

IMPACT-SPEECH 2026

**IMPACT-SPEECH: Identifying, Measuring, Preventing, and
Assessing Consequences of Bias in Speech LLMs**

Proceedings of the Workshop

October 29, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-479-8

Preface

This is the first edition of the IMPACT-SPEECH Workshop, co-located with the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026) and held on October 29, 2026, in Budapest, Hungary.

We are excited to welcome attendees to the IMPACT-SPEECH 2026 Workshop on Identifying, Measuring, Preventing, and Assessing Consequences of Bias in Speech LLMs. The workshop brings together researchers and practitioners working across speech processing, natural language processing, machine learning, fairness, and responsible AI to examine the growing challenges associated with bias in speech-enabled large language models and multimodal speech technologies.

As speech AI systems become increasingly integrated into real-world applications, they interact with speakers from diverse linguistic, demographic, cultural, and social backgrounds. These interactions raise important questions about how bias in such systems can be identified and measured, how developers can prevent or mitigate it, and how biased system behavior can affect individuals and communities. IMPACT-SPEECH 2026 provides a forum for researchers to discuss these questions and share approaches for building fairer, more inclusive, and reliable speech technologies.

This year’s program features 22 archival and non-archival papers, including 8 oral and 14 poster presentations, alongside 2 invited talks. Together, these contributions cover a broad range of topics related to bias and fairness in speech LLMs. We are grateful to all authors and participants for contributing their research, ideas, and perspectives to the workshop.

We also extend our sincere thanks to the EMNLP 2026 organizers for providing the opportunity to hold IMPACT-SPEECH alongside the conference and for their considerable efforts in creating an environment that supports discussion and collaboration across the NLP and speech communities.

Finally, we thank all members of the IMPACT-SPEECH 2026 Program Committee for their time, expertise, and thoughtful reviews. Their contributions have helped us curate a diverse and engaging program. We hope that this workshop will stimulate new collaborations, encourage further research, and contribute to the development of more equitable and responsible speech AI.

The IMPACT-SPEECH 2026 Organizing Committee: Ravi Shekhar, Monorama Swain, Jagabandhu Mishra, Sandipan Dhar, Haralambos Mouratidis, and Matthew Purver

Organizing Committee

Workshop Organizers

Ravi Shekhar, University of Essex, UK

Monorama Swain, JKU Linz, Austria

Jagabandhu Mishra, University of Eastern Finland, Finland

Sandipan Dhar, Fraunhofer IIS, Germany

Haralambos Mouratidis, University of Essex, UK

Matthew Purver, QMUL, UK and JSI, Slovenia

Program Committee

Program Committee

Badr M. Abdullah, Saarland University, Germany
Mohd Mujtaba Akhtar, IIIT Delhi, India
Iqra Ali, Queen Mary University of London, UK
Kalluri Shareef Babu, Manipal Institute of Technology Bengaluru
Alessio Brutti, Fondazione Bruno Kessler (FBK), Italy
Sergio Burdisso, Idiap Research Institute, Switzerland
Shammur Absar Chowdhury, Qatar Computing Research Institute (QCRI), Qatar
Lorenzo Concina, Fondazione Bruno Kessler (FBK), Italy
Sudhansu Bala Das, University of Galway, Ireland
Giovanni Filomeno, JKU Linz, Austria
Girish, IIIT Delhi, India
Reyhaneh Hashempour, QA Ltd, UK
Daniel Hershcovich, University of Copenhagen, Denmark
Nina Hosseini-Kivanani, Radio Télévision Luxembourg (RTL), Luxembourg
Parisa Jamadi, Imperial College London, UK
Vanja Karan, University of Vienna, Austria
Santosh Kesiraju, Brno University of Technology, Czechia
Sandeep Kumar, MBZUAI, UAE
Spyretta Leivaditi, University of Groningen, Netherlands
Yunfei Long, Queen Mary University of London, UK
Fernando López, Telefónica Innovación Digital, Spain
Ben Malin, Brunel University, UK
Shah Nawaz, Johannes Kepler University Linz, Austria
Shekhar Nayak, University of Groningen, Netherlands
Ludovica Pannitto, University of Bologna, Italy
Tanvina Patel, TU Delft, Netherlands
Orchid Chetia Phukan, IIIT Delhi, India
Senja Pollak, Jožef Stefan Institute, Slovenia
Priyanshu Priya, IIT Patna, India
Tariq Rahim, Kingston University London, UK
Tharindu Ranasinghe, Lancaster University, UK
Subhankar Roy, University of Bergamo, Italy
Markus Schedl, Johannes Kepler University Linz, Austria
Divya Sharma, IIIT Delhi, India
Samara Souza, Federal University of Pará, Brazil
Bo Su, Indiana University Bloomington, USA
Abhirama Subramanyam, MBZUAI Abu Dhabi
Richard Sutcliffe, University of Essex, UK
Eva Szekely, KTH Royal Institute of Technology, Sweden
Venktesh V, Stockholm, Sweden
Esaú Villatoro Tello, Idiap Research Institute, Switzerland
Sarenne Wallbridge, UK AI Security Institute, UK
Tariq Yousef, University of Southern Denmark, Denmark
Juntao Yu, Queen Mary University of London, UK
Yuchen Zhang, University of Essex, UK

Keynote Talk

Every Voice Matters: Towards Inclusive Speech LLMs

Marco Matassoni

Fondazione Bruno Kessler (FBK), Trento, Italy



Abstract: The increasing use of Speech Large Language Models raises fundamental questions about fairness, inclusiveness, and trustworthiness. This talk presents research conducted within the ELOQUENCE project, which aims to develop multilingual, context-aware, and bias-controlled conversational AI for applications ranging from customer service and smart homes to safety-critical scenarios. We will focus on inequalities in multilingual and low-resource speech processing, where differences in data availability, language coverage, accents, and speaker characteristics can lead to significant performance disparities. The talk will discuss approaches to identify and measure speech bias, as well as strategies to improve robustness and inclusiveness through cross-lingual adaptation, data-efficient learning, and privacy-preserving approaches.

Bio: Marco Matassoni is a Senior Researcher at the Fondazione Bruno Kessler (FBK) in Trento, Italy, where he plays a leading role within the Speech Technology (SpeechTeK) research unit of the Centre for Augmented Intelligence. Over nearly three decades, he has become an experienced figure in the field of speech processing, contributing to advancements in automatic speech recognition (ASR) and human-machine interaction systems. Marco Matassoni graduated in Electronic Engineering from the University of Padua in 1995, where his thesis focused on microphone arrays, a topic that would later define part of his professional research. In 1996, he joined ITC-IRST, the research institute that later became part of FBK, and began working on speech recognition technologies. His research was driven by the challenge of making speech systems perform robustly in real-world environments marked by noise, reverberation, and distance between speakers and microphones. During his career, he has contributed to a wide range of European research projects, including SpeechDat-Car, VICO, HIWIRE, DICIT, DIRHA. These initiatives focused on developing speech interfaces for complex acoustic conditions such as in-car communication or home automation. His work on distant-talking and far-field speech recognition has been particularly influential, especially on data augmentation techniques. In recent years, he has expanded his focus to include multilingual and low-resource ASR, speaker verification and anonymization, and bias-aware speech technologies; in the ELOQUENCE project, he is working on multilingual and cross-lingual speech recognition, addressing the challenges of building scalable models for European languages. These efforts reflect a growing concern with inclusive artificial intelligence, ensuring that speech technologies respect linguistic diversity and privacy. In addition, he regularly serves as a reviewer for major conferences and journals in the field, contributing his expertise to the peer-review process and supporting the dissemination of high-quality research.

Keynote Talk

Bias before the model: Reference noise as source of unfair speech systems

Jesper Rindom Jensen
Aalborg University



Abstract: Bias in speech systems is often attributed to the models themselves, but it can be introduced long before training begins. In this talk, I will use speech enhancement as a case study, focusing on one such example: the clean references used for training are, in practice, often noisy. Drawing on our recent research, we will see how this can bias models to let noise pass through while still scoring well on objective metrics. We will also examine current evaluation practices, showing that common metrics struggle to differentiate between state-of-the-art methods and often fail to reflect the preferences and perception of real users.

Bio: Jesper Rindom Jensen was born in Ringkøbing, Denmark in August 1984. He received the M.Sc. degree cum laude in 2009 and the Ph.D. degree in 2012, both from Aalborg University, Denmark. He is currently an Associate Professor at the Department of Electronic Systems at Aalborg University, where he is enrolled in the Professor Promotion Programme and serves as Head of the Doctoral Programme in Electrical and Electronic Engineering at the Technical Doctoral School of IT and Design. He is also Head of Research at the Audio Analysis Lab, which he has been a member of since its foundation in 2012. Before this, he was with the Department of Architecture, Design & Media Technology at Aalborg University as a Postdoc, Assistant Professor, and Associate Professor. He has been a Visiting Researcher at the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, at the Friedrich-Alexander Universität Erlangen-Nürnberg in Erlangen, Germany, and at the University of Surrey, UK. His research interests include speech and audio processing, multichannel and array processing, robot and drone audition, spatial audio, estimation theory, noise reduction, and machine learning for signal processing. He has published more than 120 papers in top-tier, peer-reviewed conference proceedings and journals, holds a patent, and is the co-author of two books, namely, *Speech Enhancement – A Signal Subspace Perspective* and *Signal Enhancement with Variable Span Linear Filters*. He has led a number of externally funded research projects, including as Principal Investigator and Coordinator of a European Defence Fund project on AI-based natural language processing. Dr. Jensen has received a highly competitive individual postdoc grant from the Independent Research Fund Denmark, has been selected as an AAU Talent at Aalborg University (awarded to young research talents), was named Teacher of the Year by the Study Board of Media Technology, and has received several travel grants from private foundations. He is a member of the IEEE Speech and Language Processing Technical Committee and of the EURASIP Technical Area Committee on Acoustic, Speech, and Music Signal Processing, and is a Senior Member of the IEEE. He currently served as an Associate Editor of the Elsevier Signal Processing Open journal and has previously served as a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, and as an Affiliate Member of the IEEE Signal Processing Theory and Methods Technical Committee, and he has served as area or track chair for ICASSP, WASPAA, and EUSIPCO.

