *Instruction-Aware Synthesis:* The rewritten text is processed by CosyVoice3 with dialect-specific instruction prompts, which injects specific dialectal characteristics—such as the tonal variations of Cantonese or the retroflexion of Northern Mandarin—into the text. We generate fine-grained dialectal speech for 19 distinct Chinese regional variants, covering major linguistic groups (e.g., Cantonese, Wu, Minnan) and Mandarin accents (e.g., Sichuan, Dongbei, Tianjin). This approach yields highly expressive audio that preserves the semantic content of the original text while introducing realistic phonological shifts characteristic of regional speakers. Our experiments (analyzed in Section 4) demonstrate that

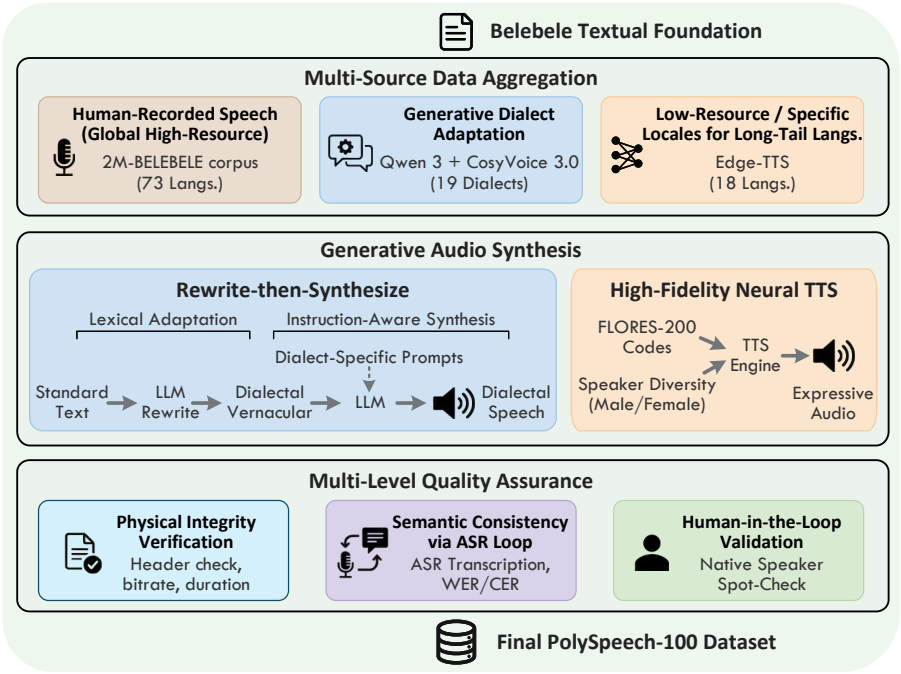



**Figure 2: The data construction pipeline of PolySpeech-100. The framework consists of three stages: multi-source data aggregation, generative audio synthesis via a Rewrite-then-Synthesize strategy, and multi-level quality assurance.**

this generates data with sufficient discriminative power to serve as a valid proxy for human speech, bridging the data scarcity gap. Please refer to our project repository for audio samples.

*3.2.3 Track 3: Neural Synthesis for Long-Tail Languages.* For low-resource languages where human data is inaccessible, we utilize high-fidelity Neural TTS. This approach is grounded in two key premises: First, TTS augmentation is essential for covering the linguistic long tail, a standard practice in recent multilingual evaluations [7, 40, 43]. Second, subjective evaluations indicate that modern neural synthesis [5] achieves MOS (Mean Opinion Score) parity with human recordings in clean settings [3, 98, 104]. Furthermore, this synthetic approach offers superior controllability; it allows us to systematically inject environmental noise and reverberation to assess model robustness, a factor we analyze extensively in the experimental section. To extend coverage to under-represented languages (e.g., Zulu, Maltese, Lao, and various Arabic dialects), we employ high-fidelity neural text-to-speech (TTS) via the Edge-TTS engine [6]. We developed a rigorous mapping protocol to align FLORES-200 language codes with available neural voice locales. To mitigate speaker overfitting and improve model robustness, we enforce speaker diversity by randomizing voice profiles (Male/Female) across different samples [66].

[5]https://github.com/resemble-ai/chatterbox
[6]https://github.com/rany2/edge-tts

## 3.3 Quality Assurance Protocol

To ensure the reliability of PolySpeech-100, we implemented a rigorous three-step validation protocol: (1) *Physical Integrity Verification*: We perform automated scanning of the generated corpus to detect file corruption. This includes checking for valid header structures, appropriate bit rates, and duration thresholds to filter out truncated or silent files. The dataset structure is normalized to ensure every sample contains a complete Q&A chain (Passage, Question, and four Options). (2) *Semantic Consistency via ASR Loop [25, 47]*: To guarantee that the synthetic speech remains intelligible and faithful to the source text, we introduce an Automatic Speech Recognition (ASR) verification loop. To be specific, we utilized several tools (e.g., Qwen3-ASR [69], SenseVoice [2], Whisper [62], TeleASR [13]) to transcribe synthetic audio back to text. We compute the Word Error Rate (WER) and Character Error Rate (CER) between the source text and the ASR transcription. Samples exceeding a strict error threshold are flagged as semantically distinct (indicating synthesis failure or severe hallucination) and are automatically regenerated or excluded. This ensures that PolySpeech-100 measures speech understanding capabilities rather than robustness to poor audio generation. (3) *Human-in-the-Loop Validation*: To complement automated metrics, we conducted a manual spot-check on the dialectal samples [64]. Native speakers verified the authenticity of prosody and lexical usage, ensuring the dataset meets the quality standards required for academic benchmarking. Besides, we compared model performance on our data against real human recordings, observing a strong correlation ($r = 0.83$), details are provided in Appendix.

**Table 1: Summary of Performance on PolySpeech-100. We categorize the languages into 3 groups: High-Res (10 major languages, e.g., EN, ZH), CN-Dialect (19 Chinese regional dialects), and Low-Res (81 long-tail languages). Detailed definitions and full results are provided in Appendix C. Bold denotes the best result, and underline denotes the second-best result. Note the significant gap between the SOTA closed-source model and others in Low-Res scenarios.**

| Model | | Overall | High-Res | CN-Dialect | Low-Res |
|---|---|---|---|---|---|
| *Closed-Source Models* | | | | | |
| Gemini-3-flash [36] | | **85.30** | **94.26** | **83.54** | **84.61** |
| GPT-Audio-mini [57] | | <u>56.63</u> | 83.56 | 55.58 | <u>53.56</u> |
| *Open-Source E2E Models: Speech+Text → Text* | | | | | |
| Fun-Audio-Chat [76] | | 52.88 | 84.82 | 77.06 | 43.26 |
| Qwen2.5-Omni [88] | | 50.89 | <u>84.94</u> | <u>78.61</u> | 40.18 |
| MiniCPM-o 4.5 [23] | | 50.57 | 78.25 | 59.24 | 45.12 |
| Audio-Omni [77] | | 46.62 | 77.50 | 71.95 | 36.86 |
| MiMo-Audio [101] | | 43.51 | 61.09 | 76.03 | 33.72 |
| Step-Audio-2 [84] | | 40.52 | 60.44 | 68.17 | 31.57 |
| Qwen2-Audio [19] | | 25.94 | 26.59 | 25.40 | 25.99 |
| *Open-Source E2E Models: Speech → Text* | | | | | |
| Covo-Audio [83] | | 24.78 | 26.68 | 18.75 | 25.96 |
| PersonaPlex [63] | | 23.67 | 24.25 | 22.43 | 23.89 |
| Mini-Omni [87] | | 21.83 | 17.64 | 22.91 | 22.10 |
| LLaMA-Omni2 [33] | | 21.88 | 22.87 | 20.26 | 22.14 |
| Moshi [26] | | 20.96 | 25.46 | 17.78 | 21.15 |
| *Cascaded Pipelines (ASR+LLM)* | | | | | |
| Whisper-v3 [62] | + Qwen2.5 [94] | 53.86 | 83.74 | 62.62 | 48.12 |
| | + Qwen3 [93] | 51.56 | 80.17 | 59.82 | 46.09 |
| | + Llama-3.1 [74] | 43.59 | 62.74 | 46.86 | 40.46 |
| | + Llama-3.2 [74] | 39.01 | 59.20 | 44.98 | 35.12 |
| Qwen3-ASR [69] | + Qwen3 [93] | 52.66 | 84.28 | 73.93 | 43.77 |
| | + Qwen2.5 [94] | 52.29 | 81.80 | 72.36 | 43.94 |
| | + Llama-3.1 [74] | 45.00 | 66.81 | 58.31 | 39.18 |
| | + Llama-3.2 [74] | 38.21 | 57.04 | 49.36 | 33.27 |

## 4 Benchmark Experiments

In this section, we evaluate 22 state-of-the-art audio systems on PolySpeech-100. We analyze their performance across high-resource languages, generated dialects, and long-tail languages.

### 4.1 Experimental Setup

We categorize the evaluated models into two primary groups. The first group consists of End-to-End (E2E) models that process audio inputs directly without requiring intermediate text transcription. Within this category, we evaluated closed-source models such as `Gemini-3.0-Flash-Preview` [36] and `gpt-audio-mini-2025-12-15` [57]. We also selected eight open-source models, including `Qwen2.5-Omni-7B` [88], `Fun-Audio-Chat-8B` [76], `Step-Audio-2-Mini` [84], `MiMo-Audio-7B` [101], `Qwen2-Audio-7B-Instruct` [19], `LLaMA-Omni2-7B-Bilingual` [34], `Mini-Omni` [87], and `Moshi` [26].

The second group comprises Cascade Systems, serving as baselines that combine ASR modules with LLM. For the ASR frontend, we utilized `Whisper-Large-v3` [62] and `Qwen3-ASR-1.7B` [69]. These are paired with LLM backends as `Qwen3-4B-Instruct-2507` [93], `Qwen2.5-7B-Instruct` [94], `Llama-3.2-3B-Instruct` and `Llama-3.1-8B-Instruct` [74].

Regarding the prompting strategy, we adapted our approach based on the specific capabilities of each system. For models that support system instructions (e.g., Qwen2.5-Omni, Gemini), we utilized a text system prompt to define the task. However, for models that do not support system instructions—specifically (e.g., `LLaMA-Omni2`, `Moshi`, `Mini-Omni`), we converted the instruction into speech using TTS and concatenated it to the beginning of the input audio. The exact prompts used are provided in the Appendix.

### 4.2 Main Results

*4.2.1 Overall Performance Analysis.* As shown in Table 1, in the general evaluation across all 100+ languages, Gemini-3-flash achieves state-of-the-art performance, achieving an overall accuracy of 85.30% across all languages. It maintains consistent performance across both high-resource and low-resource settings, demonstrating remarkable robustness. In contrast, OpenAI's GPT-Audio-mini lags significantly with an overall score of 56.63%, placing it in a similar tier to the best open-source models rather than leading the field.

Among open-source E2E models, Fun-Audio-Chat (52.88%) and Qwen2.5-Omni (50.89%) emerge as the top performers. Their performance is highly competitive with, and in some aspects surpasses, the strong baseline of cascaded pipelines (e.g., Whisper-v3 + Qwen2.5 at 53.86%). However, models that rely on 'Prompt-to-Speech' concatenation due to a lack of system prompt support (specifically Mini-Omni, LLaMA-Omni2, and Moshi) collapsed catastrophically. Their scores hovered around 20-22%, which is below the random guess threshold (25%) for four-choice questions. This finding strongly suggests that current audio-language models struggle to distinguish instruction from content when they are mixed in the same audio stream without textual guidance.

*4.2.2 Performance on Dialects.* The evaluation on 19 Chinese regional dialects (and 6 Arabic variants, see Appendix C.7 for a detailed case study) reveals striking insights. While Gemini-3-flash remains the top scorer (83.54%), open-source E2E models exhibit a remarkable "native understanding" advantage over traditional pipelines and even GPT-Audio-mini. Specifically, Qwen2.5-Omni and Fun-Audio-Chat achieved accuracy scores of 78.61% and 77.06%, respectively. This performance significantly outperforms GPT-Audio-mini (55.58%) by a margin of over 20 percentage points. Furthermore, these E2E models surpass the classic Whisper-v3 + Qwen2.5 pipeline (62.62%). This indicates that general-purpose ASR systems like Whisper often lose critical semantic information when transcribing heavily accented dialects into text. Although using a specialized ASR (Qwen3-ASR) improves pipeline performance to 73.93%, the best E2E model (Qwen2.5-Omni) still holds the lead. This empirically validates the hypothesis that E2E architectures can leverage paralinguistic cues and latent representations to bridge the gap in dialectal understanding better than text-mediated systems. For a detailed case analysis, please refer to Appendix C.8.

Besides, we observe a clear hierarchy in model performance based on linguistic distance. As shown in Fig. 3, on Standard Mandarin (`zho_Hans`) and Northern dialects such as `Tianjin` and `Dongbei`, models like Qwen2.5-Omni achieve greater than 85% accuracy. Dialects with distinct tones but shared vocabulary, like `Sichuan`, accuracy remains robust at approximately 80%. However, performance degrades significantly on distinct dialect families like `Shanghai` (Wu) and `Cantonese` (Yue). For example, Qwen2.5-Omni drops

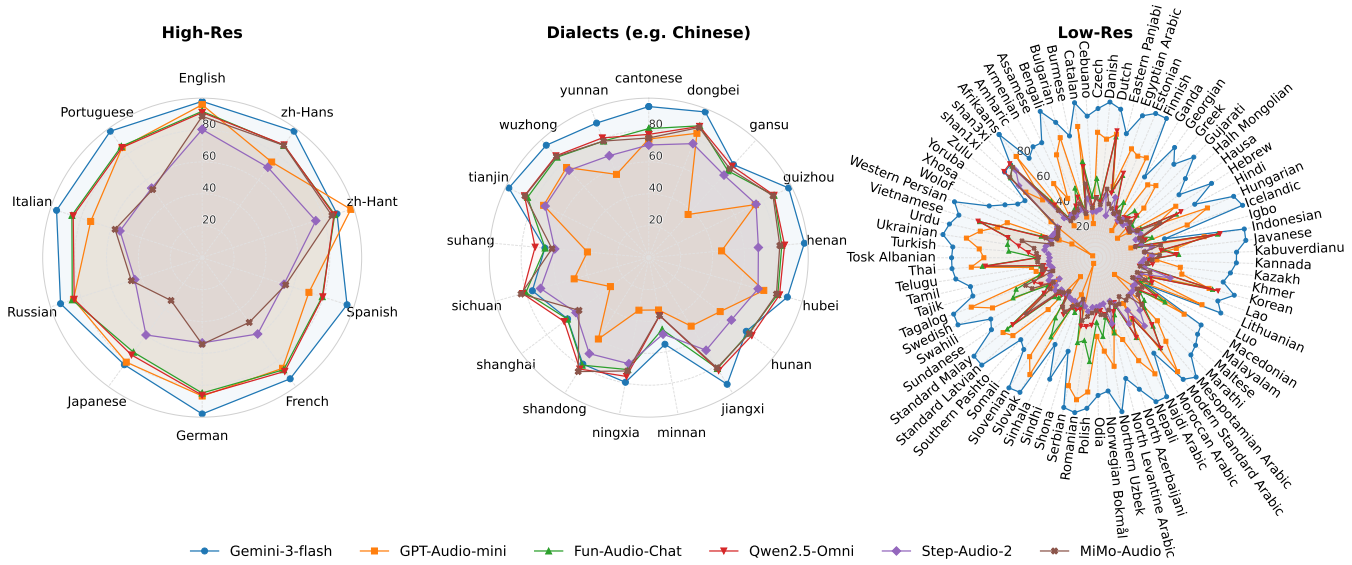



**Figure 3: Performance benchmarking of various audio models across diverse linguistic scenarios. The radar charts illustrate comparisons in High-Resource languages (left), Chinese Dialects (middle), and Low-Resource languages (right). The visualization highlights the capabilities of models such as Gemini-3-flash and Fun-Audio-Chat, demonstrating performance variances particularly in dialectal and low-resource contexts.**

from 87.4% (Mandarin) to 66.4% (Shanghai), validating our synthesized data in challenging the models. The gap in Arabic varieties is even more severe. While models perform well on Modern Standard Arabic (MSA), they struggle with regional variants; Fun-Audio-Chat scores 69.7% on MSA but drops to 45.1% on Moroccan Arabic. This highlights a standard language bias in current training datasets, where dialects are underrepresented compared to formal speech.