Table of Contents

<i>A Probe Shift Is Not a Fairness Fix: The Limits of Representation Steering in Speech Models</i> Nicolas Bourrel, Abderrahmane Issam and Gerasimos Spanakis	1
<i>AIP-Speech: Auditing Speech LLM Robustness to Background Interference</i> Aiden Dsouza, Shreyansh Mishra, Yashraj Suneja, Rohan Swain, Samyak Borkar, Arjun Jaishankar and Tanmay Urade	10
<i>Can Large Audio-Language Models Ignore Multilingual Distractors? An Evaluation of Their Selective Auditory Attention Capabilities</i> Heejoon Koo	20
<i>Early Warning for Target-Speaker Extraction: Reliability under Distribution and Failure-Definition Shift</i> Shubham Sharma and Manik Gupta	33
<i>Evaluating Bias in Phoneme-Based Automatic Speech Recognition Systems: An Analysis of IPA Transcription Models</i> Maneesha Rani Saha, Catherine Bao and Neal Patwari	42
<i>Lost in Compression: Calibration Bias and the Accent-WER Gap in Speech LLMs</i> Angad Ripudaman Singh Bajwa	61
<i>Pruning for Efficiency, Paying in Fairness: Demographic Disparities in Pruned Speech-LLMs</i> Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis and Ravi Shekhar	67
<i>Temporal Taxation Compounds Under Post-Training Compression of Whisper Models</i> Srishti Ginjala, Eric Fosler-Lussier, Christopher W. Myers and Srinivasan Parthasarathy	81
<i>TurkSV-Bench: Assessing Demographic Bias and Preprocessing Effects in Speaker Verification</i> Ahmet Berkay Kantarcı, Abdulkadir Parlak and Ahmet Karazor	95
<i>What Drives Dialectal Jailbreaks? An Ablation of Surface Form, Cultural Framing, and Strategy Banks</i> Qingyang Xu	104

A Probe Shift Is Not a Fairness Fix: The Limits of Representation Steering in Speech Models

Nicolas Bourrel Abderrahmane Issam Gerasimos Spanakis

Department of Advanced Computing Sciences

Maastricht University

{n.bourrel@student., abderrahmane.issam@, jerry.spanakis@}maastrichtuniversity.nl

Abstract

Automatic speech recognition (ASR) systems exhibit unequal error rates across speaker groups, motivating interventions on their internal representations. We ask whether speaker-linked attributes that are linearly readable from pretrained ASR encoders yield useful directions for reducing group word-error-rate (WER) gaps. Across Whisper-medium, HuBERT-large, and Wav2Vec2-large on Common Voice and the Speech Accent Archive, we probe every encoder layer for metadata-derived sex/gender, age, and native/accent labels; construct centroid and probe-derived directions; inject them at selected layers; and compare downstream probe trajectories with matched WER changes. Sex labels are highly decodable (best macro-F1 0.924–0.941), native/accent labels are also above chance (0.544–0.696), and age is weaker (0.354–0.397). Of 22 post-selected reruns, nine have 95% paired-bootstrap intervals entirely below zero, yet every absolute source-group WER reduction is below 0.7 percentage points. Conversely, a local target-class probe rate can rise from 8.09% to 99.87% while WER worsens. Linear readability is therefore neither evidence of causal use nor a reliable mitigation method. Our results motivate evaluating speech-bias interventions jointly at representation, propagation, and task levels.

1 Introduction

Automatic speech recognition (ASR) can perform unevenly across accents, genders, languages, ages, and other speaker-linked groups (Koenecke et al., 2020; Feng et al., 2021; Jahan et al., 2025). Such disparities can reduce access to voice interfaces and compound exclusion in education, employment, public services, and assistive technology. Prior audits document group-level error gaps and show that speech encoders retain information associated with demographic metadata (Attanasio et al., 2024; Lin et al., 2024; Herron et al., 2026b,a). Together,

these findings suggest an appealing intervention: if a disadvantaged group’s attribute is visible inside an ASR encoder, move its hidden representation toward a lower-error comparison group.

That proposal combines three different claims. A probe can show that an attribute is *readable*; an intervention can make later probes read a state as more target-like; and the recogniser can produce fewer word errors. None logically entails the next. Probing work has long cautioned that decodability does not establish how a model uses a feature (Belinkov, 2022; Kumar et al., 2022). For fairness interventions the distinction matters especially: a nearly complete demographic-probe shift may leave recognition unchanged, damage linguistic content, or require knowing a sensitive label at inference time.

We test this chain in three frozen ASR systems with different objectives and decoders: Whisper (Radford et al., 2023), HuBERT large (Hsu et al., 2021), and Wav2Vec2 (Baevski et al., 2020). On Mozilla Common Voice (MCV) (Ardila et al., 2020) and the Speech Accent Archive (SAA) (Weinberger, 2015), we first measure layer-wise linear access to metadata-derived sex/gender, age, and native/accent labels using linear probes. Then we apply activation steering from higher to lower Word Error Rate (WER) classes using contrastive steering (Marks and Tegmark, 2024; Rinsky et al., 2024) with the centroid difference or using the linear probe direction (Hedström et al., 2025; Chen et al., 2026). Finally, we evaluate the intervention at two downstream levels: whether its target-like shift propagates through later layers, and whether the steered model’s generation achieves a lower WER.

This paper makes three contributions. First, it provides a common layer-wise audit across three 24-block ASR encoders and two speech corpora, keeping the metadata, representation, and task measurements explicit. Second, it connects descriptive

probing to a controlled representation intervention without retraining the ASR model. Third, it reports the complete set of 22 selected follow-up reruns, including null and adverse cases, and isolates a failure mode in which a probe shift is large but ASR does not improve.

The central result is cautionary. Speaker-linked labels are often strongly readable, and fixed directions can produce dramatic target-like shifts, but WER effects are small and unstable across layers. Nine selected reruns have paired-bootstrap intervals below zero, yet none reduces source-group WER by one point. One Wav2Vec2 setting makes 99.87% of source clips target-like to the next-layer probe while WER worsens. We therefore present the method as a stress test for representation-based mitigation, not as a fairness fix.

2 Related Work

2.1 Speech representations and probes

Wav2Vec2 learns masked speech representations through contrastive prediction (Baevski et al., 2020); HuBERT predicts clustered hidden units under masking (Hsu et al., 2021); and Whisper uses large-scale weak supervision in an encoder-decoder recogniser (Radford et al., 2023). Their objectives and decoding paths differ, but all can retain information beyond lexical content. Wav2Vec2 features transfer to speaker verification and language identification (Fan et al., 2021), and layer-wise analyses show that acoustic and linguistic content is distributed non-monotonically through speech encoders (Belinkov and Glass, 2017; Pasad et al., 2021).

Linear probes offer a deliberately simple test of whether a label is accessible from a hidden state (Alain and Bengio, 2017; Belinkov, 2022). Speech probes recover phonetic, speaker, environment, emotion, and other information (Raymond et al., 2024). However, a successful probe can use a correlate that the base model ignores, and classifier performance depends on representation, probe capacity, controls, and data construction. We therefore interpret macro-F1 as linear accessibility, not a causal statement about ASR behaviour.

2.2 Fairness and representation intervention

Speech fairness studies measure error disparities and the encoding or propagation of social information. For example, multilingual ASR systems exhibit gender-related performance gaps whose rela-

tionship to internal separability varies by language and model (Attanasio et al., 2024); self-supervised speech encoders can also encode social biases affected by architecture and compression (Lin et al., 2024). These results motivate internal analysis but do not imply that demographic separability is the cause of an observed WER gap.

Nearby intervention work has stronger machinery than the fixed vectors studied here. Speaker-adaptation systems learn or retrieve speaker information to improve ASR (Sari et al., 2020). Privacy-oriented models adversarially conceal attributes while preserving speaker verification (Noé et al., 2021). Other work identifies speaker subspaces and removes them to reduce speaker identifiability while retaining phonetic information (Liu et al., 2023). Our intervention is intentionally simpler: no model parameters are updated, and a single metadata-derived vector is added at one layer. This makes it a transparent baseline for asking whether readability, movement, and task benefit actually align.

3 Experimental Design

3.1 Data and operational labels

MCV is a crowd-sourced multilingual speech corpus with speaker-contributed recordings and metadata (Ardila et al., 2020). We use English clips and construct task-specific subsets for sex/gender, age, and native/accent labels. Each task begins with 100,000 training identifiers and 1,500 test identifiers before task-label filtering. The five-class age summary retains 96,223 labelled training clips and 1,237 labelled test clips; sex/gender and native/accent each retain all 1,500 test clips. The generated MCV subsets have no shared `client_id` values between train and test.

SAA contains speakers reading a shared English passage and provides native-language and demographic metadata (Weinberger, 2015). We use a deterministic 80/20 metadata split, yielding 1,710 training and 428 test items after the available-audio and label filters. The split code operates on filename stems rather than enforcing an explicit grouped-speaker split; the archive metadata nevertheless has one `speakerid` per row. SAA is valuable for accent comparisons because lexical content is fixed, but several native-language and older-age groups are small.

Table 1 summarises the evaluation sets. The labels are operational groupings from corpus meta-

data, not complete measurements of identity, biology, accent, or linguistic background. The source-to-target notation used below only encodes an observed error-rate contrast.

3.2 Models and activation extraction

Whisper-medium¹ is an autoregressive encoder-decoder system over log-Mel features. We pad or truncate its input to 3,000 frames and record encoder hidden states. HuBERT-L² and Wav2Vec2-L³ are CTC recognisers over 16 kHz waveforms; audio is capped or zero-padded to 30 seconds. Each model exposes 25 hidden-state indices in the local Hugging Face interface: index 0 is the state before the first transformer block and indices 1–24 are block outputs. All final hidden widths are 1,024, and no ASR parameters are trained.

For the main analysis, we average each hidden sequence over time to obtain one vector per clip and layer. This makes layer-wise probing and direction estimation tractable across large subsets, but it discards temporal structure. The saved extraction pipeline also averages padded positions without a mask, a limitation revisited below.

3.3 Layer-wise probing

For every model, task, and hidden-state index, we standardise features with a `StandardScaler` fitted on training data only and train one linear head. The sex/gender probe is binary; age and native/accent probes initially retain all selected classes. The MCV probes use Adam with learning rate 10^{-3} , weight decay 10^{-8} , weighted sampling, a 10% validation split, and early stopping within 2,000 epochs. SAA binary probes use learning rate 3×10^{-4} , weight decay 10^{-4} , dropout 0.1, and longer patience. All reported runs use seed 42.

We report macro-F1 because it gives each class equal weight even when test supports differ. Uniform-random macro-F1 is computed using the observed supports and serves as a simple reference. Probe performance is used to describe accessibility and, in some exploratory analyses, to compare layer rankings; it does not select every final steering layer.

¹<https://huggingface.co/openai/whisper-medium>

²<https://huggingface.co/openai/whisper-medium>

³<https://huggingface.co/facebook/wav2vec2-large-960h>

3.4 Baseline ASR and group gaps

Each clip is first decoded with the unmodified recogniser. References and hypotheses are lowercased, stripped of punctuation and symbols, and whitespace-normalised before WER. Whisper uses 30-second chunks, six-second stride, and explicit English transcription generation. HuBERT and Wav2Vec2 use their default CTC decoding paths without an external language model.

For each intervention family, two classes are selected: a source class with higher baseline WER and a lower-WER target class with sufficient support. Examples include MCV teens→forties and India/South-Asia→England, and SAA Spanish→English. The comparison is not counterfactual: recording conditions, utterance difficulty, speaking rate, and other factors may contribute to the gap.

3.5 Steering directions

Let $h_i^\ell \in \mathbb{R}^{1024}$ be the mean-pooled training activation of clip i at hidden-state index ℓ , and let S and T denote source and target classes. The centroid direction is:

$$v_{\text{ctr}}^\ell = \frac{1}{|T|} \sum_{i \in T} h_i^\ell - \frac{1}{|S|} \sum_{i \in S} h_i^\ell. \quad (1)$$

For the probe direction, let $w_T - w_S$ be the weight difference of a binary probe trained on standardised features and let σ contain the training feature standard deviations. We map the direction back to activation coordinates as

$$v_{\text{probe}}^\ell = (w_T - w_S) \oslash \sigma, \quad (2)$$

where $\oslash$ is element-wise division. This conversion accounts for the physical scale hidden by standardisation.

At inference time, a forward pre-hook adds the normalised direction to every time step at one encoder layer:

$$H^\ell \leftarrow H^\ell + \alpha \frac{v^\ell}{\|v^\ell\|_2}. \quad (3)$$

Only clips with the known source metadata label receive the intervention; target clips are decoded in the same run but are not pushed. This sample conditioning makes the experiment a mechanism diagnostic rather than a deployable correction.

Corpus	Task	Classes or largest groups	Test n	Test distribution
MCV	Sex/gender	male, female	1,500	779 / 721
	Age	teens, twenties, thirties, forties, fifties	1,237	263 / 262 / 262 / 263 / 187
	Native/accent	England, US, India/S. Asia, Canada, Australia	1,500	404 / 403 / 403 / 162 / 128
SAA	Sex	male, female	428	233 / 195
	Age	decade bins from 0–9 to 80–89	428	largest: 20–29 (177), 30–39 (95)
	Native language	English, Spanish, Arabic, Mandarin, others	428	largest: 117 / 36 / 22 / 15

Table 1: Evaluation subsets. MCV age has 1,500 identifiers but 1,237 valid five-bin labels. SAA native language has many small classes; only the largest are listed.

3.6 Selection, uncertainty, and trajectories

Exploratory sweeps vary model, corpus, class pair, vector type, layer, strength, and direction. Promising settings and explicit probe-best or adverse controls are then decoded again with matched baseline and steered outputs. For source clips we report

$$\Delta_S = \text{WER}_{\text{steered},S} - \text{WER}_{\text{base},S}, \quad (4)$$

so negative values indicate fewer errors. Each of the 22 targeted reruns uses 2,000 paired-bootstrap resamples and a 95% percentile interval. The descriptive q_{boot} is the fraction of resampled deltas that are non-negative. It is neither a null-centred p -value nor a correction for the earlier search.

Trajectory diagnostics rerun the encoder with and without the hook and apply saved binary probes to each downstream layer. For true source clips, we track mean target probability and the target rate, i.e., the fraction classified as the target class. These quantities test whether a local intervention propagates in probe space; they do not show that a speaker attribute has literally changed.

4 Results

4.1 Metadata is readable at different depths

Table 2 reports the best MCV macro-F1 across 25 hidden-state indices. Sex/gender is highly recoverable in all three encoders, with best F1 from 0.924 to 0.941. Native/accent is also substantially above chance, especially in Whisper. Five-class age is weaker but remains above the support-aware uniform-random baseline.

The best depth depends on both task and architecture. Whisper peaks at indices 9, 13, and 17 for sex/gender, age, and native/accent; HuBERT peaks at 12, 8, and 13; Wav2Vec2 peaks at 10, 6, and 11. There is therefore no single layer that contains all speaker-linked information most cleanly. The age result should be read as recoverable age-correlated variation under coarse metadata bins, not as a robust age estimator.

Task	Whisper	HuBERT	W2V2	Chance
Sex/gender	9/.941	12/.939	10/.924	.500
Age	13/.397	8/.375	6/.354	.199
Native/accent	17/.696	13/.649	11/.544	.190

Table 2: Best MCV macro-F1. Cells give hidden-state index/F1. Chance is support-aware uniform-random macro-F1.

Corpus	Pair/task	Whisper	HuBERT	W2V2
MCV	teens/forties	6.437	9.871	14.942
MCV	India/England	3.151	9.296	16.390
MCV	male/female	.764	2.379	3.545
SAA	30–39/20–29	.133	.588	1.008
SAA	Spanish/English	2.741	8.662	11.000
SAA	male/female	.612	1.442	2.447

Table 3: Absolute baseline WER gaps (percentage points) for selected binary pairs.

4.2 Baseline gaps vary by model and corpus

Table 3 shows the absolute WER difference for the binary pairs later considered for steering. MCV sex/gender gaps are modest relative to the age and native/accent pairs, whereas SAA native-language gaps are large for all models. For example, HuBERT produces 11.957% WER for Spanish-source clips and 3.295% for English clips, an 8.662-point gap. The corresponding Wav2Vec2 gap is 11.000 points.

The same pair is not equally difficult for every model. Whisper’s SAA age gap is only 0.133 points and reverses the intended 30–39 to 20–29 direction at the displayed precision, while HuBERT and Wav2Vec2 show larger positive gaps. This model dependence cautions against treating metadata alone as the cause of an error difference.

4.3 Selected WER effects are small and mixed

Figure 1 reports all 22 targeted follow-up reruns, rather than only the supported cases. Eighteen have negative source-class point estimates, but only nine have paired-bootstrap intervals entirely below zero.

Nine other negative estimates are inconclusive, and four estimates are non-negative. Because the candidates were chosen after exploratory sweeps, these proportions are not unbiased estimates of how often steering works.

Every absolute reduction is below one point. The largest is R6, HuBERT SAA Spanish→English centroid steering: source WER changes from 11.957% to 11.272%, a saved delta of -0.684 percentage points with $q_{\text{boot}} = .0010$ and only 36 source clips. R3, HuBERT MCV teens→forties centroid steering, changes WER by -0.440 points on 263 clips. Supported Whisper sex/gender effects are smaller, about -0.22 to -0.24 points. The forest also exposes settings such as R16 and R17 whose point estimates are adverse.

The magnitude matters relative to the original gap. R6 closes roughly 0.684 of an 8.662-point Spanish–English HuBERT gap. R3 closes 0.440 of a 9.871-point teen–forties gap. These are detectable changes in selected settings, not parity and not evidence that the metadata direction uniquely caused the benefit.

4.4 Probe movement can disagree with WER

Table 4 and Figure 2 connect the internal and task-level measurements. R2, a Whisper sex/gender direction transferred from SAA to MCV, raises the next-layer target rate from 5.52% to 99.87%. The shift remains strong at the final encoder index (5.13% baseline versus 96.53% steered), but WER improves by only 0.237 points.

R3 shows a similarly large HuBERT age shift. At index 13, mean target probability changes from 0.221 to 0.973 and target rate from 17.87% to 98.48%. The probability remains elevated at the final index (0.324 to 0.696), yet the WER reduction is 0.440 points. R6 moves fewer Spanish-source clips to the English probe class (2.78% to 47.22%) but has the largest WER reduction among the selected reruns. Thus a larger internal shift does not imply a larger recognition gain.

R17 is the clearest failure. Wav2Vec2 male→female probe steering at index 4 raises the next-index target rate from 8.09% to 99.87%. The effect then decays: at the final index, target rates are 39.67% and 40.56%. Meanwhile source WER changes from 36.760% to 36.879%, a saved $+0.118$ -point adverse estimate. The intervention succeeds locally under the probe while failing as ASR control.

4.5 Probe quality is not a general layer selector

If accessibility identified causal control points, layers with stronger binary probes should tend to yield larger steering benefits. The result is configuration-dependent. For Whisper MCV sex/gender centroid steering, the Spearman correlation between probe accuracy and best saved steering benefit across strengths is essentially zero ($\rho = .0041$, two-sided permutation $p = .9858$); the probe peaks at index 9 and steering at 23. The probe-vector version is weakly negative ($\rho = -.2908$, $p = .1592$).

HuBERT MCV age is a counterexample where rankings align: centroid steering gives $\rho = .6430$, $p = .0014$, with binary-probe-best index 13 and steering-best index 12. SAA HuBERT age probe steering is also positive ($\rho = .5072$, $p = .0110$). These cases show that probes can narrow exploration in some settings, but no architecture-independent layer-selection rule emerges.

4.6 Utterance-level outcomes are mixed

Aggregate WER hides which words change. Three paired examples illustrate the range without serving as evidence on their own. In R21, the baseline transcribes the place name “Padstow” as “pad style,” while the steered hypothesis matches the reference. In R1, both outputs correctly transcribe the same utterance. In adverse R17, the baseline is correct but steering changes “Berlin work became practically” to “Berlinwork became practicalaly.” Table 5 shortens the excerpts for space.

The mixed examples are consistent with the rerun statistics. A constant speaker-linked direction can repair one local error, leave many clips unchanged, and introduce errors elsewhere. Paired resampling aggregates these heterogeneous outcomes, while trajectory probes reveal whether the internal perturbation persists. Neither replaces a user-centred assessment of error severity.

5 Discussion

5.1 Three validity tests for mitigation

The experiments separate three questions that representation-based fairness interventions should answer. First, is the attribute linearly readable? Second, does an intervention move later states in the intended representational direction? Third, does task behaviour improve for the affected group without unacceptable harm elsewhere? Our evidence gives a strong conditional “yes” to the first, mixed

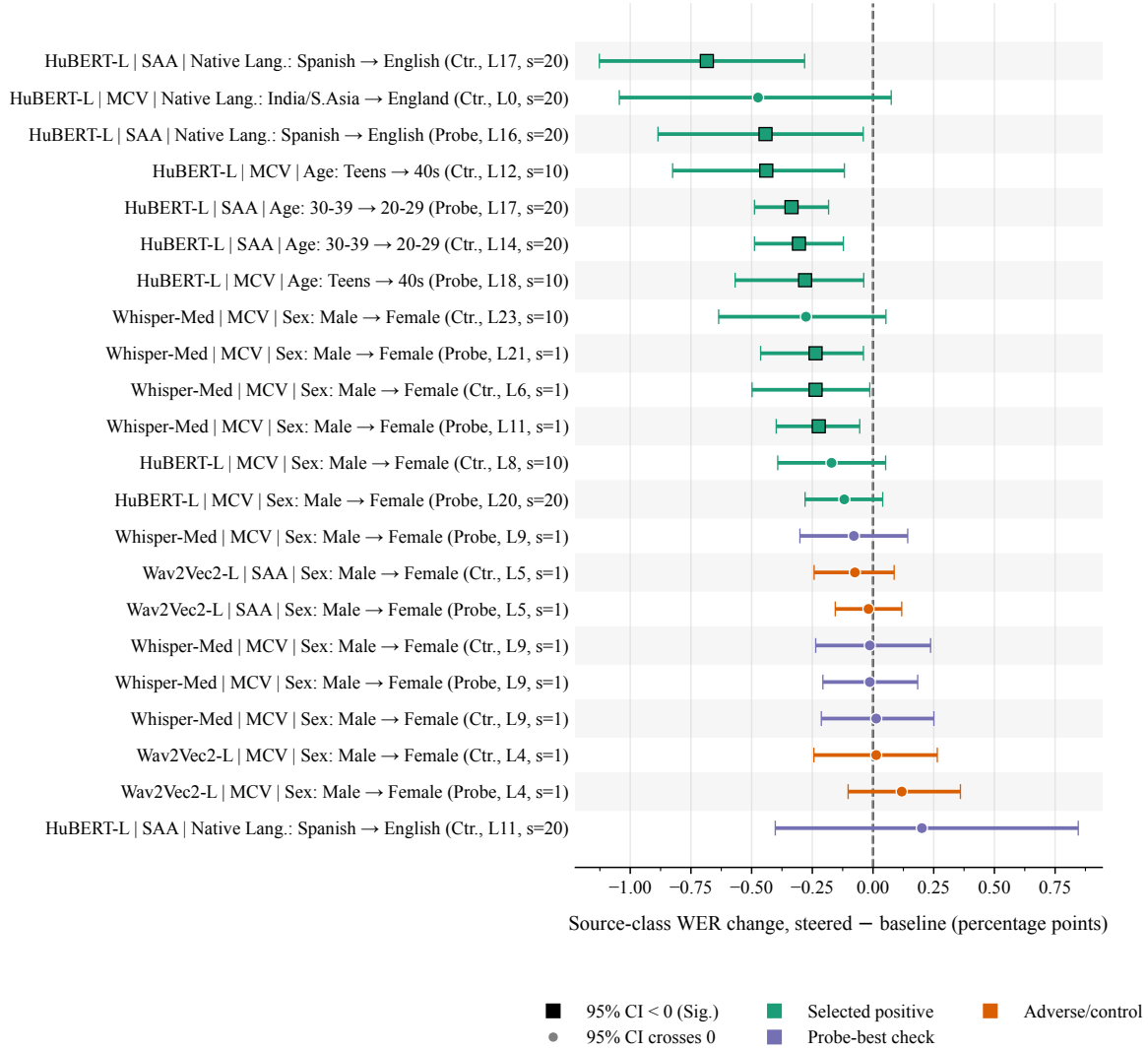


Figure 1: Post-selection source-class WER changes for all 22 targeted reruns (steered minus baseline, percentage points). Negative values favour steering; horizontal bars are paired-bootstrap 95% intervals. Squares mark intervals wholly below zero and circles intervals crossing zero. Row labels give model, evaluation set, source→target direction, vector type, injection layer, strength, and source sample size. Because configurations were chosen after exploratory sweeps, the intervals are descriptive follow-up evidence rather than family-wise confirmatory tests.

answers to the second, and weak, post-selected evidence for the third.