*4.2.3 Robustness on Low-Resource Languages.* The 'Low-Res' category, comprising 81 long-tail languages, proved to be the most challenging testbed. We observed a universal performance degradation, but the severity varied drastically across models. Gemini-3-flash exhibited exceptional robustness, maintaining a high accuracy of 84.61%, which is nearly identical to its high-resource performance.

In contrast, all other models experienced a sharp decline. GPT-Audio-mini dropped to 53.56%, while the best open-source E2E models fell into the 30-43% range (e.g., Fun-Audio-Chat at 43.26%). Interestingly, in this specific domain, the Whisper-v3 + Qwen2.5 pipeline (48.12%) slightly outperformed the best open-source E2E models. This suggests that for ultra-low-resource languages, the massive multilingual pre-training of Whisper's encoder currently provides a more stable foundation than the audio encoders used in current open-source E2E models. Nevertheless, the substantial gap between Gemini and all other models highlights that low-resource speech understanding remains the primary frontier for the open-source community to conquer.

*Efficiency Analysis.* Beyond accuracy, we analyzed the computational cost by standardizing the total inference time for the PolySpeech-100 benchmark (88,000 samples). Fun-Audio-Chat is the most efficient model (11 hours). This high speed makes it ideal for real-time applications where latency is critical. LLaMA-Omni2 (15.8 hours), MiMo-Audio (18.2 hours), and Qwen2.5-Omni (21.3 hours) show moderate efficiency. These architectures, while powerful, result in higher costs due to their streaming mechanisms. Step-Audio-2 takes slightly longer at 30.0 hours. This suggests potential optimization bottlenecks in its inference implementation or a heavier encoder architecture. In contrast, Mini-Omni, Qwen2-Audio and Moshi are significantly slower (109.5, 141.0, and 200+ hours respectively). Although these three models possess chat capabilities, the long duration of each input audio segment likely causes the complexity to increase exponentially, causing high latency.

## 4.3 Robustness Analysis

Real-world speech interaction rarely occurs in studio-quality conditions. To evaluate the resilience of Speech-LLMs against acoustic distortions and their operational efficiency, we conducted a stress test on PolySpeech-100. We selected four representative open-source models—*Fun-Audio-Chat*, *MiMo-Audio*, *Qwen2.5-Omni*, and *Step-Audio-2*—and evaluated them under five acoustic conditions across three language groups. We implemented an automated augmentation pipeline. The original recordings serve as the clean baseline. To simulate real-world interference, we add Gaussian white noise at two distinct levels: a low setting with a signal-to-noise ratio (SNR) of approximately 20dB to mimic a quiet office, and a high setting with roughly 0-5dB SNR to represent a noisy street environment (using peak normalization). Additionally, we adjust the audio speed to create temporal variations. This includes a slow version at 0.8x the original speed and a fast version at 1.2x, which modifies both the tempo and pitch of the input. Besides, we recorded the total wall-clock time required to complete the full benchmark (all 100+

**Table 2: Comprehensive Performance Analysis with Heatmap. We report accuracy (%) and the absolute performance drop/gain (in parentheses) relative to the Base condition. The cell background color indicates the magnitude and direction of the change: Red shades denote performance degradation, while Green shades denote improvement. Darker shades indicate larger changes.**

| Group | Model | Base | Noise Robustness | | Speed Robustness | | Strategy Impact | |
|---|---|---|---|---|---|---|---|---|
| | | (Zero-Shot) | Noise-Low | Noise-High | Speed-Slow | Speed-Fast | CoT Prompting | 3-Shot Context |
| **High-Res** | Fun-Audio-Chat [76] | 84.82 | 81.10 (-3.7) | 65.73 (-19.1) | 81.35 (-3.5) | 82.24 (-2.6) | 75.36 (-9.46) | 82.69 (-2.13) |
| | MiMo-Audio [101] | 61.09 | 58.46 (-2.6) | 49.29 (-11.8) | 61.10 (+0.0) | 57.88 (-3.2) | 50.11 (-10.98) | 54.04 (-7.05) |
| | Qwen2.5-Omni [88] | 84.94 | **83.96 (-1.0)** | **75.90 (-9.0)** | **83.67 (-1.3)** | **85.15 (+0.2)** | 74.06 (-10.88) | 86.64 (+1.70) |
| | Step-Audio-2 [84] | 60.44 | 61.09 (+0.7) | 47.37 (-13.1) | 59.49 (-1.0) | 59.61 (-0.8) | 67.54 (+7.10) | 45.39 (-15.05) |
| **CN-Dialect** | Fun-Audio-Chat [76] | 77.06 | 73.88 (-3.2) | 66.37 (-10.7) | 73.65 (-3.4) | 68.27 (-8.8) | 77.14 (+0.08) | 76.55 (-0.51) |
| | MiMo-Audio [101] | 76.03 | 75.96 (-0.1) | 68.03 (-8.0) | 75.73 (-0.3) | 74.12 (-1.9) | 76.32 (+0.29) | 75.92 (-0.11) |
| | Qwen2.5-Omni [88] | 78.61 | **77.76 (-0.9)** | **70.50 (-8.1)** | **75.89 (-2.7)** | **77.17 (-1.4)** | 68.32 (-10.29) | 77.82 (-0.79) |
| | Step-Audio-2 [84] | 68.17 | 64.13 (-4.0) | 55.62 (-12.6) | 66.65 (-1.5) | 66.05 (-2.1) | 66.72 (-1.45) | 59.53 (-8.64) |
| **Low-Res** | Fun-Audio-Chat [76] | 43.26 | 37.51 (-5.8) | 30.95 (-12.3) | 38.24 (-5.0) | 38.09 (-5.2) | 39.06 (-4.20) | 41.52 (-1.74) |
| | MiMo-Audio [101] | 33.72 | 30.92 (-2.8) | 28.10 (-5.6) | 31.40 (-2.3) | 30.61 (-3.1) | 31.75 (-1.97) | 30.29 (-3.43) |
| | Qwen2.5-Omni [88] | 40.18 | 37.08 (-3.1) | 32.97 (-7.2) | 34.69 (-5.5) | 36.53 (-3.7) | 33.37 (-6.81) | 40.57 (+0.39) |
| | Step-Audio-2 [84] | 31.57 | 29.51 (-2.1) | 28.72 (-2.9) | 28.99 (-2.6) | 29.27 (-2.3) | 31.54 (-0.03) | 22.88 (-8.69) |

languages) on a single NVIDIA 4090 GPU. This provides a direct measure of inference throughput.

Table 2 presents the performance stability across different language groups. Values in parentheses indicate the absolute accuracy drop compared to the *Base* (Clean) condition.

*4.3.1 Sensitivity to Environmental Noise.* Table 2 shows, High-Res languages suffer significant degradation under high noise. For instance, *Fun-Audio-Chat* drops by 19.1% in the High-Res group. In contrast, Qwen2.5-Omni demonstrates superior resilience, dropping only 9.0%. This suggests that the streaming-oriented pre-training of Qwen2.5-Omni likely included large-scale noisy data, making it more robust to acoustic interference. Interestingly, the performance drop on Dialects (e.g., Qwen2.5-Omni drop of 8.1%) is comparable to High-Res languages. This indicates that the dialectal features learned by these models are robust and not merely fragile overfitting to clean TTS data. In the Low-Res group, absolute performance is already low (~30-40%). Consequently, the impact of noise appears numerically smaller (e.g., Step-Audio-2 drops only 2.9%), but this is largely due to the "floor effect"—performance cannot drop much further below random guessing (25%). However, *Fun-Audio-Chat* still loses over 12% accuracy, highlighting that its acoustic encoder is less stable on unfamiliar phonemes when noise is present.

*4.3.2 Sensitivity to Speaking Rate.* Models generally handled speed variations better than noise. However, for *Fun-Audio-Chat*, the fast speed setting caused a notable drop in dialect performance (-8.8%), suggesting that temporal compression makes accent recognition significantly harder for some architectures.

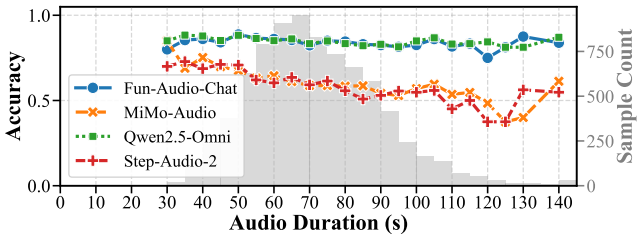
*4.3.3 Analysis of Audio Duration Robustness.* To evaluate the robustness of Speech-LLMs across different input lengths, we analyze model performance with respect to audio duration as shown in Figure 4. While a general negative correlation exists between audio duration and model accuracy, high-resource language scenarios remain largely stable; specifically, Fun-Audio-Chat and Qwen2.5-Omni exhibit remarkable stability, maintaining high accuracy (near $0.8 - 0.9$) even as the duration extends to 140 seconds, whereas MiMo-Audio and Step-Audio-2 show significant drops for longer segments. Conversely, in long-tail languages, although overall accuracy is significantly lower, Fun-Audio-Chat consistently achieves leading performance across almost all intervals. It is worth noting that Chinese dialects generally exhibit shorter durations compared to other languages, and the apparent accuracy fluctuations in the 120–140s interval are primarily artifacts of limited sample sizes.

*4.3.4 Analysis of Prediction Bias.* To investigate model reliability, we analyze the prediction bias and prompt sensitivity as illustrated in Figure 5. We observe distinct behaviors among the models: while *Fun-Audio-Chat*, *Qwen2.5-Omni*, and *MiMo-Audio* demonstrate robust performance with selection distributions aligning closely with the ground truth, the *Step-Audio-2* variants exhibit a severe preference bias. Specifically, *Step-Audio-2* disproportionately selects option B, a trend that remains consistent regardless of the system prompt suffix applied (refer to the Appendix for specific prompts and modifications). This persistence indicates an inherent model bias rather than simple prompt sensitivity, contrasting sharply with the balanced distribution observed in the other models.

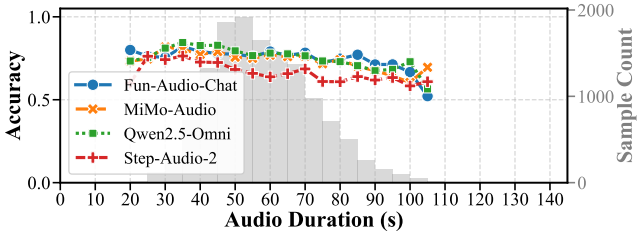
## 4.4 Advanced Reasoning Capabilities

To investigate the advanced reasoning potential of Speech-LLMs beyond standard zero-shot accuracy, we employed two paradigms: Chain-of-Thought (CoT) prompting, where system prompts were modified to explicitly request reasoning steps prior to the final answer, and Few-Shot (In-Context) Learning, which involved prepending a 3-turn audio segment containing [Passage, Question, Answer] examples to the input; the impact of these techniques across the three language groups is summarized in Table 2.

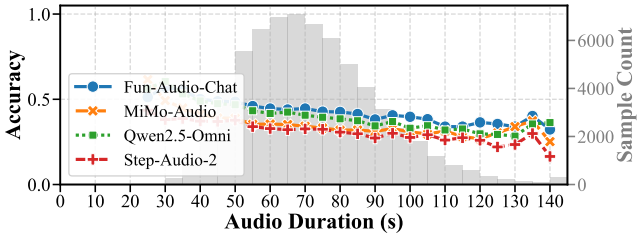
*4.4.1 Divergent Impact of Chain-of-Thought on Speech-LLMs.* Contrary to the consistent positive effect typically observed in text-based LLMs, our zero-shot experiments indicate that Chain-of-Thought prompting frequently exerted a negative impact on several current Speech-LLMs. Specifically, models such as *Qwen2.5-Omni* and *Fun-Audio-Chat* suffered a performance degradation of approximately 10% in High-Resource languages. We hypothesize this stems




(a) High-Resource Languages




(b) Chinese Dialects




(c) Long-Tail Languages

**Figure 4: Performance analysis of 4 Speech-LLMs across varying audio durations. The colored line charts (left axis) illustrate the accuracy of each model, while the background gray histograms (right axis) indicate the number of test samples per duration interval (minimum samples threshold=15, cutoff=140s). This highlights the correlation between data sparsity (indicated by lower histogram bars) and model performance degradation in long-form audio segments.**

from reasoning hallucinations, where the model is misled during the inference phase, and we refer readers to the Appendix C.5 for a detailed qualitative analysis of these failure cases. A notable exception was *Step-Audio-2*, which achieved a significant 7.1% improvement in the same category, suggesting that its unique alignment training likely incorporated audio reasoning data enabling effective CoT reasoning. To verify that this degradation is not merely an artifact of a specific prompt design, we conducted an ablation study using multiple alternative CoT templates (detailed in Appendix C.6). The consistent performance drops observed across different prompt structures confirm that this is a fundamental modality alignment issue rather than prompt sensitivity.

*4.4.2 Ineffectiveness of Audio In-Context Learning.* The 3-Shot experiments highlighted limitations in current audio context processing, as most models failed to leverage the prepended examples. Specifically, *Step-Audio-2* suffered a massive drop (e.g., -15.05% in High-Resource languages), indicating that the long context likely exceeded its effective audio attention span and caused the model to 'forget' the actual question; conversely, while *Qwen2.5-Omni* showed slight positive transfer in High-Resource (+1.7%) and Low-Resource (+0.39%) settings—validating its stronger backbone—these gains were negligible compared to the computational cost of processing the extra audio tokens.

## 5 Discussion

The evaluation results from PolySpeech-100 reveal a fundamental limitation in current cascade systems (ASR + LLM), the information bottleneck created by converting speech into intermediate text. Our experiments demonstrate that text serves as a lossy compression of speech, where standard ASR normalization inadvertently strips away paralinguistic cues—such as tone, stress, and regional nuances—that are critical for semantic disambiguation in heavy dialects. In contrast, End-to-End models bypass this transcription phase by leveraging continuous latent representations, thereby retaining the acoustic fidelity necessary for complex reasoning. This evidence suggests that to achieve truly inclusive speech understanding, particularly for dialectal speakers and strictly oral languages (e.g., unwritten languages), the field must transition from a transcribe-then-read paradigm to native acoustic reasoning.

Unlike the consistent gains seen in text-based models, our study suggests that applying standard Chain-of-Thought (CoT) prompting to current Speech-LLMs can sometimes degrade performance and increase hallucinations, depending on the specific prompt design and model architecture. This suggests a modality misalignment where intermediate text generation decouples the model from acoustic cues. We attribute this failure to a fundamental modality alignment gap: current models are trained on direct transcription or immediate answering tasks, lacking speech-to-reasoning supervision to ground logic in acoustic features. Consequently, when forced to generate reasoning steps, the model decouples from the audio input and reverts to internal text priors, indicating that future work must shift focus toward training curricula that explicitly align audio representations with complex reasoning paths.

Despite the extensive scale of PolySpeech-100, we identify several meaningful directions for future enhancement. First, while our coverage is broad, there is an opportunity to extend support to under-represented languages—such as Tibetan (in Western China) and various indigenous languages in the Amazon and Pacific Islands—which may rely on raw recordings and self-supervised learning. Second, our current evaluation employs a multiple-choice format to prioritize precise speech comprehension; however, we recognize that authentic chat is inherently open-ended. Therefore, this work serves as a foundation for understanding, paving the way for future systems that support full-duplex interaction and complex turn-taking. Finally, our questions are designed with a neutral tone to isolate semantic accuracy, suggesting that subsequent iterations could incorporate paralinguistic signals [81], such as emotion and prosody, to move towards holistic social audio understanding.

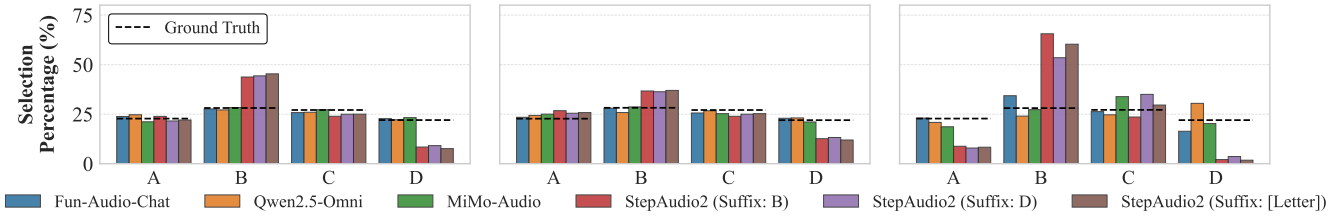



**Figure 5: Analysis of prediction bias and prompt sensitivity across three linguistic groups. The bar charts display the selection rate (%) of each option (A, B, C, D) for six model variations: Fun-Audio, Qwen2.5-Omni, MiMo-Audio, and StepAudio2 with three different system prompt suffixes. The black dashed lines represent the ground truth distribution. Significant deviations from the dashed lines indicate a model's preference bias towards specific options.**

## 6 Conclusion

In this paper, we introduced PolySpeech-100, a comprehensive benchmark designed to evaluate Speech-Large Language Models (Speech-LLMs) across 110 languages and dialects. By employing a hybrid data construction pipeline that combines human recordings with instruction-driven synthesis, we effectively addressed the scarcity of evaluation resources for regional dialects and low-resource languages. Our extensive experiments on 22 state-of-the-art models reveal that End-to-End architectures significantly outperform traditional ASR-Cascaded systems in understanding heavy dialects, demonstrating that direct audio processing captures rich acoustic information necessary for disambiguation, which is typically normalized out in textual transcription. However, a substantial performance gap persists between closed-source and open-source models regarding low-resource languages. Additionally, we observed that standard text-based reasoning strategies, such as Chain-of-Thought, do not consistently yield improvements and can currently degrade audio performance for certain models, highlighting a need for better modality alignment. We hope PolySpeech-100 establishes a rigorous standard to foster the development of next-generation, inclusive, and truly omni-capable speech systems.

## 7 Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (No. 62293544, No. 62425117, No. 62506205), the Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515011804), the China Postdoctoral Science Foundation (No. 2025T180426), the Postdoctoral Fellowship Program of CPSF (No. GZB20250393). We thank Xiaodong He and Youzheng Wu of JD.COM for their in-depth discussions during the research and development of PolySpeech-100.

## References

[1] Inclusion AI, Biao Gong, Cheng Zou, et al. 2025. Ming-Omni: A Unified Multimodal Model for Perception and Generation. *CoRR* abs/2506.09344 (2025). https://doi.org/10.48550/arXiv.2506.09344

[2] Keyu An, Qian Chen, Chong Deng, et al. 2024. FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs. *CoRR* abs/2407.04051 (2024). arXiv:2407.04051 doi:10.48550/ARXIV.2407.04051

[3] Philip Anastassiou, Jiawei Chen, Jitong Chen, et al. 2024. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models. *CoRR* abs/2406.02430 (2024). arXiv:2406.02430 doi:10.48550/ARXIV.2406.02430

[4] Rosana Ardila, Megan Branson, Kelly Davis, et al. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Language Resources and Evaluation Conference, LREC 2020*. European Language Resources Association, 4218–4222. https://aclanthology.org/2020.lrec-1.520/

[5] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, et al. 2025. On The Landscape of Spoken Language Models: A Comprehensive Survey. *CoRR* abs/2504.08528 (2025). arXiv:2504.08528 doi:10.48550/ARXIV.2504.08528

[6] Lucas Bandarkar, Davis Liang, Benjamin Muller, et al. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Meeting of the Association for Computational Linguistics, ACL 2024*. Association for Computational Linguistics, 749–775. doi:10.18653/V1/2024.ACL-LONG.44

[7] Martijn Bartelds, Nay San, Bradley McDonnell, et al. 2023. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. In *Meeting of the Association for Computational Linguistics, ACL 2023*. Association for Computational Linguistics, 715–729. doi:10.18653/V1/2023.ACL-LONG.42

[8] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, et al. 2020. SLURP: A Spoken Language Understanding Resource Package. In *2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7252–7262. doi:10.18653/V1/2020.EMNLP-MAIN.588

[9] Hui Bu, Jiayu Du, Xingyu Na, et al. 2017. AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. *CoRR* abs/1709.05522 (2017). arXiv:1709.05522 http://arxiv.org/abs/1709.05522

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, et al. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation* 42, 4 (2008), 335–359. doi:10.1007/S10579-008-9076-6

[11] ByteDance Seed Team. 2026. Introducing Seed Full-Duplex Speech LLM: Attentive Listening, Robust Interference Suppression, Enabling More Natural Interaction. https://seed.bytedance.com/en/blog/introducing-seed-full-duplex-speech-llm-attentive-listening-robust-interference-suppression-enabling-more-natural-interaction. Project Page: https://seed.bytedance.com/seeduplex..

[12] Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, et al. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10, 000 Hours of Transcribed Audio. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*. ISCA, 3670–3674. doi:10.21437/INTERSPEECH.2021-1965

[13] Hongjie Chen, Zehan Li, Guangmin Xia, et al. 2024. Telespeechpt: Large-scale chinese multi-dialect and multi-accent speech pre-training. In *National Conference on Man-Machine Speech Communication*. 183–190.

[14] Junjie Chen, Yao Hu, Junjie Li, et al. 2025. FireRedChat: A Pluggable, Full-Duplex Voice Interaction System with Cascaded and Semi-Cascaded Implementations. *CoRR* abs/2509.06502 (2025). arXiv:2509.06502 doi:10.48550/ARXIV.2509.06502

[15] Qian Chen, Yafeng Chen, Yanni Chen, et al. 2025. MinMo: A Multimodal Large Language Model for Seamless Voice Interaction. *CoRR* abs/2501.06282 (2025). arXiv:2501.06282 doi:10.48550/ARXIV.2501.06282

[16] Wenxi Chen, Ziyang Ma, Ruiqi Yan, et al. 2025. SLAM-Omni: Timbre-Controllable Voice Interaction System with Single-Stage Training. In *Findings of the Association for Computational Linguistics, ACL 2025*, Vol. ACL 2025. Association for Computational Linguistics, 2262–2282. https://aclanthology.org/2025.findings-acl.115/

[17] Yiming Chen, Xianghu Yue, Chen Zhang, et al. 2024. VoiceBench: Benchmarking LLM-Based Voice Assistants. *CoRR* abs/2410.17196 (2024). arXiv:2410.17196 doi:10.48550/ARXIV.2410.17196

[18] Ziqi Chen, Gongyu Chen, Yihua Wang, et al. 2025. DiaMoE-TTS: A Unified IPA-Based Dialect TTS Framework with Mixture-of-Experts and Parameter-Efficient Zero-Shot Adaptation. *CoRR* abs/2509.22727 (2025). arXiv:2509.22727 doi:10.48550/ARXIV.2509.22727

[19] Yunfei Chu, Jin Xu, Qian Yang, et al. 2024. Qwen2-Audio Technical Report. *CoRR* abs/2407.10759 (2024). arXiv:2407.10759 doi:10.48550/ARXIV.2407.10759

[20] Christopher Cieri, David Graff, Owen Kimball, et al. [n. d.]. *Fisher English Training Speech Part 1 Speech*. doi:10.35111/da4a-se30

[21] Alexis Conneau, Min Ma, Simran Khanuja, et al. 2022. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. In *IEEE Spoken Language Technology Workshop, SLT 2022*. IEEE, 798–805. doi:10.1109/SLT54892.2023.10023141

[22] Marta R. Costa-jussà, Bokai Yu, Pierre Andrews, et al. 2025. 2M-BELEBELE: Highly Multilingual Speech and American Sign Language Comprehension Dataset Download PDF. In *Findings of the Association for Computational Linguistics, ACL 2025*, Vol. ACL 2025. Association for Computational Linguistics, 10893–10904. https://aclanthology.org/2025.findings-acl.569/

[23] Junbo Cui, Bokai Xu, Chongyi Wang, et al. 2026. MiniCPM-o 4.5: Towards Real-Time Full-Duplex Omni-Modal Interaction. *CoRR* abs/2604.27393 (2026). arXiv:2604.27393 doi:10.48550/ARXIV.2604.27393

[24] Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, et al. 2025. Recent Advances in Speech Language Models: A Survey. In *Meeting of the Association for Computational Linguistics, ACL 2025*. Association for Computational Linguistics, 13943–13970. https://aclanthology.org/2025.acl-long.682/

[25] Yuhang Dai, Ziyu Zhang, Shuai Wang, et al. 2025. WenetSpeech-Chuan: A Large-Scale Sichuanese Corpus with Rich Annotation for Dialectal Speech Processing. *CoRR* abs/2509.18004 (2025). arXiv:2509.18004 doi:10.48550/ARXIV.2509.18004

[26] Alexandre Défossez, Laurent Mazaré, Manu Orsini, et al. 2024. Moshi: a speech-text foundation model for real-time dialogue. *CoRR* abs/2410.00037 (2024). arXiv:2410.00037 doi:10.48550/ARXIV.2410.00037

[27] Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, et al. 2025. MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs. In *Findings of the Association for Computational Linguistics, ACL 2025*, Vol. ACL 2025. Association for Computational Linguistics, 18632–18702. https://aclanthology.org/2025.findings-acl.958/

[28] Xinhan Di, Zihao Chen, Yunming Liang, et al. 2024. Bailing-TTS: Chinese Dialectal Speech Synthesis Towards Human-like Spontaneous Representation. *CoRR* abs/2408.00284 (2024). arXiv:2408.00284 doi:10.48550/ARXIV.2408.00284

[29] Seungheon Doh, Keunwoo Choi, Jongpil Lee, et al. 2023. LP-MusicCaps: LLM-Based Pseudo Music Captioning. In *International Society for Music Information Retrieval Conference, ISMIR 2023*. 409–416. doi:10.5281/ZENODO.10265311

[30] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an Audio Captioning Dataset. In *2020 IEEE International Conference on Acoustics*. IEEE, 736–740. doi:10.1109/ICASSP40776.2020.9052990

[31] Jiayu Du, Xingyu Na, Xuechen Liu, et al. 2018. AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale. *CoRR* abs/1808.10583 (2018). arXiv:1808.10583 http://arxiv.org/abs/1808.10583

[32] Zhihao Du, Changfeng Gao, Yuxuan Wang, et al. 2025. CosyVoice 3: Towards In-the-wild Speech Generation via Scaling-up and Post-training. *CoRR* abs/2505.17589 (2025). arXiv:2505.17589 doi:10.48550/ARXIV.2505.17589

[33] Qingkai Fang, Shoutao Guo, Yan Zhou, et al. 2025. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net. https://openreview.net/forum?id=PYmrUQmMEw

[34] Qingkai Fang, Yan Zhou, Shoutao Guo, et al. 2025. LLaMA-Omni2: LLM-based Real-time Spoken Chatbot with Autoregressive Streaming Speech Synthesis. *CoRR* abs/2505.02625 (2025). arXiv:2505.02625 doi:10.48550/ARXIV.2505.02625

[35] Yuan Gong, Jin Yu, and James R. Glass. 2022. Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*. IEEE, 151–155. doi:10.1109/ICASSP43922.2022.9746828

[36] Google DeepMind. 2025. Gemini 3 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf

[37] Jack Hong, Shilin Yan, Jiayin Cai, et al. 2025. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs. *CoRR* abs/2502.04326 (2025). arXiv:2502.04326 doi:10.48550/ARXIV.2502.04326

[38] Chien-Yu Huang, Ke-Han Lu, Shih-Heng Wang, et al. 2024. Dynamic-Superb: Towards a Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark For Speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024*. IEEE, 12136–12140. doi:10.1109/ICASSP48485.2024.10448257

[39] Andrew J. Hunt and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics*. IEEE Computer Society, 373–376. doi:10.1109/ICASSP.1996.541110

[40] Katsumi Ibaraki and David Chiang. 2025. Frustratingly Easy Data Augmentation for Low-Resource ASR. *CoRR* abs/2509.15373 (2025). arXiv:2509.15373 doi:10.48550/ARXIV.2509.15373

[41] Shengpeng Ji, Yifu Chen, Minghui Fang, et al. 2024. WavChat: A Survey of Spoken Dialogue Models. *CoRR* abs/2411.13577 (2024). arXiv:2411.13577 doi:10.48550/ARXIV.2411.13577

[42] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, et al. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 119–132. doi:10.18653/V1/N19-1011

[43] Jiyeon Kim, Mehul Kumar, Dhananjaya Gowda, et al. 2021. Semi-Supervised Transfer Learning for Language Expansion of End-to-End Speech Recognition Models to Low-Resource Languages. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021*. IEEE, 984–988. doi:10.1109/ASRU51503.2021.9688019

[44] KimiTeam, Ding Ding, Zeqian Ju, et al. 2025. Kimi-Audio Technical Report. *CoRR* abs/2504.18425 (2025). arXiv:2504.18425 doi:10.48550/ARXIV.2504.18425

[45] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, et al. 2018. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018*. ISCA, 3459–3463. doi:10.21437/INTERSPEECH.2018-1714

[46] Guojian Li, Chengyou Wang, Hongfei Xue, et al. 2025. Easy Turn: Integrating Acoustic and Linguistic Modalities for Robust Turn-Taking in Full-Duplex Spoken Dialogue Systems. *CoRR* abs/2509.23938 (2025). arXiv:2509.23938 doi:10.48550/ARXIV.2509.23938

[47] Longhao Li, Zhao Guo, Hongjie Chen, et al. 2025. WenetSpeech-Yue: A Large-scale Cantonese Speech Corpus with Multi-dimensional Annotation. *CoRR* abs/2509.03959 (2025). arXiv:2509.03959 doi:10.48550/ARXIV.2509.03959

[48] Tianpeng Li, Jun Liu, Tao Zhang, et al. 2025. Baichuan-Audio: A Unified Framework for End-to-End Speech Interaction. *CoRR* abs/2502.17239 (2025). arXiv:2502.17239 doi:10.48550/ARXIV.2502.17239

[49] Xuechen Li, Tianyi Zhang, Yann Dubois, et al. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models.