This separation changes how a steering result should be interpreted. A target-like probe flip is a mechanism check, not a fairness metric. It can reflect movement along correlated acoustic dimensions, removal of content useful for decoding, or a transient displacement later blocks undo. The differing Whisper, HuBERT, and Wav2Vec2 trajectories also suggest that encoder geometry cannot be evaluated independently of the downstream decoder. Whisper’s autoregressive decoder may absorb or ignore an encoder perturbation differently from CTC decoding, while later self-supervised blocks can attenuate a locally large shift.

5.2 Implications for fairness claims

The source classes were chosen because their observed WER was higher, but the comparison classes are not causal counterfactuals. The direction male→female or Spanish→English is analytical shorthand, not a claim that one group ought to sound like another. Indeed, treating lower-error groups as representational targets can reproduce assimilationist assumptions if divorced from this diagnostic context.

The true metadata label also determines which clips receive steering. That makes the experiment useful for controlled analysis but unsuitable as a deployment recipe. A deployed recogniser would

Run	Model	Set/task	Vector	Source→target	n	Base	Δ	q_{boot}	Target rate
R2	Whisper	M/S sex	Probe	male→female	779	12.806	−.237	.0080	5.52→99.87
R3	HuBERT	M/M age	Centroid	teens→forties	263	27.618	−.440	.0055	17.87→98.48
R6	HuBERT	S/S native	Centroid	Spanish→English	36	11.957	−.684	.0010	2.78→47.22
R17	W2V2	M/M sex	Probe	male→female	779	36.760	+.118	.8646	8.09→99.87

Table 4: Contrasting trajectory reruns. “Set” gives evaluation/vector data (M=MCV, S=SAA). Base and Δ are source-class WER in percent and percentage points. Target rate is measured at the first diagnostic layer after injection.

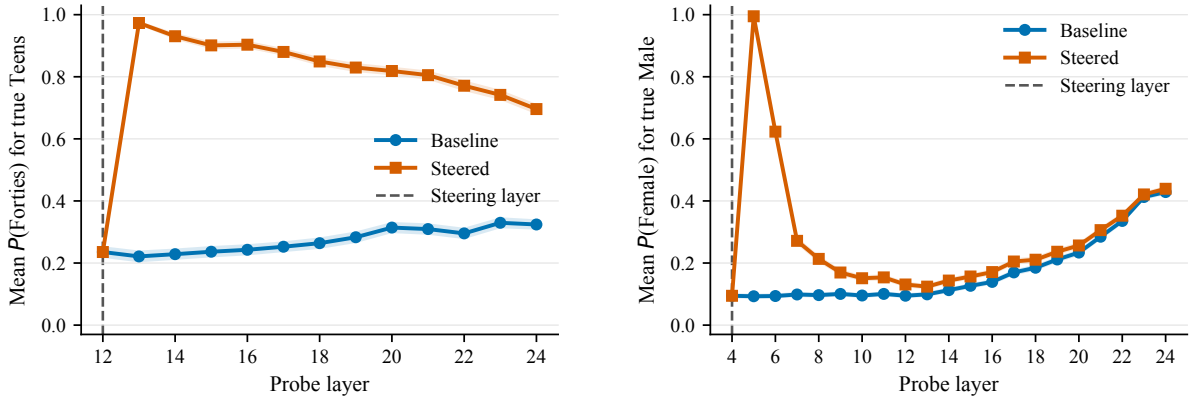


Figure 2: Downstream probe trajectories expose the gap between representation movement and ASR improvement. Left: R3 HuBERT centroid steering from teens to forties at index 12 (strength 10; $n = 263$ source clips) produces a large, persistent target-probability increase, while source WER improves by 0.440 percentage points. Right: R17 Wav2Vec2 probe steering from male to female at index 4 (strength 1; $n = 779$) produces an almost deterministic immediate shift that largely decays, while source WER worsens by 0.118 points. Lines are baseline and steered passes; dashed lines mark intervention indices.

not normally know a speaker’s age, sex/gender, or native-language label before transcription; collecting or inferring it creates privacy and fairness risks of its own. The small improvements here do not justify such a requirement.

5.3 What stronger evidence would require

Simple directions remain valuable because they are transparent and auditable baselines. Future work should nevertheless separate discovery from evaluation: estimate directions on training data, select layers and strengths on a validation set, and reserve a final test set for fairness outcomes. A systematic placebo suite should include norm-matched random directions, shuffled-label centroids, opposite directions, and within-class split directions. Applying each direction to target and unrelated groups would expose collateral harm.

Evaluation should also go beyond WER. Metrics such as match error rate (MER) and word information lost (WIL) (Morris et al., 2004), alongside phoneme error, semantic changes, named-entity errors, calibration, and user-relevant severity, can distinguish an inconsequential substitution from an

accessibility-critical failure. Temporally local interventions may better match accent or age cues than adding one vector to every frame. For Whisper, decoder cross-attention is another plausible intervention point. The core criterion should remain behavioural: an interpretable or steerable feature is useful only if it supports robust improvements under held-out evaluation and appropriate controls.

6 Conclusion

Speaker-linked metadata is linearly accessible across three major ASR encoders, but accessibility does not make the same directions reliable control mechanisms. Selected steering configurations sometimes yield small WER reductions and large internal probe shifts, yet the two effects frequently diverge. The strongest local probe shift in our diagnostic set can accompany worse recognition. Bias-mitigation studies should therefore evaluate readability, propagation, and downstream utility separately and treat success at any one level as insufficient.

Run	Case	Reference cue	Baseline cue	Steered cue
R21	Improved	opposite Padstow	opposite pad style	opposite Padstow
R1	Unchanged	the boy swore that every time	correct	correct
R17	Harmful	Berlin work became practically	correct	Berlinwork became practicalaly

Table 5: Illustrative paired transcription excerpts. Examples reveal error types but do not establish an aggregate effect.

Limitations

The metadata labels are imperfect and reductive. MCV fields are self-reported and incomplete; SAA classes are small; binary male/female labels exclude non-binary identities; and native-language, accent, age, and sex/gender are not interchangeable with acoustic properties. Our labels should be interpreted only as dataset-derived groupings.

The intervention is a diagnostic rather than deployable mitigation. It uses the true source label to decide which clips to steer, whereas a real recogniser would not generally know that label before transcription. Using inferred or declared sensitive attributes at deployment would create additional fairness and privacy risks.

The selected reruns follow exploratory sweeps over models, datasets, pairs, directions, layers, and strengths. Paired bootstrapping quantifies uncertainty for a chosen configuration but does not remove selection bias or control a search-wide error rate. A separate validation set and untouched final test set are needed for confirmatory claims. The final suite also lacks systematic placebo directions and a complete analysis of effects on non-source groups.

The SAA split audit is weaker than MCV’s explicit `client_id` check, and several SAA comparisons have small source groups, including 36 Spanish clips. The group comparisons are not counterfactual; recording quality, sentence content, speaking rate, and other difficulty factors may contribute to WER differences.

Mean pooling removes temporal structure, and the saved pipeline averages padded hidden sequences without masking. Adding one constant vector at every time step may therefore exaggerate a global speaker direction and poorly match phoneme-local accent or age cues. The evaluation covers two English datasets and three pretrained systems, not multilingual speech LLMs or production conditions. Finally, WER treats all word errors alike and does not measure semantic severity or user impact.

Ethical Considerations

This study analyses labels linked to socially sensitive attributes. High probe scores must not be interpreted as accurate inference of identity, nor as justification for demographic profiling. The source-to-target notation follows an observed WER contrast and does not imply that a higher-error group should acoustically conform to a lower-error group. Any future deployment-oriented mitigation should avoid requiring sensitive labels where possible, obtain appropriate consent, assess intersecting groups, and test whether gains for one group impose harms on others. The small, post-selected effects reported here are explicitly presented as a warning against premature fairness claims.

References

- Guillaume Alain and Yoshua Bengio. 2017. [Understanding intermediate layers using linear classifier probes](#).
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. [Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.

- Yonatan Belinkov and James Glass. 2017. [Analyzing hidden representations in end-to-end automatic speech recognition systems](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Junxi Chen, Junhao Dong, and Xiaohua Xie. 2026. [Adaptive probe-based steering for robust LLM jail-breaking](#). In *Forty-third International Conference on Machine Learning*.
- Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2021. [Exploring wav2vec 2.0 on speaker verification and language identification](#). pages 1509–1513.
- Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. [Quantifying bias in automatic speech recognition](#). *Preprint*, arXiv:2103.15122.
- Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. 2025. [To steer or not to steer? mechanistic error reduction with abstention for language models](#). In *Forty-second International Conference on Machine Learning*.
- Felix Herron, Solange Rossato, Alexandre Allauzen, Benoit Favre, and François Portet. 2026a. Speaker group encoding in self-supervised speech recognition models. In *Text, Speech, and Dialogue*, pages 121–132, Cham. Springer Nature Switzerland.
- Felix Herron, Solange Rossato, Alexandre Allauzen, and François Portet. 2026b. [Identifying and typifying demographic unfairness in phoneme-level embeddings of self-supervised speech recognition models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33603–33620, San Diego, California, United States. Association for Computational Linguistics.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [Hubert: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.
- Maliha Jahan, Priyam Mazumdar, Thomas Thebaud, Mark Hasegawa-Johnson, Jesús Villalba, Najim Dehak, and Laureano Moro-Velazquez. 2025. [Unveiling performance bias in asr systems: A study on gender, age, accent, and more](#). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. [Racial disparities in automated speech recognition](#). *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.
- Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. [Probing classifiers are unreliable for concept removal and detection](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17994–18008. Curran Associates, Inc.
- Yi-Cheng Lin, Tzu-Quan Lin, Hsi-Che Lin, Andy T. Liu, and Hung yi Lee. 2024. [On the social bias of speech self-supervised models](#). In *Interspeech 2024*, pages 4638–4642.
- Oli Liu, Hao Tang, and Sharon Goldwater. 2023. [Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces](#). pages 2968–2972.
- Samuel Marks and Max Tegmark. 2024. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). In *First Conference on Language Modeling*.
- Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. [From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition](#). In *Interspeech 2004*, pages 2765–2768.
- Paul-Gauthier Noé, Aran Emîni, Matrouf Driss, Titouan Parcollet, Andreas Nautsch, and Jean-François Bonastre. 2021. [Adversarial disentanglement of speaker representation for attribute-driven privacy preservation](#).
- Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. [Layer-wise analysis of a self-supervised speech representation model](#). In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Quentin Raymondoud, Mickael Rouvier, and Richard Dufour. 2024. [Probing the information encoded in neural-based acoustic models of automatic speech recognition systems](#). *Preprint*, arXiv:2402.19443.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Leda Sari, Niko Moritz, Takaaki Hori, and Jonathan Le Roux. 2020. [Unsupervised speaker adaptation using attention-based speaker memory for end-to-end asr](#). In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7384–7388.
- Steven Weinberger. 2015. [Speech accent archive](#). George Mason University. Retrieved from <http://accent.gmu.edu>.