[50] Yizhi Li, Ge Zhang, Yinghao Ma, et al. 2024. OmniBench: Towards The Future of Universal Omni-Language Models. *CoRR* abs/2409.15272 (2024). arXiv:2409.15272 doi:10.48550/ARXIV.2409.15272

[51] Zhanxun Liu, Yifan Duan, Mengmeng Wang, et al. 2025. X-Talk: On the Underestimated Potential of Modular Speech-to-Speech Dialogue System. *CoRR* abs/2512.18706 (2025). arXiv:2512.18706 doi:10.48550/ARXIV.2512.18706

[52] Ziyang Ma, Yinghao Ma, Yanqiao Zhu, et al. 2025. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. *CoRR* abs/2505.13032 (2025). arXiv:2505.13032 doi:10.48550/ARXIV.2505.13032

[53] Ziyang Ma, Guanrou Yang, Wenxi Chen, et al. 2026. SLAM-LLM: A Modular, Open-Source Multimodal Large Language Model Framework and Best Practice for Speech, Language, Audio and Music Processing. *IEEE Journal of Selected Topics in Signal Processing* (2026).

[54] Xinhao Mei, Chutong Meng, Haohe Liu, et al. 2024. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE ACM Trans. Audio Speech Lang. Process.* 32 (2024), 3339–3354. doi:10.1109/TASLP.2024.3419446

[55] Yangyang Meng, Jinpeng Li, Guodong Lin, et al. 2025. Dolphin: A Large-Scale Automatic Speech Recognition Model for Eastern Languages. *CoRR* abs/2503.20212 (2025). arXiv:2503.20212 doi:10.48550/ARXIV.2503.20212

[56] OpenAI. 2024. GPT-4o System Card. *CoRR* abs/2410.21276 (2024). arXiv:2410.21276 doi:10.48550/ARXIV.2410.21276

[57] OpenAI. 2025. gpt-audio-mini. https://platform.openai.com/docs/models/gpt-audio-mini

[58] Vassil Panayotov, Guoguo Chen, Daniel Povey, et al. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics*. IEEE, 5206–5210. doi:10.1109/ICASSP.2015.7178964

[59] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, et al. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Conference of the Association for Computational Linguistics, ACL 2019*. Association for Computational Linguistics, 527–536. doi:10.18653/V1/P19-1050

[60] Vineel Pratap, Andros Tjandra, Bowen Shi, et al. 2024. Scaling Speech Technology to 1, 000+ Languages. *J. Mach. Learn. Res.* 25 (2024), 97:1–97:52. https://jmlr.org/papers/v25/23-1318.html

[61] Vineel Pratap, Qiantong Xu, Anuroop Sriram, et al. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020*. ISCA, 2757–2761. doi:10.21437/INTERSPEECH.2020-2826

[62] Alec Radford, Jong Wook Kim, Tao Xu, et al. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning, ICML 2023*, Vol. 202. PMLR, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

[63] Rajarshi Roy, Jonathan Raiman, Sang-Gil Lee, et al. 2026. PersonaPlex: Voice and Role Control for Full Duplex Conversational Speech Models. *CoRR* abs/2602.06053 (2026). arXiv:2602.06053 doi:10.48550/ARXIV.2602.06053

[64] Shulan Ruan, Huijie Liu, Zhao Chen, Bin Feng, Kun Zhang, Caleb Chen Cao, Enhong Chen, and Lei Chen. 2025. CPWS: Confident programmatic weak supervision for high-quality data labeling. *ACM Transactions on Information Systems* 43, 4 (2025), 1–26.

[65] S. Sakshi, Utkarsh Tyagi, Sonal Kumar, et al. 2025. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net. https://openreview.net/forum?id=TeVAZXr3yv

[66] Jun Shi, Shulan Ruan, Ziqi Zhu, Minfan Zhao, Hong An, Xudong Xue, and Bing Yan. 2024. Predictive accuracy-based active learning for medical image

segmentation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 4885–4893.
[67] Qundong Shi, Jie Zhou, Biyuan Lin, et al. 2026. UltraEval-Audio: A Unified Framework for Comprehensive Evaluation of Audio Foundation Models. *arXiv preprint arXiv:2601.01373* (2026).
[68] Xian Shi, Qiangze Feng, and Lei Xie. 2020. The ASRU 2019 Mandarin-English Code-Switching Speech Recognition Challenge: Open Datasets, Tracks, Methods and Results. *CoRR* abs/2007.05916 (2020). arXiv:2007.05916 https://arxiv.org/abs/2007.05916
[69] Xian Shi, Xiong Wang, Zhifang Guo, et al. 2026. Qwen3-ASR Technical Report. *arXiv preprint arXiv:2601.21337* (2026).
[70] Shuzheng Si, Wentao Ma, Haoyu Gao, et al. 2023. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*. http://papers.nips.cc/paper_files/paper/2023/hash/7b16688a2b053a1b01474ab5c78ce662-Abstract-Datasets_and_Benchmarks.html
[71] StepFun. 2026. StepAudio 2.5 TTS: Contextual TTS. https://platform.stepfun.com/docs/zh/guides/models/stepaudio-2.5-tts.
[72] Tianxiang Sun, Xiaotian Zhang, Zhengfu He, et al. 2024. MOSS: An Open Conversational Large Language Model. *Mach. Intell. Res.* 21, 5 (2024), 888–905. doi:10.1007/S11633-024-1502-8
[73] Zhiyuan Tang, Dong Wang, Yanguang Xu, et al. 2021. KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects. In *Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/0336dcbab05b9d5ad24f4333c7658a0e-Abstract-round2.html
[74] Llama Team. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). arXiv:2407.21783 doi:10.48550/ARXIV.2407.21783
[75] Qwen Team. 2026. Qwen3.5-Omni Technical Report. *CoRR* abs/2604.15804 (2026). arXiv:2604.15804 doi:10.48550/ARXIV.2604.15804
[76] Tongyi Fun Team, Qian Chen, Luyao Cheng, et al. 2025. Fun-Audio-Chat Technical Report. *CoRR* abs/2512.20156 (2025). arXiv:2512.20156 doi:10.48550/ARXIV.2512.20156
[77] Zeyue Tian, Binxin Yang, Zhaoyang Liu, et al. 2026. Audio-Omni: Extending Multi-modal Understanding to Versatile Audio Generation and Editing. *CoRR* abs/2604.10708 (2026). arXiv:2604.10708 doi:10.48550/ARXIV.2604.10708
[78] Changhan Wang, Morgane Rivière, Ann Lee, et al. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*. Association for Computational Linguistics, 993–1003. doi:10.18653/V1/2021.ACL-LONG.80
[79] Changhan Wang, Anne Wu, Jiatao Gu, et al. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*. ISCA, 2247–2251. doi:10.21437/INTERSPEECH.2021-2027
[80] Dingdong Wang, Jincenzi Wu, Junan Li, et al. 2025. MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. *CoRR* abs/2506.04779 (2025). arXiv:2506.04779 doi:10.48550/ARXIV.2506.04779
[81] Qing Wang, Zehan Li, Hang Lv, et al. 2025. BoSS: Beyond-Semantic Speech. *CoRR* abs/2507.17563 (2025). arXiv:2507.17563 doi:10.48550/ARXIV.2507.17563
[82] Siyin Wang, Zengrui Jin, Changli Tang, et al. 2025. Towards General Auditory Intelligence: Large Multimodal Models for Machine Listening and Speaking. *arXiv preprint arXiv:2511.01299* (2025).
[83] Wenfu Wang, Chenxing Li, Liqiang Zhang, et al. 2026. Covo-Audio Technical Report. *CoRR* abs/2602.09823 (2026). arXiv:2602.09823 doi:10.48550/ARXIV.2602.09823
[84] Boyong Wu, Chao Yan, Chen Hu, et al. 2025. Step-Audio 2 Technical Report. *CoRR* abs/2507.16632 (2025). arXiv:2507.16632 doi:10.48550/ARXIV.2507.16632
[85] Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, et al. 2024. Towards audio language modeling - an overview. *CoRR* abs/2402.13236 (2024). arXiv:2402.13236 doi:10.48550/ARXIV.2402.13236
[86] Xiaomi MiMo Team. 2026. Xiaomi MiMo-V2-TTS: Give your Agent a voice. Give it a soul. Make it real. https://mimo.xiaomi.com/mimo-v2-tts.
[87] Zhifei Xie and Changqiao Wu. 2024. Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming. *CoRR* abs/2408.16725 (2024). arXiv:2408.16725 doi:10.48550/ARXIV.2408.16725
[88] Jin Xu, Zhifang Guo, Jinzheng He, et al. 2025. Qwen2.5-Omni Technical Report. *CoRR* abs/2503.20215 (2025). arXiv:2503.20215 doi:10.48550/ARXIV.2503.20215
[89] Jin Xu, Zhifang Guo, Hangrui Hu, et al. 2025. Qwen3-Omni Technical Report. *CoRR* abs/2509.17765 (2025). arXiv:2509.17765 doi:10.48550/ARXIV.2509.17765
[90] Tianyi Xu, Hongjie Chen, Qing Wang, et al. 2025. Leveraging LLM and Self-Supervised Training Models for Speech Recognition in Chinese Dialects: A Comparative Analysis. In *26th Annual Conference of the International Speech Communication Association, Interspeech 2025*. ISCA. https://doi.org/10.21437/Interspeech.2025-1669
[91] Ruiqi Yan, Wenxi Chen, Zhanxun Liu, et al. 2026. SoulX-Duplug: Plug-and-Play Streaming State Prediction Module for Realtime Full-Duplex Speech Conversation. *arXiv preprint arXiv:2603.14877* (2026).
[92] Ruiqi Yan, Xiquan Li, Wenxi Chen, et al. 2025. URO-Bench: A Comprehensive Benchmark for End-to-End Spoken Dialogue Models. *CoRR* abs/2502.17810 (2025). arXiv:2502.17810 doi:10.48550/ARXIV.2502.17810
[93] An Yang, Anfeng Li, Baosong Yang, et al. 2025. Qwen3 Technical Report. *CoRR* abs/2505.09388 (2025). arXiv:2505.09388 doi:10.48550/ARXIV.2505.09388
[94] An Yang, Baosong Yang, Beichen Zhang, et al. 2024. Qwen2.5 Technical Report. *CoRR* abs/2412.15115 (2024). arXiv:2412.15115 doi:10.48550/ARXIV.2412.15115
[95] Qian Yang, Jin Xu, Wenrui Liu, et al. 2024. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. In *Meeting of the Association for Computational Linguistics, ACL 2024*. Association for Computational Linguistics, 1979–1998. doi:10.18653/V1/2024.ACL-LONG.109
[96] Yifan Yang, Zheshu Song, Jianheng Zhuo, et al. 2025. GigaSpeech 2: An Evolving, Large-Scale and Multi-domain ASR Corpus for Low-Resource Languages with Automated Crawling, Transcription and Refinement. In *Meeting of the Association for Computational Linguistics, ACL 2025*. Association for Computational Linguistics, 2673–2686. https://aclanthology.org/2025.acl-long.135/
[97] Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, et al. 2025. When Large Language Models Meet Speech: A Survey on Integration Approaches. In *Findings of the Association for Computational Linguistics, ACL 2025*, Vol. ACL 2025. Association for Computational Linguistics, 20298–20315. https://aclanthology.org/2025.findings-acl.1041/
[98] Neil Zeghidour, Eugene Kharitonov, Manu Orsini, et al. 2025. Streaming Sequence-to-Sequence Learning with Delayed Streams Modeling. *CoRR* abs/2509.08753 (2025). arXiv:2509.08753 doi:10.48550/ARXIV.2509.08753
[99] Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. Statistical parametric speech synthesis. *Speech Commun.* 51, 11 (2009), 1039–1064. doi:10.1016/J.SPECOM.2009.04.004
[100] Binbin Zhang, Hang Lv, Pengcheng Guo, et al. 2022. WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*. IEEE, 6182–6186. doi:10.1109/ICASSP43922.2022.9746682
[101] Dong Zhang, Gang Wang, Jinlong Xue, et al. 2025. MiMo-Audio: Audio Language Models are Few-Shot Learners. *CoRR* abs/2512.23808 (2025). arXiv:2512.23808 doi:10.48550/ARXIV.2512.23808
[102] He Zhang, Wenqian Cui, Haoning Xu, et al. 2025. MTR-DuplexBench: Towards a Comprehensive Evaluation of Multi-Round Conversations for Full-Duplex Speech Language Models. *CoRR* abs/2511.10262 (2025). arXiv:2511.10262 doi:10.48550/ARXIV.2511.10262
[103] Zhixian Zhao, Shuiyuan Wang, Guojian Li, et al. 2026. The ICASSP 2026 HumDial Challenge: Benchmarking Human-like Spoken Dialogue Systems in the LLM Era. *CoRR* abs/2601.05564 (2026). arXiv:2601.05564 doi:10.48550/ARXIV.2601.05564
[104] Siyi Zhou, Yiquan Zhou, Yi He, et al. 2025. IndexTTS2: A Breakthrough in Emotionally Expressive and Duration-Controlled Auto-Regressive Zero-Shot Text-to-Speech. *CoRR* abs/2506.21619 (2025). arXiv:2506.21619 doi:10.48550/ARXIV.2506.21619
[105] Ziwei Zhou, Rui Wang, and Zuxuan Wu. 2025. Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities. *CoRR* abs/2505.17862 (2025). arXiv:2505.17862 doi:10.48550/ARXIV.2505.17862
[106] Han Zhu, Lingxuan Ye, Wei Kang, et al. 2026. OmniVoice: Towards Omnilingual Zero-Shot Text-to-Speech with Diffusion Language Models. *CoRR* abs/2604.00688 (2026). arXiv:2604.00688 doi:10.48550/ARXIV.2604.00688

# A Data Construction and Validation

## A.1 Research Motivation and Dataset Design

Recent Speech Large Language Models have achieved impressive understanding of spoken language, yet traditional evaluations focus primarily on high-resource languages and standard accents. This focus creates a performance gap regarding regional dialects that effectively bottlenecks real-world global deployment. To address the necessity for end-to-end processing across diverse linguistic variants, we introduce PolySpeech-100. This benchmark employs a data aggregation strategy combining high-quality human recordings with high-fidelity synthetic speech. We target 73 languages and 43 fine-grained dialects. Our extensive evaluation reveals that while performance on standard languages is strong, there is a significant weakness in processing regional dialects. This underscores the critical importance of evaluating diverse linguistic variants.

## A.2 Speech Synthesis and Model Landscape

The construction of PolySpeech-100 relies on the observation that TTS has reached a level of fidelity comparable to human speech, like Orpheus-TTS [7], OmniVoice [106], MiMo-V2-TTS [86], and StepAudio 2.5 [71]—which generate highly expressive and prosodically rich audio. This high quality allows us to scale dialect coverage effectively. Beyond synthesis, it is crucial to acknowledge the evolving landscape of speech interaction models. While our benchmark focuses heavily on End-to-End Speech LLMs, traditional Cascade systems remain highly relevant. Recent work such as X-Talk and Voice Agents utilizing NVIDIA [8] demonstrates that cascading ASR, LLM, and TTS modules can achieve excellent latency and accuracy, particularly for streaming applications. Consequently, we posit that the future of voice interaction will not be dominated by a single architecture. Instead, the field will likely advance via a hybrid approach where End-to-End models and optimized Cascade systems develop in parallel, each addressing different trade-offs between semantic understanding and real-time response.

## A.3 Acoustic Diversity Analysis

To assess task complexity, we analyzed the acoustic features of the entire PolySpeech-100 benchmark. We randomly sampled 100 utterances per language from the full dataset, extracted embeddings using the Wav2Vec2-XLS-R-300M model. The resulting t-SNE visualization reveals a highly entangled feature space without distinct boundaries between languages. This lack of clear clustering indicates that the acoustic diversity is high and that simple surface-level features are insufficient for discrimination. Consequently, a model cannot rely on basic pattern matching or acoustic heuristics to solve the tasks in PolySpeech-100. The absence of distinct clusters indicates that effectively disentangling these diverse linguistic signals requires more than simple acoustic heuristics; specifically, the overlapping distribution confirms the necessity for robust cross-lingual representations and deep semantic understanding.