AIP-Speech: Auditing Speech LLM Robustness to Background Interference

Aiden Dsouza¹, Shreyansh Mishra¹, Yashraj Suneja¹, Rohan Swain¹,
Samyak Borkar¹, Arjun Jaishankar¹, Tanmay Urade¹

¹IvLabs, Visvesvaraya National Institute of Technology (VNIT), India

Correspondence: aiden.dsouza@ivlabs.in

Abstract

Speech Large Language Models (Speech LLMs) are increasingly deployed in real-world acoustic environments, yet it remains unclear which measurable properties of background sounds best explain performance degradation. We present AIP-Speech, an inference-only benchmark that evaluates Speech LLM robustness under controlled mixtures of real-world background recordings. Rather than comparing robustness across background categories alone, we characterize each recording using interpretable acoustic descriptors and examine their associations with degradation. Across five open Speech LLMs, we find that simple acoustic descriptors—including Zero Crossing Rate, Spectral Centroid, and Spectral Rolloff—consistently exhibit strong associations with degradation. We further quantify robustness disparities across accents and evaluate whether inference-time prompting improves resilience under adverse acoustic conditions. AIP-Speech provides a reproducible benchmark and descriptor-annotated background corpus for analyzing Speech LLM robustness.

1 Introduction

Speech large language models integrate speech encoders with pretrained language models to perform tasks such as transcription, keyword spotting, and spoken question answering directly from raw audio (Chu et al., 2024; Xu et al., 2025; Abouelenin et al., 2025). Modern Speech LLMs adopt an end-to-end architecture that unifies speech perception and language reasoning within a single model (Peng et al., 2025). Speech, however, is rarely encountered in isolation.

How Speech LLMs behave under background interference remains poorly understood. Existing benchmarks evaluate Speech LLMs under environmental variation (Chen et al., 2026), speaker-side phenomena such as speech disfluency (Liu et al.,

2025), natural in-the-wild recordings (Zhang et al., 2025), or systematic perturbations (Shah et al., 2025). Broader benchmarks additionally assess trustworthiness across dimensions such as robustness, safety, fairness, and hallucination (Li et al., 2025). However, analyses of background interference typically report aggregate performance under a limited set of acoustic conditions, providing little insight into which measurable acoustic properties are associated with degradation. Moreover, it remains unclear whether background interference affects demographic groups uniformly or amplifies existing disparities among them.

To address these gaps, we introduce **Acoustic Interference Profiling (AIP)-Speech**, an inference-only benchmark for systematically auditing Speech LLM robustness under controlled background interference¹. By holding the spoken content fixed while systematically varying the background audio, we isolate the effect of different classes of acoustic interference on Speech LLM behavior. Specifically, our work makes three contributions: (i) we introduce a reproducible benchmark and evaluation framework for auditing Speech LLMs under controlled real-world background interference;

(ii) we present descriptor-level analysis of Speech LLM robustness, relating interpretable acoustic descriptors to degradation across automatic speech recognition and keyword spotting;

(iii) we evaluate robustness disparities across demographic groups and investigate inference-time prompting under controlled acoustic interference.

2 AIP-Speech Benchmark

Figure 1 summarizes the AIP-Speech evaluation pipeline. A clean speech utterance is mixed with a controlled background recording at multiple signal-to-noise ratios (SNRs) to produce paired noisy examples. Speech LLMs are evaluated under identi-

¹<https://github.com/IvLabs/AIP-Speech>

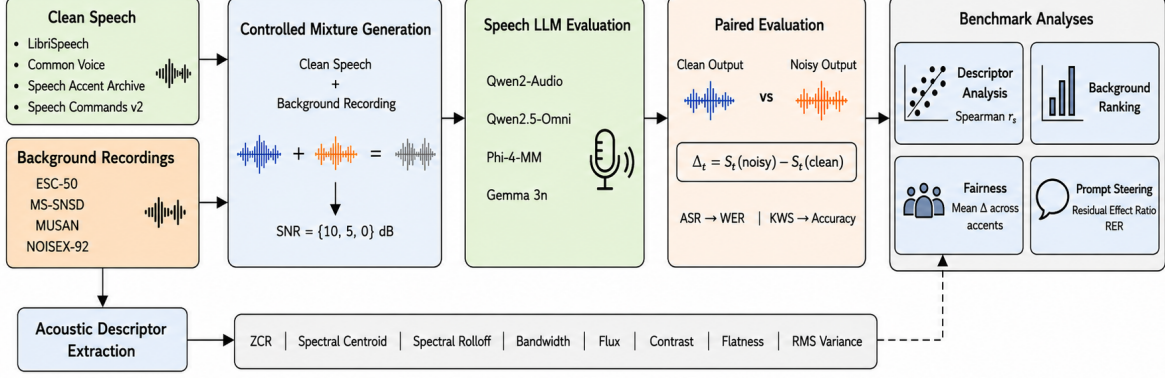


Figure 1: Overview of the AIP-Speech benchmark. Clean speech is combined with background recordings to evaluate Speech LLM robustness under controlled acoustic interference. Acoustic descriptors are extracted independently from background recordings and incorporated only during the analysis stage to explain robustness patterns.

cal inference settings, and paired clean–noisy predictions are used to compute robustness, fairness, prompting, and descriptor-based analyses. Pairing every noisy example with its clean counterpart isolates the effect of background interference from intrinsic model accuracy.

2.1 Background Audio and Descriptor Extraction

AIP-Speech uses a fixed set of 20 publicly available real-world background recordings, ensuring reproducibility while capturing diverse acoustic conditions. Recordings are drawn from ESC-50 (Piczak, 2015), MS-SNSD (Reddy et al., 2019), NOISEX-92 (Varga and Steeneken, 1993), and MUSAN (Snyder et al., 2015). Each clean speech utterance is combined with a background recording at multiple target signal-to-noise ratios (SNRs) to produce deterministic noisy counterparts. All recordings are resampled and loudness-normalized prior to mixture generation. The complete list of background recording classes is provided in Appendix A.

Each background recording is further characterized using a set of standard acoustic descriptors computed directly from the waveform. Specifically, we compute spectral centroid, spectral rolloff, spectral bandwidth, spectral flatness, zero crossing rate (ZCR), spectral flux, spectral contrast, and RMS energy variance (McFee et al., 2015). These descriptors capture complementary spectral and temporal properties of the background signal and are computed once for each recording prior to evaluation. Descriptor values are subsequently related to

task degradation using Spearman rank correlation.

2.2 Robustness Measures

For each evaluation task t , let $S_t(\mathbf{x})$ denote the task-specific scoring function evaluated on speech input $\mathbf{x}$. Robustness is quantified as the performance degradation Δ_t between paired clean and noisy examples $(\mathbf{x}_{\text{clean}}, \mathbf{x}_{\text{noisy}})$, formulated such that positive values consistently represent a performance drop relative to the clean baseline:

$$\Delta_t = \begin{cases} S_{\text{ASR}}(\mathbf{x}_{\text{noisy}}) - S_{\text{ASR}}(\mathbf{x}_{\text{clean}}), & \text{for } t = \text{ASR} \\ S_{\text{KWS}}(\mathbf{x}_{\text{clean}}) - S_{\text{KWS}}(\mathbf{x}_{\text{noisy}}), & \text{for } t = \text{KWS} \end{cases} \quad (1)$$

where $S_{\text{ASR}}(\cdot)$ denotes the Word Error Rate (WER) and $S_{\text{KWS}}(\cdot)$ denotes classification accuracy. Consequently, Δ_{ASR} captures the absolute increase in WER, whereas Δ_{KWS} measures the drop in accuracy. KWS uses a closed-set protocol over the ten Google Speech Commands v2 keywords, where target utterances require an exact keyword match and non-target utterances are scored correct only if no target keyword is predicted.

To quantify demographic robustness, we compute the mean performance degradation for each demographic group g ,

$$\overline{\Delta}_{t,g} = \frac{1}{N_g} \sum_{i \in g} \Delta_t^{(i)}, \quad (2)$$

where N_g denotes the number of evaluation examples belonging to group g . Larger values indicate that the corresponding demographic experiences greater performance degradation under background interference.

Prompt effectiveness is summarized using the Residual Effect Ratio,

$$RER = \frac{\overline{\Delta_t}^{prompt}}{\overline{\Delta_t}^{baseline}}, \quad (3)$$

where values below one indicate reduced degradation under prompting.

3 Experimental Setup

We evaluate AIP-Speech on two speech processing tasks: automatic speech recognition (ASR) and keyword spotting (KWS). The ASR evaluation comprises 180 speech utterances, including 100 from LibriSpeech (Panayotov et al., 2015) and 80 accent-labelled utterances from Mozilla Common Voice 17 (Ardila et al., 2020). Robustness across demographic groups is evaluated using these accent-labelled utterances. The KWS evaluation comprises 113 audio clips from Google Speech Commands v2 (Warden, 2018).

We evaluate five publicly available Speech LLMs: Qwen2.5-Omni-3B, Qwen2.5-Omni-7B (Xu et al., 2025), Qwen2-Audio-Instruct (Chu et al., 2024), Phi-4-Multimodal (Abouelenin et al., 2025), and Gemma 3n (Gemma Team, 2025). All models are evaluated using deterministic greedy decoding. All models are evaluated using the same background recordings at signal-to-noise ratios (SNRs) of 10, 5, and 0 dB. Complete model-wise performance across all evaluated SNR conditions is provided in Appendix G. Robustness metrics are computed from paired clean and noisy predictions. For descriptor-level analysis, degradation is averaged across the evaluated SNR levels before computing Spearman rank correlations with the corresponding acoustic descriptors. The steering prompt is applied consistently across all evaluated models, and its complete text is provided in Appendix F.

4 Results

We evaluate Speech LLM robustness under controlled acoustic interference using AIP-Speech through descriptor analysis, background ranking, fairness evaluation, and inference-time prompt steering.

4.1 Descriptor-Level Analysis

Figure 2 demonstrates that robustness degradation is not randomly distributed across background recordings. Instead, recordings with larger descriptor values consistently tend to produce greater

degradation for both ASR and KWS, despite the inherent variability expected from real-world acoustic environments. The similar trends observed across two fundamentally different speech tasks suggest that the relationship reflects a property of the background signal rather than task-specific behavior.

Figure 3 shows that Zero Crossing Rate (ZCR) exhibits the strongest marginal association with degradation ($r_s = 0.645$ for ASR and 0.723 for KWS), closely followed by Spectral Centroid and Spectral Rolloff, while descriptors such as Spectral Flatness, Spectral Contrast, and RMS Energy Variance show weaker associations. Moreover, the ordering of descriptor associations is consistent across ASR and KWS. To assess consistency across architectures, we compute these correlations independently for each model in Appendix C. The strongest descriptors remain positively associated with degradation across all five models, although the correlation magnitudes vary by architecture. Collectively, these results indicate that simple, low-level acoustic descriptors provide informative signals for characterizing degradation.

4.2 Background Ranking

Figure 4 provides a recording-level perspective on Speech LLM robustness. The observed variability across recordings is consistent with the descriptor analysis, where Zero Crossing Rate, Spectral Centroid, and Spectral Rolloff exhibit the strongest associations with degradation across both tasks. The complete descriptor values for every background recording are provided in Appendix D.

4.3 Fairness

Figure 5 compares the robustness of Speech LLMs across accent subgroups by measuring the mean increase in Word Error Rate (ΔWER) between clean and noisy speech. Computing degradation relative to each subgroup’s clean baseline isolates the effect of background interference from differences in baseline recognition performance. The figure reveals substantial variation in robustness across both models and accent subgroups. While some models exhibit relatively uniform degradation, others show pronounced sensitivity to specific accents, indicating that background interference does not affect all demographic groups equally. Complete subgroup-wise statistics are reported in Appendix E.

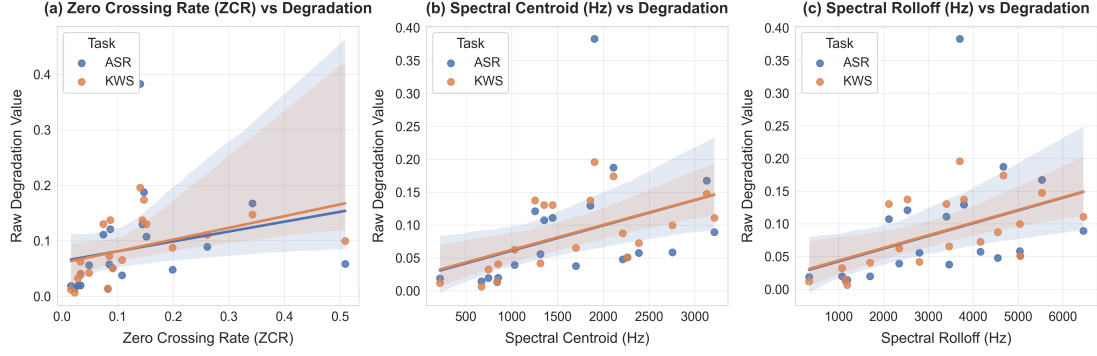


Figure 2: Relationship between three representative acoustic descriptors and task degradation. Each point corresponds to one background recording. Larger descriptor values generally correspond to greater degradation for both ASR and KWS.

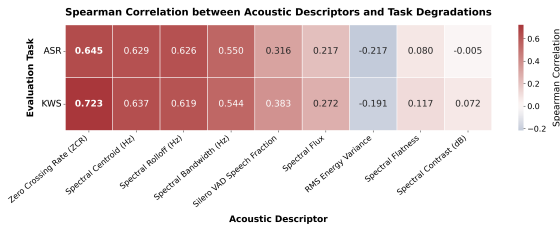


Figure 3: Spearman rank correlation between acoustic descriptors and task degradation.

4.4 Prompt-based Steering

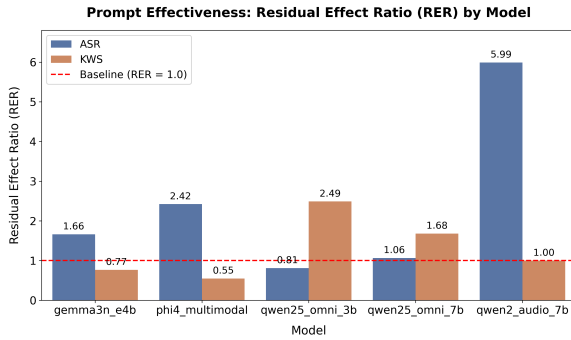


Figure 6: Residual Effect Ratio (RER) after applying inference-time prompt steering. Values below one indicate reduced degradation relative to the baseline instruction.

We evaluate whether inference-time prompt steering improves robustness without modifying model parameters. Figure 6 reports the Residual Effect Ratio (RER), which measures the degradation remaining after steering relative to the baseline instruction. Lower RER values indicate greater robustness gains, with the magnitude of improvement varying across the evaluated Speech LLMs. Complete Residual Effect Ratio (RER) values for

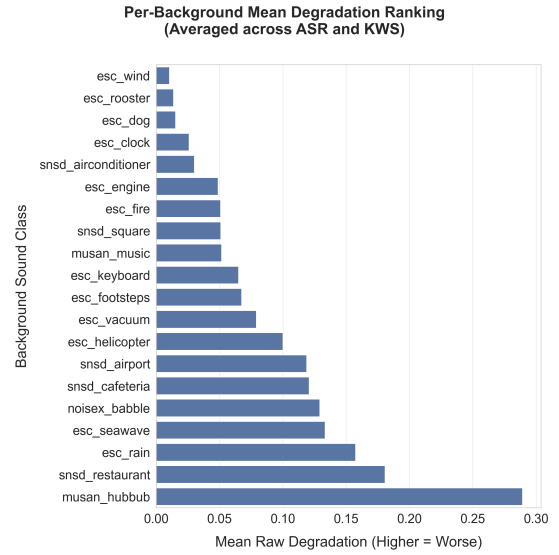


Figure 4: Mean degradation ranking of individual background recordings averaged across ASR and KWS.

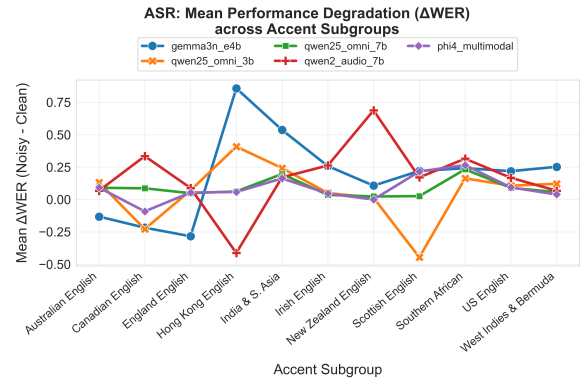


Figure 5: Mean ASR performance degradation (Δ WER) across accent subgroups under background interference. Larger values indicate greater degradation relative to the corresponding clean condition.

all evaluated models are reported in Table 5 (Appendix F).

5 Conclusion

We introduced AIP-Speech, an inference-only benchmark for evaluating Speech LLM robustness under controlled real-world background interference. By pairing clean and noisy utterances, annotating backgrounds with interpretable acoustic descriptors, and evaluating ASR, keyword spotting, demographic subgroups, and prompt-based steering, AIP-Speech provides a systematic framework for analyzing acoustic interference. Our results show that simple acoustic descriptors are strongly associated with degradation, robustness varies substantially across accent groups, and inference-time prompting yields limited improvements.

Limitations

This work is limited to inference-time evaluation of five open Speech LLMs on automatic speech recognition and keyword spotting, and our findings may not generalize to conversational or reasoning-intensive speech tasks. Although the benchmark covers diverse real-world background recordings, it does not exhaust the full range of acoustic environments encountered in practice. The fairness analysis is limited to English accent subgroups and relies on the demographic annotations provided by the underlying datasets. Finally, descriptor-level analysis identifies statistical associations with robustness degradation but does not establish causal relationships between acoustic properties and model behavior.

References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.
- Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2026. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Gemma Team. 2025. **Gemma 3n**. Google. On-device multimodal model with a Universal Speech Model audio encoder.
- Kai Li, Can Shen, Yile Liu, Jirui Han, Kelong Zheng, Xuechao Zou, Lionel Z Wang, Shun Zhang, Xingjian Du, Hanjun Luo, et al. 2025. Audiotrust: Benchmarking the multifaceted trustworthiness of audio large language models. *arXiv preprint arXiv:2505.16211*.
- Hongcheng Liu, Yixuan Hou, Heyang Liu, Yuhao Wang, Yanfeng Wang, and Yu Wang. 2025. Vocalbench-df: A benchmark for evaluating speech llm robustness to disfluency. *arXiv preprint arXiv:2510.15406*.
- Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, et al. 2015. librosa: Audio and music signal analysis in python. *SciPy*, 2015(18-24):7.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.
- Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, et al. 2025. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*.
- Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.
- Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. 2019. A scalable noisy speech dataset and online subjective test framework. *arXiv preprint arXiv:1909.08050*.
- Muhammad Shah, David Solans Noguero, Mikko Heikkilä, Bhiksha Raj, and Nicolas Kourtellis. 2025. Speech robust bench: A robustness benchmark for speech recognition. In *International Conference on Learning Representations*, volume 2025, pages 38625–38651.
- David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

- Andrew Varga and Herman JM Steeneken. 1993. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251.
- Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Ke qin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. [Qwen2.5-omni technical report](#). *ArXiv*, abs/2503.20215.
- Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. 2025. Wildspeech-bench: Benchmarking end-to-end speechllms in the wild. *arXiv preprint arXiv:2506.21875*.

Background	Source
rain	ESC-50
sea waves	ESC-50
engine	ESC-50
vacuum	ESC-50
footsteps	ESC-50
fire	ESC-50
helicopter	ESC-50
clock tick	ESC-50
keyboard	ESC-50
dog	ESC-50
rooster	ESC-50
wind	ESC-50
cafeteria	MS-SNSD
restaurant	MS-SNSD
public square	MS-SNSD
airport	MS-SNSD
air conditioner	MS-SNSD
babble	NOISEX-92
music	MUSAN
hubbub	MUSAN

Table 1: Background recordings used throughout the benchmark.

A Background Noise Set

Table 1 lists the 20 background recordings used in AIP-Speech. Recordings are drawn from four publicly available datasets to provide a diverse range of real-world acoustic environments.

B Acoustic Descriptor Definitions

Each background recording is characterized using a set of standard acoustic descriptors computed directly from the waveform. Unless otherwise specified, descriptor values are averaged over the recording after frame-level computation.

- **Zero Crossing Rate (ZCR):** Measures the rate at which the waveform changes sign and serves as a simple indicator of high-frequency activity.
- **Spectral Centroid:** Represents the center of mass of the magnitude spectrum and is commonly interpreted as the perceptual brightness of a sound.
- **Spectral Bandwidth:** Measures the spread of spectral energy around the spectral centroid, indicating how broadly energy is distributed across frequencies.
- **Spectral Rolloff:** The frequency below which 85% of the spectral energy is contained, providing a measure of the effective upper frequency content.