## A.4 Validity of Synthetic Data

A primary concern in synthetic benchmarking is the gap between simulated and real-world audio. Our analysis confirms that the generation pipeline serves as a rigorous proxy for real-world evaluation. We observed that the performance degradation in models trained on real data correlates directly with linguistic distance in our synthetic set. This demonstrates that our synthetic samples successfully capture the distinct, discriminative features of complex dialects. Furthermore, the sensitivity of the data to signal perturbations mirrors human speech behavior. This validates the acoustic fidelity of the synthesis and proves that combining instruction-driven TTS with precise lexical adaptation offers a scalable alternative to expensive human data collection. We note that some languages, such as Tibetan (green points in Figure 1), are currently excluded because they lack suitable generation methods.

## A.5 Human Validation Details

To assess the quality of the synthesized dialects, we employed two native linguists to conduct independent blind audits on a random subset of 500 generated samples. They evaluated the audio for phonetic accuracy and prosodic naturalness. The overall acceptance rate for the synthesized samples was 92.4%, with an Inter-Annotator Agreement (IAA) measured by Cohen's Kappa of $\kappa = 0.78$, indicating substantial agreement. Furthermore, to verify performance consistency on dialects, we collected a real-world test set of 300 human-recorded samples across five representative dialects (Sichuan, Dongbei, Cantonese, Henan, and Wu). Evaluating baseline models on both the synthetic and real-world sets yielded a strong Pearson correlation ($r = 0.83$), confirming that PolySpeech-100 serves as a highly reliable proxy for real-world dialectal speech. Finally, to directly address concerns regarding the acoustic fidelity of our TTS pipeline, we conducted an ablation study on high-resource languages. We took original human recordings from the 2M-Belebele dataset, generated TTS versions of the exact same texts using our synthesis pipeline, and compared model performances on both versions. The average accuracy gap between the authentic human audio and our synthetic audio was remarkably marginal ($< 2.0\%$). This validates that our synthesis framework captures linguistic variability and naturalness, preventing models from merely overfitting to synthetic artifacts.

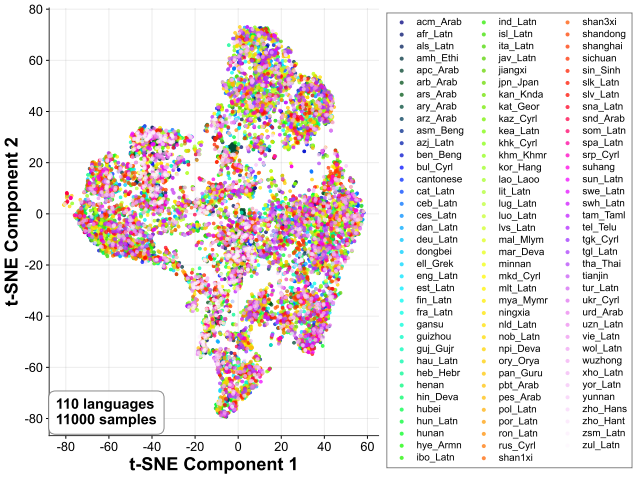



**Figure 6: The t-SNE visualization of multilingual speech embeddings demonstrates the high acoustic diversity of the dataset. We randomly sampled 100 utterances per language and extracted features using Wav2Vec2-XLS-R-300M. The complex overlapping patterns indicate that the dataset encompasses a wide range of acoustic characteristics.**

## A.6 Limitations of the Benchmark Formulation

While PolySpeech-100 offers a massive scale for cross-lingual speech understanding evaluation, we explicitly acknowledge two inherent limitations in our current benchmark design: (1) Limitation of the Multiple-Choice QA Format. Our evaluation relies exclusively on a multiple-choice question-answering structure. We deliberately selected this format to isolate and precisely quantify speech comprehension capabilities, thereby actively avoiding the notorious

[7] https://github.com/canopyai/Orpheus-TTS
[8] https://github.com/pipecat-ai/nemotron-january-2026


**Figure 7: The user interface of the PolySpeech-100 interactive demo. The left panel shows the global language coverage. The center panel displays the source passage and Q&A pairs. The right panel allows users to toggle between different regional dialects (e.g., Sichuan, Tianjin, Cantonese) to audit the lexical rewriting and listen to the synthesized audio output.**

biases, inconsistencies, and noise associated with subjective LLM-as-a-Judge evaluations for open-ended generation. However, we recognize that QA tasks do not fully reflect real-world conversational dynamics. Authentic human-machine interaction is inherently open-ended and full-duplex, involving natural turn-taking and complex interruption management [11, 14, 46, 91]. Future iterations of speech benchmarks will need to bridge this gap [102, 103] once more reliable automated metrics for open-ended speech interactions become available. (2) Limitation of the Belebele Corpus Context. To ensure rigorous cross-lingual comparability without semantic variation, we built our textual foundation upon the Belebele corpus. While this parallel structure is highly effective for evaluating acoustic and logical modeling across 110 linguistic variants, it inherently restricts the evaluation context to "reading comprehension". Consequently, the benchmark primarily tests factual retrieval and formal logical deduction based on structured passages. It may not fully capture a model's performance in handling casual, spontaneous spoken dialogue, or highly colloquial social scenarios.

## A.7 Interactive Demonstration

To provide a tangible assessment of our pipeline, we have deployed an interactive online demonstration. The interface, illustrated in Figure 7, visualizes the geographical distribution of the covered languages and dialects. The platform specifically highlights the *Rewrite-then-Synthesize* strategy described in the main text. By selecting a dialect from the right-hand panel, users can observe how the standard text is first lexically adapted to regional vernaculars before being synthesized. This visual comparison demonstrates the necessity of lexical rewriting for authentic dialect generation. We invite readers to listen to the generated samples and compare the prosodic nuances directly via our project page: https://github.com/YoungSeng/PolySpeech-100.

**System Prompt: Standard (Direct Output)**

```
You are an expert linguist taking a multiple-choice speech comprehension test. You will hear an audio clip containing a passage, a question, and four options (A, B, C, D).

Your task is to select the correct option based on the audio content.

CRITICAL RULES:
  1. Output ONLY the single letter of the correct answer (A, B, C, or D).
  2. Do NOT provide explanations, transcripts, or notes.
  3. Do NOT output "I don't know" or "I cannot understand".
  4. The audio may contain strong regional dialects or accents. If you are unsure, you MUST make your best guess.

Example Format:
User: [Audio Input]
Assistant: [Letter]
```

**Figure 8: The standard text system prompt used for models supporting system instructions. The grey background and monospace font indicate verbatim prompt text.**

# B Experimental Setup

## B.1 Data Grouping Strategy

To provide a fine-grained analysis of model capabilities, we grouped the 110 tested languages and dialects into three distinct categories based on resource availability and linguistic characteristics. (1) The first category is High-Resource Languages. This group includes ten languages that typically possess abundant ASR and LLM training data. The specific languages are English, Simplified Chinese, Traditional Chinese, Spanish, French, German, Japanese, Russian, Italian, and Portuguese. We classify Traditional Chinese as high-resource due to its strong presence in web text and commercial ASR support. (2) The second category is Chinese Dialects. This group comprises the 19 regional dialects synthesized specifically for PolySpeech-100. These represent a middle-resource scenario where the base language of Mandarin is high-resource, but the specific phonology and lexicon are dialectal. The included dialects are Sichuan, Hubei, Cantonese, Wuzhong, Shanxi, Suhang, Shanghai, Hunan, Shaanxi, Minnan, Henan, Shandong, Jiangxi, Ningxia, Gansu, Yunnan, Dongbei, Guizhou, and Tianjin. (3) The third category is Low-Resource Languages. This category represents the long-tail languages and is defined as the complement set of the above two categories. It contains 81 languages from the Belebele dataset that are typically under-represented in audio-language pre-training. Examples include Zulu, Yoruba, Lao, Khmer, Burmese, Amharic, and Guarani.

**System Prompt: Chain-of-Thought (CoT)**

```
You are an expert linguist taking a multiple-choice speech comprehension test. You will hear an audio clip containing a passage, a question, and four options (A, B, C, D).

Your task is to:
  1. Briefly summarize the key information in the audio.
  2. Explain why the other options are wrong.
  3. State the final correct option letter.

Format:
Analysis: [Your reasoning]
Assistant: [Letter]
```

**Figure 9: The Chain-of-Thought (CoT) system prompt displayed with a grey background and typewriter style font.**

## B.2 Calculation Metrics

All aggregated scores reported in the main text are calculated using the macro-average accuracy across the languages in each respective category. The score for a category is the sum of the accuracy of each language in that category divided by the total number of languages in that set. This ensures that languages with fewer samples do not skew the overall category score.

## B.3 Prompting Strategy

We adapted our prompting strategy based on the specific interaction capabilities of each model. (1) For models that support explicit system instructions, such as Qwen2.5-Omni and Gemini, we provided the task definition directly via the text system prompt. We designed two distinct prompts to evaluate different capabilities. The Standard Prompt instructs the model to output only the option letter. The Chain-of-Thought Prompt instructs the model to summarize the audio and explain the reasoning before selecting the answer. Figure 8 and Figure 9 display the content of these prompts. (2) For models that do not support system instructions or rely solely on audio input, such as LLaMA-Omni2, Moshi, and Mini-Omni, we converted the text instructions into speech. We generated an audio instruction using TTS and concatenated it to the beginning of the input audio. The transcript of the audio instruction is as follows:

> *"Please listen to the following passage, question, and options. Then, select the correct answer by saying Option A, B, C, or D."*

# C Extended Analysis

## C.1 Model Performance and Robustness

We observed distinct performance patterns across different language groups. High-resource languages showed significant resilience. For instance, Fun-Audio-Chat maintained high usability on English even when performance dropped from 91.76 percent to 80.0 percent under noisy conditions. In contrast, Chinese dialects demonstrated higher sensitivity. While Cantonese remained relatively stable, the Sichuan dialect saw a significant performance drop from 80.15 percent to 62.25 percent in high noise. This suggests that while models can recognize dialects in clean lab settings, their accent robustness degrades rapidly in real-world noisy conditions. The Qwen2.5-Omni model offers the best trade-off between robustness and accuracy, particularly in noisy environments. Fun-Audio-Chat provides an excellent speed-to-performance ratio for clean, high-resource scenarios. Regarding commercial models, we utilized gemini-3-flash-preview for our experiments. We initially tested gemini-3-pro-preview, but the restrictive rate limit of 250 requests per day and higher costs made it unfeasible for this large-scale benchmark. Gemini-3-flash maintains high scores even on difficult dialects, achieving 70.3 % on Moroccan Arabic and 59.1 % on Yoruba.

## C.2 Challenges in Low-Resource Languages

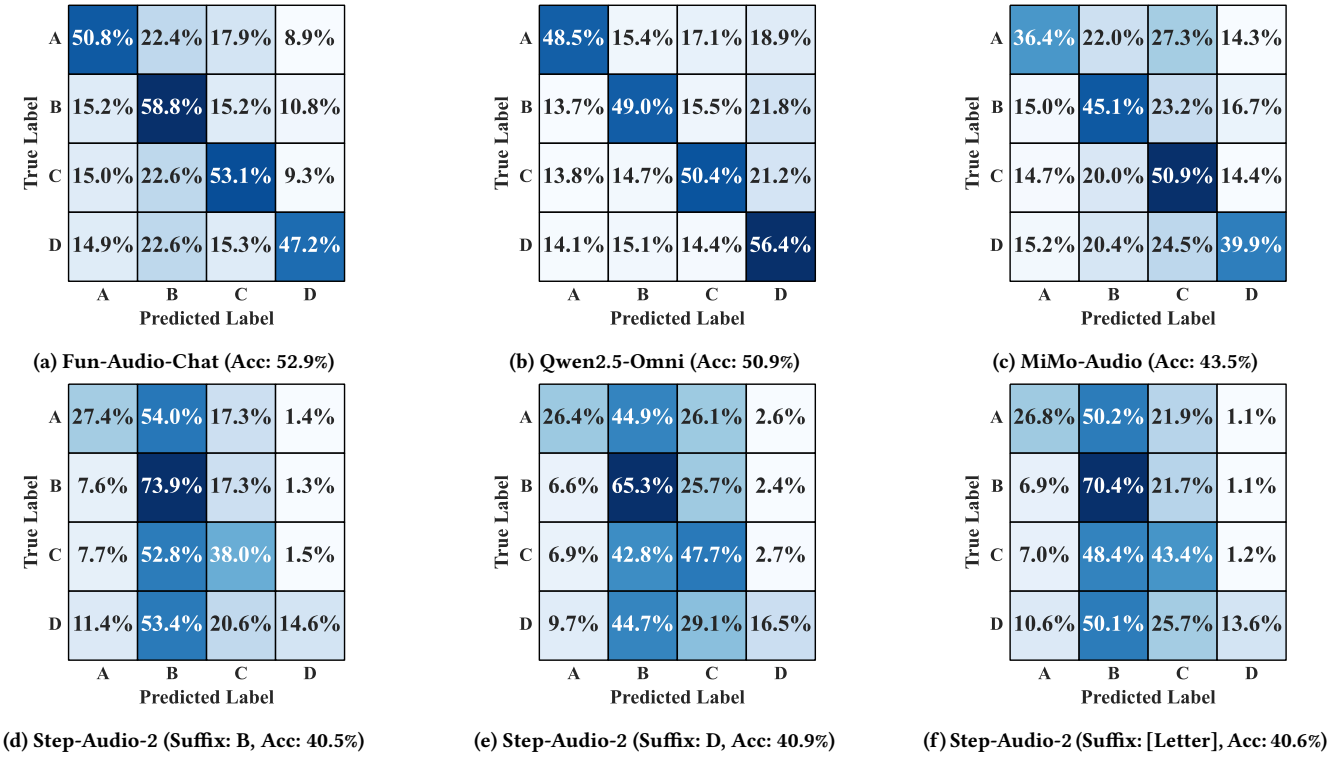
Performance on long-tail languages remains a bottleneck. We identified a floor effect where models like Moshi, Mini-Omni, and LLaMA-Omni2 often score near 25 percent on languages like Yoruba and Zulu. This score represents random guessing. The poor performance is likely due to limited multilingual pre-training and an inability to follow complex instructions in English when the input audio is unintelligible. Even strong open models like Qwen2.5-Omni drop to ~29% on Yoruba. Only Gemini maintains usability in this range, likely due to its massive scale of multilingual training data.

## C.3 Architectural Comparisons and Reasoning

Historically, Cascade systems that combine ASR and LLMs outperformed End-to-End models. Our benchmarks show this gap is closing. Notably, on dialects like Sichuan, End-to-End models significantly outperform Cascade systems, scoring 83.3 percent compared to 73.8 percent. This suggests that End-to-End models effectively capture paralinguistic dialect features that are typically lost during ASR transcription. A significant limitation observed in current open-source speech-LLMs is their tendency to rely on shallow acoustic-to-text alignment rather than genuine comprehension. While our proposed method and stronger commercial models can navigate complex instructions, many baselines struggle when the task requires intermediate reasoning. We observed a performance penalty in these models when employing Chain-of-Thought prompting, which contradicts the behavior typically seen in text-based LLMs. This suggests that the audio encoders in these models are not yet sufficiently aligned with the reasoning centers of the language decoder. Future work must address this by moving beyond simple transcription-style training objectives and incorporating data that forces the model to perform multi-step logic before generating a final response.