- **Spectral Flux:** Quantifies frame-to-frame changes in the magnitude spectrum and reflects temporal spectral variation.
- **Spectral Flatness:** Measures how noise-like a signal is by comparing the geometric and arithmetic means of the spectrum. Higher values indicate flatter, more noise-like spectra.
- **Spectral Contrast:** Measures the difference between spectral peaks and valleys across frequency bands, capturing harmonic structure and spectral richness.
- **RMS Energy Variance:** Measures the temporal variability of signal energy throughout the recording.
- **Speech Fraction:** Computed using the Silero Voice Activity Detector (VAD) as the proportion of frames classified as speech.

C Per-Model Descriptor Correlations

To assess whether descriptor–degradation relationships are consistent across architectures, Table 2 reports Spearman rank correlations computed independently for each Speech LLM. Degradation is averaged across the evaluated SNR levels before computing the correlations, consistent with the descriptor-level analysis in the main paper.

D Per-Background Acoustic Descriptor Values

Table 3 reports the acoustic descriptor values computed for every background recording used in AIP-Speech. Providing these values enables direct inspection of the descriptor profile associated with each recording and facilitates reproducibility of the descriptor-level analyses presented in the main paper.

E Complete Fairness Results

Table 4 reports the complete subgroup-wise robustness results underlying Figure 5. Values correspond to the mean increase in Word Error Rate (Δ WER) relative to each subgroup’s clean baseline. Reporting the complete statistics enables reproducibility beyond the visualization shown in the main paper.

F Prompt-based Steering

For the prompt-based robustness experiment, each model is evaluated under two inference settings:

Descriptor	Task	Pooled (All 5)	Qwen2-Audio	Qwen2.5-Omni-7B	Qwen2.5-Omni-3B	Phi-4-MM	Gemma 3n
Zero Crossing Rate	ASR	+0.645	+0.711	+0.704	+0.693	+0.672	+0.469
	KWS	+0.723	-0.126	+0.585	+0.647	+0.710	+0.678
Spectral Centroid	ASR	+0.629	+0.693	+0.597	+0.558	+0.642	+0.451
	KWS	+0.637	-0.322	+0.501	+0.488	+0.674	+0.567
Spectral Rolloff	ASR	+0.626	+0.687	+0.586	+0.544	+0.669	+0.448
	KWS	+0.619	-0.228	+0.444	+0.439	+0.656	+0.581
Spectral Bandwidth	ASR	+0.550	+0.621	+0.522	+0.462	+0.636	+0.337
	KWS	+0.544	-0.118	+0.406	+0.402	+0.576	+0.519
Spectral Flux	ASR	+0.217	+0.248	+0.263	+0.250	+0.192	+0.107
	KWS	+0.272	-0.173	+0.384	+0.278	+0.277	+0.226
Spectral Flatness	ASR	+0.080	+0.150	+0.029	-0.030	+0.099	+0.177
	KWS	+0.117	-0.243	+0.025	-0.029	+0.165	+0.038
Spectral Contrast	ASR	-0.005	+0.053	+0.057	+0.087	-0.065	-0.068
	KWS	+0.072	-0.168	-0.013	+0.075	+0.062	-0.003
RMS Energy Variance	ASR	-0.217	-0.233	-0.247	-0.179	-0.251	-0.096
	KWS	-0.191	+0.311	-0.156	-0.205	-0.217	-0.173
Silero VAD Speech Fraction	ASR	+0.316	+0.314	+0.406	+0.397	+0.275	+0.254
	KWS	+0.383	+0.017	+0.457	+0.431	+0.362	+0.369

Table 2: Spearman rank correlation coefficients (r_s) between acoustic descriptors and task degradation. The **Pooled (All 5)** column reports the aggregate correlations used in Figure 3, while the remaining columns report correlations computed independently for each Speech LLM. Degradation is averaged across the evaluated SNR levels before computing the correlations.

(i) the default task instruction (baseline), and (ii) a steering prompt that instructs the model to ignore background interference and attend only to the intended foreground speaker. The same steering prompt is applied consistently across all evaluated Speech LLMs.

F.1 Automatic Speech Recognition (ASR)

Baseline Prompt

Transcribe the speech in this audio clip exactly as spoken. Return only the transcription text.

Steering Prompt

*You are an automatic speech recognition (ASR) system operating in noisy real-world environments. Transcribe **only** the speech spoken by the intended foreground speaker. Ignore background voices, overlapping conversations, music, television, announcements, and all environmental sounds.*

If any foreground speech is unintelligible because of background interference, omit it rather than guessing or hallucinating.

Return only the final transcript without explanations, confidence scores, or additional text.

F.2 Keyword Spotting (KWS)

Baseline Prompt

*Listen to this audio clip. Respond with **only** the spoken word (one word). If no clear word is spoken, respond with SILENCE.*

Steering Prompt

Identify the keyword spoken by the intended foreground speaker. Ignore all background voices, music, television, announcements, and environmental sounds.

If no foreground keyword can be identified, respond with SILENCE.

Return only the keyword or SILENCE without any additional text.

F.3 Prompt Steering Results

Table 5 reports the Residual Effect Ratio (RER) for each evaluated Speech LLM. Values below one indicate that prompt steering reduces the degradation caused by background interference.

G Performance Across Signal-to-Noise Ratios

To evaluate robustness across different interference levels, we report model performance under clean conditions together with additive background interference at 10, 5, and 0 dB SNR.

Figure 7 summarizes the average performance of all evaluated Speech LLMs across the two benchmark tasks. As expected, lower signal-to-noise ratios consistently degrade model performance. The

Background	ZCR	Centroid	Bandwidth	Rolloff	Speech	Flux	Flatness	Contrast	RMS Var.
esc_wind	0.0226	669.76	1330.72	1190.62	0.0000	0.9022	0.0011	17.2040	0.000658
esc_rooster	0.0822	841.66	606.18	1153.43	0.0000	0.6738	0.4599	27.7246	0.001060
esc_dog	0.0161	218.08	142.86	342.65	0.0000	0.3479	0.9091	16.0911	0.000359
esc_clock	0.0288	746.37	1351.25	1076.89	0.0021	1.7850	0.0013	19.2353	0.001402
snsd_airconditioner	0.0333	851.80	1375.42	1699.16	0.0000	0.9540	0.0032	15.7306	0.000312
esc_engine	0.0488	1314.33	1544.96	2801.37	0.0000	0.8921	0.0088	18.0697	0.002576
esc_fire	0.0912	2264.88	2236.70	5048.61	0.0000	2.1282	0.0931	16.1431	0.001323
snsd_square	0.0332	1033.03	1645.28	2351.40	0.1014	1.1128	0.0049	15.2164	0.004249
musan_music	0.1081	1701.98	1668.04	3462.07	0.3362	1.3270	0.0215	20.7004	0.002288
esc_keyboard	0.0848	2388.24	1807.72	4160.08	0.0000	3.3657	0.1094	17.3824	0.000738
esc_footsteps	0.1987	2212.21	1959.91	4548.02	0.0000	1.6759	0.1182	16.8711	0.001814
esc_vacuum	0.5089	2755.58	2109.97	5039.44	0.0000	0.9197	0.0533	19.2603	0.000078
esc_helicopter	0.2609	3211.55	2666.81	6452.88	0.0000	0.9847	0.0935	18.6647	0.000018
snsd_airport	0.1516	1354.49	766.39	2120.14	0.0000	1.1973	0.0002	20.3577	0.001119
snsd_cafeteria	0.0744	1446.31	1905.25	3401.42	0.8367	1.1776	0.0211	15.6066	0.000375
noisex_babble	0.0870	1255.53	1428.23	2535.08	0.4696	1.1716	0.0114	18.5808	0.000327
esc_seawave	0.1446	1859.74	1834.55	3791.06	0.0000	0.8943	0.0655	17.5634	0.000899
esc_rain	0.3420	3129.37	1921.83	5528.74	0.0000	0.9045	0.2122	18.5032	0.000029
snsd_restaurant	0.1472	2110.20	2106.19	4668.49	0.0171	1.0586	0.0909	14.9841	0.000437
musan_hubbub	0.1406	1904.30	1705.12	3701.78	0.6201	1.9962	0.0648	19.4705	0.002083

Table 3: Per-recording acoustic descriptor values for every background recording used in AIP-Speech. These descriptor profiles underlie the descriptor-level correlation analysis presented in the main paper and provide a recording-level acoustic characterization for every background recording.

Accent Group	Gemma 3n	Phi-4-MM	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B	Qwen2-Audio
Australian English	-0.1316	+0.0924	+0.1309	+0.0906	+0.0681
Canadian English	-0.2173	-0.0915	-0.2280	+0.0866	+0.3350
England English	-0.2835	+0.0542	+0.0705	+0.0490	+0.0888
Hong Kong English	+0.8583	+0.0583	+0.4083	+0.0625	-0.4125
Indian English	+0.5370	+0.1628	+0.2425	+0.1967	+0.1713
Irish English	+0.2583	+0.0461	+0.0500	+0.0356	+0.2628
New Zealand English	+0.1071	0.0000	+0.0143	+0.0238	+0.6881
Scottish English	+0.2211	+0.2164	-0.4461	+0.0261	+0.1708
US English	+0.0854	+0.0480	+0.0579	+0.0522	+0.0490
Other / Misc.	+0.2470	+0.1526	+0.1425	+0.1453	+0.1915

Table 4: Complete subgroup-wise robustness statistics (mean Δ WER) used in the fairness analysis. Larger positive values indicate greater performance degradation caused by background interference relative to clean baseline.

Model	ASR	KWS
Gemma 3n	1.664	0.767
Phi-4-MM	2.423	0.549
Qwen2.5-Omni-3B	0.812	2.489
Qwen2.5-Omni-7B	1.064	1.681
Qwen2-Audio	5.989	1.000

Table 5: Residual Effect Ratio (RER) after prompt-based steering. Values below one indicate reduced degradation relative to the baseline prompt.

figure provides additional context for the descriptor-level analysis presented in the main paper by illustrating the overall effect of increasing acoustic interference.

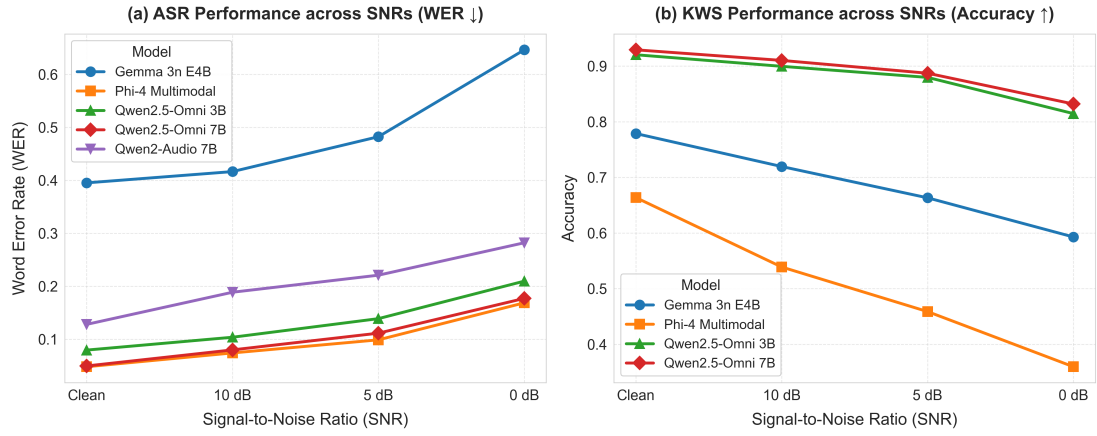


Figure 7: Average model performance under clean conditions and progressively more challenging acoustic environments (10, 5, and 0 dB SNR). (a) Automatic Speech Recognition performance measured using Word Error Rate (WER; lower is better). (b) Keyword Spotting performance measured using classification accuracy (higher is better). Performance consistently deteriorates as background interference increases, motivating the descriptor-level robustness analysis presented in the main paper.