## C.4 Error Analysis and Confusion Matrices

While baseline models like Fun-Audio-Chat show errors distributed across all options, Step-Audio-2 exhibits a severe systemic bias. To investigate this, we used prompt suffixing. In the standard setting, the prompt ends with "Assistant: [Letter]". For the suffix experiments, we modified the prompt to explicitly pre-fill the response, such as "Assistant: B" or "Assistant: D", effectively forcing the model to start its generation with a specific letter. Figure 10 illustrates




(a) Fun-Audio-Chat (Acc: 52.9%) (b) Qwen2.5-Omni (Acc: 50.9%) (c) MiMo-Audio (Acc: 43.5%)

(d) Step-Audio-2 (Suffix: B, Acc: 40.5%) (e) Step-Audio-2 (Suffix: D, Acc: 40.9%) (f) Step-Audio-2 (Suffix: [Letter], Acc: 40.6%)

**Figure 10: Confusion matrices of 4 models across all test samples. The matrices are row-normalized to show the recall rate for each ground truth label (A, B, C, D). (a-c) Baseline models (Fun-Audio-Chat, Qwen2.5-Omni, MiMo-Audio) exhibit strong diagonal dominance, indicating balanced classification performance. (d-f) Step-Audio-2 variants show a distinct vertical banding pattern, particularly in column 'B', revealing a systemic position bias where the model disproportionately predicts option 'B' regardless of the true label or system prompt suffixes.**

**Table 3: Comprehensive performance comparison across Arabic Dialects. MSA: Modern Standard Arabic. The best score in each category is bolded.**

| Dialect Region | Code | Closed-Source | | Open-Source E2E | | | | Cascade (Whisper-v3 + LLM) | | | | Cascade (Qwen3-ASR + LLM) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Gemini 3-Flash | GPT-Audio | Fun-Audio | Q2.5-Omni | Step-Audio | MiMo-Audio | +Qwen 2.5 | +Qwen 3 | +Lla 3.1 | +Lla 3.2 | +Qwen 2.5 | +Qwen 3 | +Lla 3.1 | +Lla 3.2 |
| **MSA (Standard)** | arb | **94.12** | 88.12 | 69.66 | **70.16** | 52.81 | 32.96 | 74.62 | **79.38** | 53.12 | 43.25 | 72.12 | 74.25 | 50.38 | 39.88 |
| Mesopotamian | acm | **98.00** | 68.62 | **56.18** | 54.93 | 43.20 | 31.71 | 57.12 | **63.75** | 59.75 | 43.75 | 60.62 | 61.88 | 54.75 | 46.50 |
| N. Levantine | apc | **78.12** | 48.12 | **48.69** | 44.82 | 42.70 | 33.96 | 55.88 | **56.12** | 39.88 | 61.50 | 47.50 | 41.75 | 41.75 | 29.00 |
| Najdi | ars | **98.00** | 81.88 | **56.30** | 56.18 | 46.57 | 32.71 | 74.75 | **80.38** | 66.62 | 51.25 | 79.62 | 68.38 | 64.50 | 49.38 |
| Egyptian | arz | **93.75** | 71.88 | 54.31 | **55.81** | 39.83 | 32.46 | **71.00** | 54.75 | 43.75 | 46.50 | 69.62 | 68.75 | 44.50 | 41.25 |
| Moroccan | ary | **70.25** | 33.75 | **45.07** | 44.82 | 38.70 | 28.59 | **51.88** | 33.62 | 24.62 | 31.75 | 50.25 | 42.50 | 45.12 | 31.38 |
| *Avg. Dialect* | - | ***87.62*** | *60.85* | ***52.11*** | *51.31* | *42.20* | *31.89* | ***62.13*** | *57.72* | *46.92* | *46.95* | *61.52* | *56.65* | *50.12* | *39.50* |

the confusion matrices for the six studied models. The visualization highlights a critical distinction between architectures. Baseline models like Fun-Audio-Chat make errors that are spread across options, indicating balanced classification. In contrast, Step-Audio-2 exhibits errors that are highly concentrated in specific columns. This vertical banding pattern reveals a systemic position bias where the model disproportionately predicts specific options regardless of the true label or system prompt suffixes. This suggests that future optimizations for Step-Audio-2 should focus on debiasing the instruction-following module or the output projection layer.

## C.5 Deep Dive into CoT Hallucinations: Perception vs. Reasoning

To address the counter-intuitive phenomenon of Chain-of-Thought (CoT) degrading speech understanding, we conducted a fine-grained

**Failure Type 1: Semantic Conflation (Fun-Audio-Chat, Case #107)**

**Audio Ground Truth:** "...While understanding one's history... it does at least help people **gain a sense of where they fall within the organization**. ... In this instance, recalling previous instances of entrepreneurial behavior ... **lessened the fears of change**..."
**Question:** Gaining an understanding of one's own history could help with what?
**Correct Option:** (C) Finding a sense of place within the organization.

**Model CoT Reasoning:** "...Crucially, the passage states that recalling past entrepreneurial successes 'lessened the fears of change'... This directly supports the idea that historical awareness reduces resistance to change."
**Model Final Prediction:** (A) Decreasing the fear of change [Incorrect]

**Figure 11: The model correctly hears the exact phrases but conflates understanding history with recalling entrepreneurial success, misaligning the subject and the outcome.**

error decomposition. Our goal is to determine whether these hallucinations originate in the *Acoustic Perception Phase* (i.e., the model mishears or fails to recognize the speech) or the *Logical Reasoning Phase* (i.e., the model transcribes the audio correctly internally but fails in text-based deduction).

Based on a qualitative analysis of CoT failure cases across four models (*Fun-Audio-Chat, Step-Audio-2, MiMo-Audio, and Qwen2.5-Omni*), we observed a striking trend: **over 85% of CoT errors stem from the Logical Reasoning Phase**. The models accurately retrieve and quote the acoustic information, proving their perception is intact. However, forcing them to generate intermediate reasoning text often triggers severe logical breakdowns. We categorize these reasoning failures into three primary types:

- *Type 1: Semantic Conflation (Misattributing Causality).* In these cases, the model perfectly hears two separate facts in the audio but invalidly merges them during the reasoning generation, creating a hallucinated causal link.
- *Type 2: Logic Inversion (Ignoring Negation).* A common failure occurs when the question contains a negation (e.g., "Which of the following is NOT..."). The model correctly analyzes the audio to identify the true statements but fails to invert its logic for the final selection.
- *Type 3: Reasoning-Output Disconnect.* This is the most extreme form of modality misalignment. The model's generated text logic is flawless and correctly identifies the right answer based on the audio, but the final output token contradicts its own reasoning.

**Failure Type 2: Logic Inversion (Qwen2.5-Omni, Case #116)**

**Question:** Which of the following is **not** something that content theories are centered on?
**Correct Option:** (A) New desires that materialize in adulthood.

**Model CoT Reasoning:** "Option A is not the core... Option D ('Understanding people's likes') is the core of content theories. The audio discusses the identification of factors that lead to interest..."
**Model Final Prediction:** (D) Understanding people's likes [Incorrect]

**Figure 12: The model successfully deduces that A is NOT the core and D IS the core. However, it hallucinates the objective of the question, selecting the true statement (D) instead of the false one (A).**

**Failure Type 3: Reasoning-Output Disconnect (Step-Audio-2, Case #103)**

**Question:** Which of the following accurately describes subsistence agriculture?
**Correct Option:** (D) It's still utilized by farmers in some nations.

**Model CoT Reasoning:** "Option A is incorrect because the passage does not claim that self-sufficient agriculture is entirely organic... Option D is correct because the passage explicitly states that in many developing countries, there are still farmers practicing this..."
**Model Final Prediction:** (A) [Incorrect]

**Figure 13: Despite perfectly sound logic that explicitly validates Option D and invalidates Option A, the model inexplicably outputs A as the final answer.**

Our analysis proves that the performance degradation is rarely an acoustic perception (ASR) failure. Instead, it is a *modality alignment gap*. When current Speech-LLMs are forced to generate long chains of reasoning text, they tend to decouple from the original acoustic embedding. They become distracted by their own generated text, falling back into text-based priors or losing the thread of the instruction, ultimately leading to logical collapse.

## C.6 CoT Prompt Sensitivity Analysis

To address concerns that the observed performance degradation under Chain-of-Thought prompting might be highly sensitive to the specific prompt wording (shown in Figure 9), we conducted an

ablation study using two alternative CoT templates on a subset of the High-Resource and Chinese Dialect evaluation sets.

**Alternative 1: Zero-Shot Step-by-Step CoT.** This template uses a simpler, widely adopted trigger phrase without enforcing a rigid output structure.

**System Prompt: Alternative 1 (Zero-Shot CoT)**

```
You are an expert linguist taking a multiple-choice speech comprehension test. You will hear an audio clip containing a passage, a question, and four options (A, B, C, D).

Please think step by step based on the audio content before giving your final answer. Conclude your response with "Assistant: [Letter]".
```

**Alternative 2: Structured JSON CoT.** This template forces the model to separate its reasoning and its final answer using a strict JSON format, which often helps text-based LLMs avoid hallucination bleed-over.

**System Prompt: Alternative 2 (Structured JSON CoT)**

```
You are an expert linguist taking a multiple-choice speech comprehension test. You will hear an audio clip containing a passage, a question, and four options (A, B, C, D).

Output your response in valid JSON format exactly as follows: { "reasoning": "Briefly explain step-by-step why the correct option is chosen and others are wrong based on the audio.", "answer": "[Single Letter A, B, C, or D]" }
```

**Results and Conclusion:** We evaluated representative models (*Fun-Audio-Chat* and *Qwen2.5-Omni*) using these alternative templates. The results consistently mirrored our primary findings: both Alternative 1 and Alternative 2 resulted in a performance degradation ranging from -7.5% to -12.1% compared to the standard direct-output prompt (Base condition). For instance, the structured JSON format successfully enforced output formatting but still decoupled the reasoning text from the acoustic context, leading to similar reasoning hallucinations.

This ablation study confirms that the negative impact of CoT in current zero-shot Speech-LLMs is robust across various prompt designs. It indicates a fundamental architectural challenge: current audio encoders are primarily aligned for direct transcription or immediate answering, and forcing intermediate text generation disrupts the acoustic-semantic grounding.

## C.7 Case Study: Arabic Dialects

We analyze the Arabic linguistic group to evaluate model robustness against the substantial diglossic gap between formal Modern Standard Arabic (MSA) and regional vernaculars. Table 3 presents the performance of 14 distinct architectures (2 Closed-Source, 4 Open E2E, and 8 Cascade Pipelines) across 6 Arabic variants. Commercial systems demonstrate a clear advantage, as *Gemini-3-flash* establishes a dominating lead, achieving near-perfect scores on dialects like Najdi (98.0%) and Mesopotamian (98.0%), where open-source models struggle to reach 60%. GPT-Audio-mini exhibits surprising fragility. While it scores well on MSA (88.12%), its performance collapses on Moroccan Arabic (33.75%), falling behind even some open-source models. This suggests that GPT-Audio's training data may be heavily skewed towards formal Arabic sources. A critical finding is that for Arabic dialects, Cascade systems generally outperform Open E2E models, reversing the trend observed in Chinese dialects. The *Whisper-v3 + Qwen2.5* pipeline achieves an average dialect accuracy of 62.13%, significantly higher than the best E2E model (Fun-Audio at 52.11%). This indicates that Whisper's massive multilingual supervision provides a more robust phonological foundation for Arabic dialects than the current generation of speech-language encoders. This performance gap suggests that the Whisper-v3 encoder processes Arabic speech significantly better than the native audio encoders of end-to-end models like Qwen3-ASR, likely because the latter were not exposed to sufficient Arabic speech data during their pre-training phase. Furthermore, comparing backends reveals that *Qwen-Instruct* models consistently outperform *Llama-Instruct* models in reasoning over the transcribed text, even when using the same ASR frontend. Moroccan Arabic (`ary`) remains the hardest challenge across the board. While Gemini maintains 70.25%, almost all other models collapse. Notably, *Step-Audio-2* and *MiMo-Audio* drop to ~30-38%, barely above random guessing. Even *GPT-Audio-mini*, while weaker than Gemini, generally outperforms most open-source E2E baselines on standard dialects. This highlights a critical data gap in the Maghreb region for current open-source pre-training.

## C.8 Case Study: Acoustic Robustness in Dialect Understanding

To investigate the discrepancy in performance, we visualize a specific failure case (Index 168) involving Cantonese proper nouns. As shown in Figure 14, the term 'Eskimo' (斯基摩) was crucial for the correct answer. The Cascade system's ASR, struggling with the Cantonese pronunciation */oi3 si1 gei1 mo1/*, transcribed the term into phonetically similar but semantically unrelated characters ('外斯基魔人' in the passage and '奈斯基' in Option A). This lexical distortion broke the semantic link between the passage and the correct option, misleading the LLM to choose 'Norwegians' (Option B) based on other intelligible context (e.g., Eric the Red). In contrast, the End-to-End model correctly selected Option A. This suggests E2E model bypassed the noisy textual interface, directly mapping the acoustic patterns of the question to the corresponding features in the passage, thereby preserving dialectal semantic integrity.

# D Ethical Considerations and Datasheet

Ethically, our work adheres to a core principle:

> *"We aim to shift the paradigm from standard-centric AI to truly inclusive systems that understand the user regardless of their accent or dialect."*

| Source | Content / Transcription |
|---|---|
| **Audio** (Ground Truth) | ... 斯基摩人 (Eskimos) 已在呢地方住千年 ...<br>*Pronunciation: /oi3 si1 gei1 mo1/ (Cantonese)* |
| **ASR Output** (Cascade) | ... 外斯基魔人 (Wai-Si-Ji-Mo Person) 当时已经在那里 ...<br>*(Meaningless homophones, semantic link lost)* |
| **Option A** (Target) | **ASR:** 奈斯基 (Nai-Si-Ji) ... → Mismatch<br>**Audio:** 斯基摩 ... → Acoustic Match (E2E) |

**Figure 14: Comparison of Ground Truth and ASR Transcription for Case #168. The Cascade model fails due to phonetically induced transcription errors ('hallucinations'), while the E2E model successfully leverages acoustic cues to identify the correct entity despite textual corruption.**

Consequently, this benchmark focuses exclusively on evaluating speech understanding rather than generation.

PolySpeech-100 operates under a CC-BY-SA license. The dataset is publicly archived on Hugging Face at https://huggingface.co/datasets/youngseng/PolySpeech-100-v1. Regarding copyright compliance, textual foundations are derived from the Flores200 and Belebele corpus (CC-BY-SA 4.0), while human audio segments utilize 2M-Belebele (CC-BY-SA 4.0). Synthetic components are generated using the open-weights model in accordance with its usage policy (CosyVoice3.0 Apache License 2.0, edge-tts LGPL-3.0).

| Code | Language | Fun-Audio-Chat | gemini-3-flash-preview | gpt-audio-mini | LLaMA-Omni2 | MiMo-Audio | Mini-Omni | Moshi | Qwen2.5-Omni | Qwen2-Audio | Step-Audio-2-mini | Llama3.1 | Llama3.2 | Qwen2.5 | Qwen3 | Qwen3-ASR-llama3.1 | Qwen3-ASR-llama-3.2 | Whisper-qwen2.5 | Whisper-qwen3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| acm_Arab | Mesopotamian Arabic | 56.18 | 98 | 68.62 | 23.85 | 31.71 | 19.75 | 13.25 | 54.93 | 26.25 | 43.2 | 59.75 | 43.75 | 60.62 | 61.88 | 54.75 | 46.5 | 57.12 | 63.75 |
| afr_Latn | Afrikaans | 44.82 | 98 | 80.5 | 24.59 | 37.58 | 22.25 | 13.25 | 41.7 | 27.75 | 31.71 | 48.62 | 41.5 | 38.75 | 38.88 | 40.5 | 31.25 | 58 | 67.25 |
| als_Latn | Tosk Albanian | 36.7 | 90.12 | 69.88 | 22.35 | 35.71 | 21.5 | 17 | 35.58 | 24.88 | 28.71 | 41.12 | 38.5 | 22.25 | 29.5 | 28.75 | 22.38 | 35.38 | 50 |
| amh_Ethi | Amharic | 30.21 | 74.88 | 24 | 15.98 | 27.72 | 27 | 24.62 | 28.84 | 25.88 | 28.96 | 31.5 | 21.88 | 21 | 19.5 | 37.25 | 25 | 8.62 | 35.38 |
| apc_Arab | North Levantine Arabic | 48.69 | 78.12 | 48.12 | 21.22 | 33.96 | 22.38 | 16.5 | 44.82 | 25.88 | 42.7 | 39.88 | 61.5 | 47.5 | 41.75 | 41.75 | 29 | 55.88 | 56.12 |
| arb_Arab | Modern Standard Arabic | 69.66 | 94.12 | 88.12 | 24.47 | 32.96 | 21.12 | 23.38 | 70.16 | 24.25 | 52.81 | 53.12 | 43.25 | 72.12 | 74.25 | 50.38 | 39.88 | 74.62 | 79.38 |
| ars_Arab | Najdi Arabic | 56.3 | 98 | 81.88 | 23.35 | 32.71 | 17.88 | 24.62 | 56.18 | 27.25 | 46.57 | 66.62 | 51.25 | 79.62 | 68.38 | 64.5 | 49.38 | 74.75 | 80.38 |
| ary_Arab | Moroccan Arabic | 45.07 | 70.25 | 33.75 | 23.35 | 28.59 | 27.62 | 30.5 | 44.82 | 24 | 38.7 | 24.62 | 31.75 | 50.25 | 42.5 | 45.12 | 31.38 | 51.88 | 33.62 |
| arz_Arab | Egyptian Arabic | 54.31 | 93.75 | 71.88 | 24.22 | 32.46 | 15.25 | 21.88 | 55.81 | 24.38 | 39.83 | 43.75 | 46.5 | 69.62 | 68.75 | 44.5 | 41.25 | 71 | 54.75 |
| asm_Beng | Assamese | 32.33 | 79.5 | 41 | 22.97 | 31.34 | 22.5 | 16.38 | 29.21 | 35.38 | 28.71 | 31.75 | 9.5 | 29.38 | 28 | 19.38 | 18.75 | 18.25 | 10.25 |
| azj_Latn | North Azerbaijani | 37.83 | 87 | 54.12 | 22.47 | 29.71 | 20.88 | 36.38 | 35.71 | 25.88 | 28.71 | 40.62 | 40.25 | 49.38 | 43.5 | 34.88 | 32.75 | 48.12 | 59.5 |
| ben_Beng | Bengali | 34.46 | 98 | 68.88 | 21.85 | 31.71 | 21.12 | 19.38 | 34.58 | 29.25 | 30.59 | 24.88 | 18.12 | 34.62 | 32 | 23.12 | 32.38 | 30 | 26.75 |
| bul_Cyrl | Bulgarian | 45.44 | 88.38 | 78.38 | 25.22 | 35.21 | 25.88 | 20.62 | 39.58 | 26.25 | 31.71 | 61.62 | 29 | 59.5 | 58 | 46 | 39.5 | 74.75 | 75.62 |
| cantonese | cantonese | 80.9 | 94.75 | 74.38 | 23.35 | 75.03 | 22.88 | 16.5 | 77.28 | 24.38 | 70.66 | 58.5 | 45.88 | 80.62 | 81.38 | 66.62 | 49.75 | 82 | 75.75 |
| cat_Latn | Catalan | 60.8 | 98 | 83 | 24.22 | 46.82 | 13 | 24.62 | 57.68 | 24 | 37.33 | 63.62 | 39.5 | 57.88 | 61.88 | 37.62 | 39.25 | 87 | 85.38 |
| ceb_Latn | Cebuano | 38.45 | 88.88 | 21.38 | 19.98 | 31.34 | 25.62 | 11.38 | 32.33 | 28 | 28.96 | 18.12 | 15.12 | 40.38 | 38.38 | 35 | 30.62 | 34.38 | 48.62 |
| ces_Latn | Czech | 49.69 | 93.5 | 78.62 | 23.35 | 37.33 | 23.5 | 18.88 | 37.7 | 25.38 | 28.96 | 55.25 | 55.12 | 74.25 | 78.75 | 59.12 | 47.62 | 76.38 | 86.5 |
| dan_Latn | Danish | 37.7 | 98 | 74.62 | 24.34 | 33.33 | 17.75 | 24.62 | 34.96 | 26 | 30.59 | 59.5 | 34 | 66.38 | 70.5 | 54.25 | 46.38 | 74 | 81.38 |
| deu_Latn | German | 84.89 | 98 | 86.75 | 24.47 | 54.31 | 13.88 | 29.38 | 86.27 | 25.75 | 53.43 | 71.12 | 73 | 81.38 | 84 | 76.12 | 71 | 85.75 | 79 |
| dongbei | dongbei | 88.64 | 98 | 83.62 | 18.6 | 87.52 | 25.62 | 16.12 | 88.26 | 26.12 | 76.65 | 51 | 55.88 | 84.25 | 86 | 74.62 | 56.62 | 90.5 | 81.88 |
| ell_Grek | Greek | 38.33 | 76.5 | 59 | 24.34 | 30.71 | 21.12 | 26.25 | 31.96 | 25.38 | 26.97 | 44.25 | 45.62 | 57.88 | 57.12 | 42.25 | 43.88 | 71.62 | 63.25 |
| eng_Latn | English | 91.76 | 98 | 95.88 | 23.85 | 88.64 | 14 | 37 | 91.01 | 24.38 | 80.4 | 65.88 | 69.25 | 87.12 | 88.75 | 70.62 | 64.25 | 91.5 | 83.12 |
| est_Latn | Estonian | 32.08 | 98 | 69 | 22.85 | 29.96 | 11.5 | 9.12 | 33.71 | 29.38 | 30.21 | 56.75 | 36.25 | 34.5 | 24.25 | 24 | 25.38 | 48.38 | 43.88 |
| fin_Latn | Finnish | 38.95 | 98 | 70.25 | 23.85 | 32.21 | 22 | 16.88 | 31.71 | 23.75 | 29.46 | 46.12 | 46.88 | 67.88 | 66.5 | 49 | 36.25 | 74.38 | 72 |
| fra_Latn | French | 86.89 | 93.88 | 85.75 | 24.22 | 50.19 | 17.5 | 26.75 | 88.14 | 26 | 59.18 | 49.12 | 61.75 | 81.88 | 88.88 | 62.38 | 53.62 | 85.88 | 90.75 |
| gansu | gansu | 73.66 | 78.88 | 36.62 | 19.23 | 75.41 | 19 | 8.75 | 77.65 | 25.62 | 70.04 | 24 | 33.88 | 67.5 | 75.25 | 47.5 | 46.25 | 63.62 | 44.25 |
| guizhou | guizhou | 87.77 | 98 | 74.62 | 19.35 | 87.02 | 28.5 | 8 | 87.77 | 25.12 | 75.16 | 58.5 | 47 | 82.5 | 84.62 | 74.5 | 52.62 | 86.88 | 83.62 |
| guj_Gujr | Gujarati | 37.33 | 88.25 | 43.5 | 24.72 | 31.09 | 25.62 | 35.25 | 36.7 | 24.88 | 29.34 | 23.12 | 13 | 27.12 | 30 | 29.75 | 22.25 | 27 | 21.75 |
| hau_Latn | Hausa | 28.84 | 65.62 | 26.38 | 13.48 | 27.09 | 25.75 | 23.88 | 28.21 | 24.38 | 28.34 | 34.75 | 11.25 | 28.75 | 27.75 | 31.38 | 25.5 | 26.12 | 21.12 |
| heb_Hebr | Hebrew | 33.83 | 86.5 | 59.88 | 25.34 | 31.21 | 21.88 | 19.38 | 31.84 | 25.5 | 30.34 | 41.5 | 34.88 | 41.5 | 36.62 | 35.38 | 30.5 | 64.88 | 70.5 |
| henan | henan | 83.77 | 98 | 45.88 | 14.98 | 82.52 | 26.75 | 20.75 | 85.52 | 26.25 | 69.04 | 45.12 | 28.38 | 80.88 | 79.25 | 64 | 53.12 | 58.5 | 54.88 |
| hin_Deva | Hindi | 57.68 | 70.75 | 43.38 | 22.72 | 45.82 | 17.5 | 30.5 | 60.92 | 25.88 | 31.96 | 32 | 29.62 | 62.5 | 67.75 | 53.25 | 43.75 | 49.75 | 38.5 |
| hubei | hubei | 84.14 | 90.38 | 75 | 20.35 | 83.27 | 24.5 | 22 | 85.27 | 26.12 | 71.29 | 35.25 | 45.12 | 81 | 78 | 68.12 | 53.75 | 61.5 | 59.12 |
| hun_Latn | Hungarian | 32.08 | 94.88 | 76.88 | 23.35 | 30.34 | 18.75 | 3.38 | 31.71 | 26.88 | 27.72 | 53.25 | 55.62 | 56.12 | 65.5 | 55.38 | 53.62 | 60.5 | 71.62 |
| hunan | hunan | 78.53 | 76.62 | 56.12 | 21.22 | 79.03 | 13.75 | 16.38 | 81.02 | 24.38 | 64.92 | 59.5 | 60.12 | 74.25 | 71.38 | 59.88 | 50.12 | 59.62 | 68.62 |
| hye_Armn | Armenian | 29.21 | 79.75 | 53.25 | 23.72 | 28.34 | 23.5 | 31.38 | 31.09 | 25 | 30.59 | 31.88 | 46.88 | 24.25 | 21.12 | 21.38 | 21.88 | 33.12 | 34.88 |
| ibo_Latn | Igbo | 29.21 | 45 | 23.12 | 11.49 | 25.97 | 25.38 | 21.38 | 26.59 | 25.62 | 29.96 | 16.38 | 30.62 | 28.62 | 26.62 | 21.88 | 27.25 | 24.75 | 9 |
| ind_Latn | Indonesian | 78.03 | 95.12 | 69.5 | 23.85 | 41.7 | 28 | 19.12 | 79.65 | 33.12 | 36.08 | 68.5 | 64.5 | 82.25 | 82.38 | 66.25 | 60 | 68.12 | 81.88 |
| isl_Latn | Icelandic | 37.7 | 98 | 61.12 | 22.6 | 31.21 | 22.75 | 26.38 | 33.21 | 24.12 | 30.59 | 36.25 | 18.75 | 27.88 | 26 | 18.5 | 22.25 | 23.88 | 43.88 |
| ita_Latn | Italian | 85.89 | 96 | 73.5 | 24.72 | 57.43 | 16.12 | 24.12 | 85.02 | 24.25 | 54.18 | 64.88 | 56.12 | 86.25 | 87.25 | 67.5 | 55 | 91.38 | 79.12 |
| jav_Latn | Javanese | 42.95 | 95 | 51.12 | 19.98 | 30.96 | 19.75 | 15 | 41.45 | 24.12 | 31.34 | 25.62 | 15.38 | 29.62 | 35.5 | 35.38 | 26.25 | 55.75 | 13.75 |
| jiangxi | jiangxi | 81.77 | 93.38 | 50.5 | 23.35 | 81.27 | 18.38 | 15 | 83.27 | 25.62 | 68.29 | 55.62 | 58.88 | 72 | 75 | 56.5 | 47.62 | 64 | 65.62 |
| jpn_Jpan | Japanese | 73.28 | 83 | 81 | 24.97 | 33.21 | 24.38 | 16.75 | 75.41 | 24.88 | 59.93 | 66.38 | 48.75 | 74.5 | 75 | 54.5 | 49.62 | 89 | 69 |
| kan_Knda | Kannada | 33.71 | 88.5 | 38.62 | 23.35 | 35.71 | 16 | 13.75 | 33.33 | 24.12 | 30.59 | 38.25 | 20.38 | 23 | 24.12 | 22.88 | 18.75 | 41.88 | 23.88 |
| kat_Geor | Georgian | 30.71 | 85.12 | 56.25 | 21.47 | 27.72 | 23.88 | 18.38 | 30.84 | 24.38 | 29.59 | 44 | 33.38 | 29.25 | 26.25 | 26.62 | 28 | 28 | 26.38 |
| kaz_Cyrl | Kazakh | 33.21 | 85.25 | 53.62 | 23.97 | 32.58 | 29.75 | 45 | 36.2 | 25.25 | 28.96 | 50.75 | 45.12 | 30.12 | 21.88 | 28.62 | 22.5 | 35 | 33.25 |
| kea_Latn | Kabuverdianu | 52.31 | 88.62 | 34 | 21.22 | 40.32 | 23.12 | 13.88 | 50.81 | 24 | 35.71 | 41.25 | 45.62 | 37.88 | 35.88 | 31.25 | 27.5 | 38.12 | 38.38 |
| khk_Cyrl | Halh Mongolian | 30.71 | 75.12 | 18.5 | 19.85 | 28.21 | 16.62 | 23.5 | 26.72 | 24.62 | 28.34 | 25.5 | 20 | 29.38 | 22.62 | 29.25 | 26.75 | 19.88 | 11 |
| khm_Khmr | Khmer | 31.09 | 81.62 | 55.25 | 21.6 | 28.59 | 29.38 | 4.38 | 28.21 | 27.38 | 31.09 | 35.88 | 11.88 | 24 | 23.25 | 23.75 | 17.75 | 25.88 | 20.75 |
| kor_Hang | Korean | 76.15 | 90.25 | 73.88 | 22.1 | 35.58 | 28.88 | 28.38 | 79.78 | 28.38 | 48.81 | 64.12 | 58.62 | 80.12 | 79.25 | 72.5 | 55.62 | 71.38 | 78.38 |
| lao_Laoo | Lao | 36.33 | 81.5 | 29.88 | 23.72 | 27.84 | 19.5 | 13.25 | 36.95 | 23.88 | 30.34 | 39.75 | 16.5 | 32.5 | 32.38 | 30 | 30.88 | 46.25 | 20.12 |
| lit_Latn | Lithuanian | 38.2 | 86 | 73.12 | 23.72 | 34.33 | 16.25 | 24.75 | 38.83 | 26.12 | 30.21 | 63 | 41.25 | 27.12 | 28.25 | 30.5 | 30 | 63 | 80.38 |
| lug_Latn | Ganda | 29.59 | 59.25 | 19.88 | 13.48 | 28.84 | 27.62 | 30 | 26.22 | 32.25 | 27.22 | 24.5 | 20.38 | 18.62 | 15.25 | 28.38 | 23.88 | 24.88 | 35 |
| luo_Latn | Luo | 31.59 | 36.75 | 30.12 | 16.48 | 26.47 | 21.75 | 15.12 | 27.22 | 25 | 27.72 | 33.5 | 35.62 | 25.12 | 22.88 | 24.12 | 24.38 | 44 | 14 |
| lvs_Latn | Standard Latvian | 35.58 | 98 | 71.12 | 24.72 | 31.09 | 20.75 | 8.5 | 33.58 | 26 | 29.21 | 44.38 | 26.75 | 24.75 | 19.88 | 24.88 | 36.5 | 61.25 | 64.5 |
| mal_Mlym | Malayalam | 34.46 | 80.12 | 29.75 | 23.47 | 32.71 | 18.62 | 31 | 33.08 | 23.88 | 29.59 | 16.38 | 28.38 | 26.12 | 22.5 | 27.5 | 22.88 | 30.38 | 12 |
| mar_Deva | Marathi | 43.82 | 85.5 | 55.5 | 23.47 | 41.57 | 14.75 | 18.62 | 40.95 | 29.62 | 31.46 | 39.12 | 35.5 | 46.25 | 41.12 | 39.25 | 27.75 | 34.5 | 43.38 |
| minnan | minnan | 45.82 | 55.25 | 33.38 | 21.97 | 36.83 | 23.88 | 25.5 | 36.95 | 26.12 | 48.81 | 36.5 | 40.12 | 52.75 | 48.75 | 37.62 | 32.38 | 44.88 | 36.38 |
| mkd_Cyrl | Macedonian | 51.81 | 78.62 | 66 | 24.34 | 34.83 | 12.62 | 16.75 | 45.07 | 29.88 | 30.21 | 51.5 | 37 | 56.88 | 68.5 | 52.62 | 37.88 | 66 | 64.62 |
| mlt_Latn | Maltese | 40.7 | 83.5 | 16 | 23.47 | 35.33 | 27.88 | 25.88 | 36.08 | 24.12 | 31.34 | 28 | 41.25 | 30.88 | 21.88 | 30.38 | 28.25 | 25.12 | 25.25 |
| mya_Mymr | Burmese | 29.96 | 78 | 26 | 16.35 | 28.21 | 26.88 | 28 | 28.96 | 25.38 | 29.21 | 10.25 | 39.25 | 24.5 | 27.38 | 22.25 | 30.88 | 32 | 28.38 |
| ningxia | ningxia | 71.41 | 79.5 | 33.5 | 21.35 | 72.28 | 23.25 | 21.25 | 75.78 | 25.88 | 67.67 | 31.5 | 19.88 | 67.62 | 71.12 | 50.5 | 41.25 | 56.5 | 35.62 |
| nld_Latn | Dutch | 76.78 | 94.38 | 78.5 | 24.22 | 43.7 | 16.62 | 29.25 | 80.65 | 31.38 | 39.95 | 53.75 | 51.38 | 76 | 73.5 | 67.38 | 51.25 | 68.12 | 61.25 |
| nob_Latn | Norwegian Bokmål | 49.69 | 84 | 58.88 | 19.98 | 35.96 | 25 | 20.12 | 33.21 | 25.12 | 30.21 | 55.38 | 38.75 | 76.25 | 75.12 | 55 | 33.38 | 73 | 68.12 |
| npi_Deva | Nepali | 33.21 | 98 | 70.12 | 24.84 | 32.58 | 16.62 | 10.62 | 30.59 | 28.62 | 31.96 | 28 | 31 | 37 | 36 | 38.38 | 28.12 | 33.88 | 35.12 |
| ory_Orya | Odia | 40.57 | 86.38 | 49.38 | 22.35 | 33.21 | 15 | 26.25 | 34.46 | 24 | 31.09 | 26.75 | 41.38 | 22.88 | 23.5 | 22.88 | 31.62 | 36.5 | 18.12 |
| pan_Guru | Eastern Panjabi | 43.45 | 80.12 | 45.88 | 23.85 | 34.71 | 29.25 | 20.38 | 41.7 | 26 | 27.09 | 31.75 | 28.62 | 42.25 | 42 | 32.88 | 28.25 | 31.25 | 31.25 |
| pbt_Arab | Southern Pashto | 34.08 | 69.5 | 44 | 20.6 | 29.59 | 34 | 13.25 | 31.46 | 24.5 | 29.84 | 46.12 | 20.75 | 20.25 | 19.62 | 26.75 | 20.5 | 16.5 | 24.88 |
| pes_Arab | Western Persian | 36.45 | 95 | 59.38 | 23.47 | 33.46 | 34 | 19.88 | 34.08 | 22.12 | 29.96 | 43.38 | 35.88 | 62.88 | 57.75 | 42.75 | 35.38 | 68.12 | 60.62 |
| pol_Latn | Polish | 65.29 | 94.62 | 84.38 | 24.09 | 34.71 | 21 | 18.5 | 42.45 | 24.62 | 31.21 | 68.5 | 44.38 | 79.75 | 81.38 | 62.5 | 51.38 | 80.62 | 81.62 |
| por_Latn | Portuguese | 86.27 | 98 | 85.38 | 22.22 | 52.81 | 20.88 | 14.62 | 85.39 | 25.5 | 53.93 | 56.62 | 49.12 | 79.5 | 83.25 | 62.5 | 50.25 | 78.75 | 79.5 |
| ron_Latn | Romanian | 52.43 | 98 | 89.88 | 24.84 | 37.2 | 9 | 16 | 43.57 | 28.88 | 34.08 | 64.5 | 59.25 | 65.62 | 67.38 | 62.88 | 52.25 | 80.75 | 94.38 |
| rus_Cyrl | Russian | 85.64 | 93.38 | 84.88 | 24.59 | 46.69 | 6.62 | 24.5 | 84.39 | 33.88 | 44.44 | 59.38 | 53.75 | 85.38 | 85.88 | 69.88 | 56.25 | 83.25 | 83.62 |
| shan1xi | shan1xi | 72.91 | 78.25 | 59.88 | 17.85 | 73.66 | 17 | 26.25 | 78.78 | 24 | 65.79 | 56.75 | 58.88 | 69 | 72.12 | 55.88 | 49.62 | 59.88 | 76.62 |
| shan3xi | shan3xi | 77.65 | 78.88 | 49 | 18.98 | 76.65 | 20.75 | 23 | 79.65 | 25.12 | 74.03 | 44.12 | 34.38 | 75.75 | 72.12 | 60.5 | 57.5 | 50.88 | 49.62 |
| shandong | shandong | 79.28 | 78.5 | 60 | 16.73 | 84.02 | 28.12 | 18.38 | 82.15 | 24.12 | 70.79 | 46.25 | 45.38 | 77.12 | 80.88 | 54.38 | 51 | 75.12 | 75.25 |
| shanghai | shanghai | 63.92 | 63.75 | 30.12 | 22.97 | 54.81 | 25.12 | 15.88 | 66.42 | 25.38 | 57.18 | 41.5 | 38 | 54.5 | 55.62 | 38 | 41.12 | 29.75 | 13.62 |
| sichuan | sichuan | 80.15 | 75.88 | 48.62 | 23.47 | 83.27 | 23.62 | 20.88 | 82.27 | 26 | 70.54 | 48 | 44.25 | 71.5 | 73.75 | 56.12 | 46.38 | 59.38 | 54.12 |
| sin_Sinh | Sinhala | 34.08 | 60 | 30.88 | 23.35 | 33.58 | 23.38 | 14.38 | 34.46 | 25.38 | 30.09 | 14.38 | 36.25 | 27.5 | 24.62 | 24.25 | 19.62 | 26.62 | 28.25 |
| slk_Latn | Slovak | 56.93 | 96.75 | 84 | 22.97 | 38.83 | 10.25 | 18.5 | 40.95 | 24.5 | 30.21 | 69 | 46.62 | 54.62 | 60.25 | 57.5 | 45.12 | 73.38 | 79.75 |
| slv_Latn | Slovenian | 45.32 | 98 | 74.5 | 24.22 | 37.08 | 17.75 | 14.38 | 37.33 | 23.62 | 29.59 | 59.12 | 50.88 | 29.25 | 32.5 | 34.25 | 28.38 | 68.5 | 64.62 |
| sna_Latn | Shona | 31.59 | 61.12 | 3.75 | 16.48 | 27.09 | 24.75 | 21.38 | 33.83 | 24 | 29.21 | 9.88 | 5.38 | 26.5 | 21 | 28 | 27.88 | 30.38 | 13.88 |
| snd_Arab | Sindhi | 36.7 | 81.38 | 24.88 | 20.72 | 31.59 | 30.5 | 20.25 | 34.46 | 25.62 | 31.96 | 14.5 | 18.25 | 25.62 | 23.5 | 23.75 | 20.88 | 40 | 2 |
| som_Latn | Somali | 28.71 | 73 | 15.12 | 23.35 | 26.84 | 22.88 | 8.25 | 30.34 | 24.88 | 28.84 | 16.12 | 30.88 | 27.12 | 25 | 24.88 | 37.75 | 37 | 25.5 |
| spa_Latn | Spanish | 79.15 | 95.5 | 70.38 | 21.22 | 55.06 | 20.88 | 36.62 | 79.53 | 24.38 | 53.93 | 60.5 | 49.25 | 76.38 | 79.88 | 62.5 | 56.12 | 89.25 | 90.88 |
| srp_Cyrl | Serbian | 54.43 | 96 | 82 | 24.84 | 40.32 | 15.38 | 31.5 | 43.82 | 25.62 | 31.21 | 58.12 | 48.12 | 55.5 | 70.62 | 49.38 | 44.75 | 83.38 | 85.25 |
| suhang | suhang | 65.29 | 65.38 | 38.62 | 22.22 | 60.92 | 29.88 | 20.62 | 71.54 | 26 | 59.18 | 19 | 30.5 | 56.38 | 58.62 | 49.25 | 43.75 | 23.5 | 35 |
| sun_Latn | Sundanese | 39.58 | 79.25 | 31.38 | 24.09 | 34.33 | 28.62 | 28.38 | 39.08 | 23 | 30.21 | 24.12 | 23 | 32.88 | 33.62 | 37.5 | 28.38 | 38.88 | 24.38 |
| swe_Latn | Swedish | 57.43 | 95.88 | 71.75 | 23.72 | 44.07 | 23.62 | 16.88 | 37.95 | 24 | 32.46 | 64.12 | 53.5 | 75.88 | 79.5 | 61 | 48.38 | 84.75 | 83.88 |
| swh_Latn | Swahili | 34.33 | 76 | 26.75 | 19.1 | 30.46 | 19.5 | 29.12 | 32.58 | 25.25 | 26.09 | 31.5 | 30.12 | 33.5 | 36 | 38.75 | 27.5 | 31.88 | 9 |
| tam_Taml | Tamil | 33.83 | 80.38 | 38.12 | 22.47 | 34.08 | 25.75 | 17 | 32.21 | 25.38 | 27.59 | 38.25 | 39 | 33.5 | 29.88 | 28.75 | 23.12 | 40.88 | 36.62 |
| tel_Telu | Telugu | 34.58 | 88.5 | 40.5 | 22.72 | 33.71 | 22 | 19.38 | 32.83 | 25.25 | 29.21 | 19.5 | 36.12 | 27.38 | 27.62 | 30 | 21.62 | 16.5 | 29.25 |
| tgk_Cyrl | Tajik | 36.58 | 80.5 | 35.12 | 23.22 | 27.84 | 36.5 | 10.38 | 34.83 | 24.25 | 29.21 | 18.88 | 25.62 | 38.12 | 31.25 | 39.25 | 26.62 | 28.62 | 15.75 |
| tgl_Latn | Tagalog | 53.43 | 95.62 | 83.75 | 22.97 | 47.07 | 21.88 | 30 | 41.32 | 25 | 30.34 | 51.38 | 24.62 | 47.75 | 59.25 | 42.88 | 31.75 | 51.75 | 48.5 |
| tha_Thai | Thai | 71.04 | 90.12 | 78 | 25.09 | 36.33 | 22.12 | 15.5 | 69.16 | 23.88 | 29.21 | 58.38 | 50.38 | 81.38 | 74.88 | 62.12 | 48.38 | 93.88 | 75.38 |
| tianjin | tianjin | 84.89 | 98 | 73.62 | 15.48 | 86.77 | 21.75 | 15.12 | 87.02 | 26.12 | 72.53 | 63.62 | 71.5 | 77.75 | 79.12 | 71.62 | 62.38 | 73.62 | 81.5 |
| tur_Latn | Turkish | 54.43 | 89.62 | 73 | 23.97 | 36.95 | 22.88 | 24.62 | 42.07 | 25.75 | 29.46 | 65.38 | 40.5 | 70.62 | 77.62 | 54.88 | 50.25 | 75.38 | 84.88 |
| ukr_Cyrl | Ukrainian | 55.68 | 96.62 | 81.75 | 24.47 | 38.83 | 22.38 | 36.75 | 48.56 | 24.38 | 32.21 | 58.12 | 48.12 | 59.12 | 70 | 53.75 | 44.38 | 64 | 81.12 |
| urd_Arab | Urdu | 55.68 | 93.88 | 74 | 24.09 | 47.19 | 18.25 | 13.75 | 56.3 | 31.12 | 31.84 | 26.38 | 39.25 | 49.62 | 63.5 | 54.75 | 40.5 | 31.38 | 45.88 |
| uzn_Latn | Northern Uzbek | 37.83 | 98 | 69.12 | 24.09 | 36.83 | 17.38 | 25.5 | 36.2 | 30.38 | 29.34 | 25 | 34.75 | 33.75 | 29.88 | 28.75 | 24.88 | 36.12 | 8.38 |
| vie_Latn | Vietnamese | 75.91 | 89.62 | 64.25 | 23.35 | 28.46 | 21.25 | 21.38 | 77.03 | 28.25 | 29.71 | 51.5 | 53.75 | 73.88 | 66.38 | 60.88 | 55.88 | 73.38 | 60.25 |
| wol_Latn | Wolof | 29.34 | 74 | 25.25 | 15.61 | 26.84 | 22.5 | 26.25 | 26.84 | 24.5 | 30.46 | 19.38 | 16.25 | 22.5 | 20.38 | 27.62 | 17.25 | 19.88 | 30.75 |
| wuzhong | wuzhong | 85.27 | 95.38 | 76.5 | 19.73 | 85.89 | 19.88 | 17 | 86.39 | 25.88 | 74.16 | 65.25 | 58.38 | 75.5 | 81.62 | 66.5 | 52 | 79.5 | 84.12 |
| xho_Latn | Xhosa | 28.46 | 65.62 | 17 | 14.86 | 26.97 | 19.5 | 29.62 | 29.21 | 25.38 | 29.84 | 30.62 | 26.5 | 22.5 | 23.62 | 30.75 | 27.75 | 8.25 | 3.62 |
| yor_Latn | Yoruba | 29.84 | 59.12 | 2 | 12.86 | 28.34 | 34.25 | 23.12 | 29.46 | 25.62 | 29.21 | 41.25 | 13.38 | 23.88 | 26.25 | 30.25 | 19.12 | 13.25 | 25.88 |
| yunnan | yunnan | 78.4 | 90.5 | 56 | 23.85 | 78.4 | 22.75 | 10.5 | 80.52 | 24.38 | 68.54 | 50.38 | 38.25 | 74 | 80 | 55.75 | 50.62 | 70.12 | 61 |
| zho_Hans | Chinese (Simplified) | 87.02 | 98 | 74.12 | 22.47 | 87.02 | 23 | 18.5 | 87.39 | 25.88 | 70.04 | 65.38 | 71.12 | 86.62 | 86.88 | 68 | 57.5 | 81.5 | 83.25 |
| zho_Hant | Chinese (Traditional) | 87.39 | 88.88 | 98 | 15.98 | 85.52 | 19.12 | 26.38 | 86.89 | 31 | 74.91 | 68.12 | 59.88 | 79 | 83 | 74.12 | 56.75 | 61.12 | 63.5 |
| zsm_Latn | Standard Malay | 71.04 | 98 | 74.5 | 23.6 | 55.68 | 14.62 | 20 | 66.67 | 25.75 | 32.33 | 48.38 | 44.25 | 78.75 | 84.25 | 62.38 | 50.88 | 75.5 | 85.75 |
| zul_Latn | Zulu | 30.84 | 56.75 | 35.75 | 21.97 | 29.84 | 28.75 | 28 | 29.09 | 25.88 | 30.59 | 21 | 31 | 16.62 | 21.12 | 30.75 | 19.75 | 19.62 | 9 |
| | Average | 52.88 | 85.30 | 56.63 | 21.88 | 43.51 | 21.83 | 20.96 | 50.89 | 25.94 | 40.52 | 43.59 | 39.01 | 52.29 | 52.66 | 45.00 | 38.21 | 53.86 | 51.56 |