Can Large Audio-Language Models Ignore Multilingual Distractors? An Evaluation of Their Selective Auditory Attention Capabilities

Heejoon Koo

University of Illinois Urbana-Champaign
heejoon4@illinois.edu

Abstract

Robust selective auditory attention under multilingual interference is critical for reliable LALM deployment. We introduce MUSI, a cocktail party-inspired controlled diagnostic evaluation of source-grounded spoken-language understanding and reasoning. Each item pairs an English target dialogue with a plausible distractor in English, Spanish, Korean, or Chinese, and evaluates models under (1) single-stream, (2) separation-based, and (3) end-to-end cocktail party settings across controlled SNRs. Across four open-weight and two closed-source LALMs, we find model-dependent language-conditioned variation and heightened vulnerability at adverse SNRs. Errors are dominated by distractor-grounded source confusion, while separation reduces acoustic overlap but often leaves source attribution unresolved. These findings highlight selective auditory attention as a primary capability for reliable LALMs deployment.

1 Introduction

Overlapping speech has long been studied in ASR and speech separation through multi-speaker recognition, diarization, and source separation (Hershey et al., 2016; Cosentino et al., 2020; Watanabe et al., 2020). Auditory neuroscience frames the same challenge as the cocktail party problem: listeners must organize acoustic mixtures and selectively attend to task-relevant sources while suppressing competing speech (Cherry, 1953; McDermott, 2009; Bronkhorst, 2015). For Large Audio-Language Models (LALMs), however, separating or transcribing speech does not guarantee source-grounded reasoning: a model may rely on the wrong stream, conflate sources, or generate fluent but source-incoherent answers.

This failure mode is critical in high-stakes domains such as aviation and healthcare, where audio-based systems are increasingly used for tasks including respiratory sound analysis (Koo et al.,

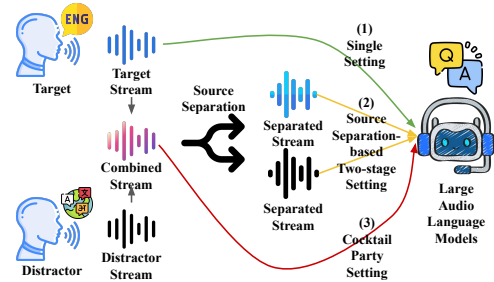


Figure 1: Our MUSI evaluation framework.

2026a,b), and concurrent same- or cross-lingual speech may contain semantically salient but task-irrelevant information. It also raises a multilingual reliability concern: performance may vary with the language of surrounding speech. Existing benchmarks address related but distinct capabilities: multi-speaker ASR and separation benchmarks focus on transcription, diarization, or signal reconstruction (Hershey et al., 2016; Watanabe et al., 2020; Cosentino et al., 2020; Borsdorf et al., 2021; Nguyen et al., 2026); general LALM benchmarks primarily evaluate single-stream audio understanding (Yang et al., 2024; Wang et al., 2025; Yang et al., 2025; Chen et al., 2026); and trustworthiness benchmarks study acoustic perturbations or instruction-level threats (Li et al., 2025; Song et al., 2025; Hou et al., 2025; Cheng et al., 2025; Peng et al., 2025). As summarized in Table 1, none directly evaluates whether LALMs can follow a target dialogue while suppressing plausible multilingual distractors. Such controlled interference reveals whether answers are grounded in the target, a competing stream, or neither, and how performance varies across distractor languages.

We introduce **MUSI** (**M**ultilingual **S**elective **A**ttention), a multilingual multiple-choice question-answering (MCQ) testbed for source-grounded LALM understanding and reasoning. Each item pairs an English target dialogue with a concurrent distractor in English, Spanish, Korean, or Chinese, covering both same-language and cross-lingual

Category	Literature	Task	Gap relative to MUSI
Multi-speaker ASR & Separation	CHiME-6 (Watanabe et al., 2020)	Multi-speaker far-field ASR and diarization	Transcription/diarization quality, not source-grounded reasoning
	WSJ0-2mix (Hershey et al., 2016) / LibriMix (Cosentino et al., 2020)	Single-channel two-speaker separation	Signal reconstruction (SI-SDR); no LALM reasoning
Cocktail Party Listening	TLE (Borsdorf et al., 2021)	Target-language extraction in multilingual mixtures	Language-conditioned separation; no MCQ reasoning or source attribution
	MCoRec (Nguyen et al., 2026)	Multi-modal multi-party conversation transcription	Transcription- and clustering-oriented; no reasoning or source attribution
General LALM Evaluation	AIR-Bench (Yang et al., 2024)	LALM audio understanding (MCQ + chat)	Clean single-stream audio; no concurrent distractors
	AudioBench (Wang et al., 2025)	Universal LALM benchmark, 8 tasks	Single-source audio; no multilingual interference
	VoiceBench (Chen et al., 2026)	Voice assistant evaluation under varied conditions	Acoustic/speaker variations but no semantic concurrent distractors
Trustworthiness & Reasoning	AudioTrust (Li et al., 2025)	LALM trustworthiness across 6 dimensions	Non-semantic acoustic risks; not source misattribution
	SAKURA (Yang et al., 2025)	Multi-hop reasoning over audio attributes	Single-stream input; no selective attention under interference
Source-grounded QA	MUSI (Ours)	MCQ-based selective attention with concurrent multilingual semantic distractors	—

Table 1: Positioning of MUSI across related work, from signal-level multi-speaker ASR and source separation to recent LALM benchmarks.

interference (Chiswick and Miller, 2005; Littell et al., 2017). We evaluate six LALMs under single-stream, separation-based two-stage, and end-to-end cocktail party settings, with controlled target-to-distractor SNRs.

Our results reveal substantial model-dependent variation across distractor languages and strong sensitivity to adverse SNRs. Errors are dominated by distractor-grounded source confusion, while confidence becomes miscalibrated under interference. Although separation reduces acoustic overlap, it often leaves source attribution unresolved and yields confident wrong-stream answers.

In summary, our contributions are as follows:

- We introduce MUSI, a multilingual controlled diagnostic evaluation for target-grounded reasoning under plausible concurrent speech distractors.
- We characterize model-dependent variation across distractor languages and trace model errors to the target, distractor, or neither stream.
- We show that strong single-stream performance does not ensure cocktail party robustness and that separation-based inference retains a source-attribution bottleneck.

2 MUSI Dataset

2.1 Dataset Composition

MUSI covers four safety-critical domains: Aviation & Maritime, Healthcare & Medicine, Finance & Business, and Construction & Plant. Each item contains an English target dialogue, a same-domain

semantically plausible distractor in English, Spanish, Korean, or Chinese, a target-grounded question, and four answer options. The question is answerable only from the target, whereas the distractor provides plausible but task-irrelevant evidence. MUSI contains 200 cases, with 50 per domain. To control speaker-related confounders, all utterances are synthesized using OpenAI API-based TTS (Hurst et al., 2024) with fixed speaker assignments. We use a male target and female distractor across all distractor-language conditions. Mixtures use a default target-to-distractor SNR of 0 dB, with additional levels for robustness analysis.

2.2 Dataset Statistics

Target–distractor cosine similarity, measured using multilingual-e5-large (Wang et al., 2024), ranges from 0.796 to 0.842 across distractor languages, indicating broadly comparable semantic similarity across conditions. However, non-English distractors differ in duration: Spanish and Korean distractors are approximately 3.3 s and 2.6 s longer, respectively, than the 21.4 s target average, which may affect temporal overlap. Separation quality is strongly SNR-dependent: average SI-SDR increases from 2.95 dB at −10 dB to 16.16 dB at +10 dB, with greater language-dependent variation at lower SNRs. These statistics contextualize language-conditioned performance differences beyond distractor language alone. Further details are provided in Appendix A.

Group	Model	Setting	Target	Distractor				Average	Lang. Gap
			English	English	Spanish	Korean	Chinese		
Open-weight	Qwen2-Audio	Single	0.773 / 0.132	–	–	–	–	0.773 / 0.132	–
		Separation	–	0.482 / 0.497	0.477 / 0.497	0.707 / 0.245	0.450 / 0.527	0.529 _(−24.4) / 0.441	25.7
		Cocktail	–	0.550 / 0.270	0.412 / 0.418	0.683 / 0.156	0.220 / 0.627	0.466 _(−30.7) / 0.366	46.3
	MERaLiON-2	Single	0.757 / 0.176	–	–	–	–	0.757 / 0.176	–
		Separation	–	0.592 / 0.385	0.678 / 0.290	0.768 / 0.187	0.732 / 0.229	0.693 _(−6.5) / 0.273	17.6
		Cocktail	–	0.500 / 0.371	0.608 / 0.270	0.655 / 0.222	0.642 / 0.231	0.601 _(−15.6) / 0.274	15.5
	Audio-Flamingo-3	Single	0.908 / 0.030	–	–	–	–	0.908 / 0.030	–
		Separation	–	0.750 / 0.199	0.705 / 0.241	0.827 / 0.112	0.752 / 0.199	0.758 _(−15.0) / 0.187	12.2
		Cocktail	–	0.560 / 0.164	0.440 / 0.268	0.753 / 0.027	0.567 / 0.163	0.580 _(−32.8) / 0.141	31.3
	Qwen2.5-Omni	Single	0.650 / 0.061	–	–	–	–	0.650 / 0.061	–
		Separation	–	0.358 / 0.458	0.500 / 0.245	0.635 / 0.104	0.577 / 0.142	0.518 _(−13.3) / 0.236	27.7
		Cocktail	–	0.250 / 0.365	0.300 / 0.254	0.547 / 0.046	0.307 / 0.222	0.351 _(−29.9) / 0.216	29.7
Closed-source	GPT-4o mini Audio	Single	0.772 / –	–	–	–	–	0.772 / –	–
		Separation	–	0.553 / –	0.577 / –	0.592 / –	0.623 / –	0.586 _(−18.6) / –	7.0
		Cocktail	–	0.587 / –	0.593 / –	0.682 / –	0.683 / –	0.636 _(−13.6) / –	9.6
	Gemini-2.0-Flash	Single	0.955 / –	–	–	–	–	0.955 / –	–
		Separation	–	0.892 / –	0.967 / –	0.973 / –	0.977 / –	0.952 _(−0.3) / –	8.5
		Cocktail	–	0.677 / –	0.043 / –	0.077 / –	0.172 / –	0.242 _(−71.3) / –	63.4

Table 2: Main performance across three settings at 0 dB target-to-distractor SNR. Each cell reports Accuracy ($\uparrow$) / ECE ($\downarrow$). ECE is reported only for open-weight models with accessible token-level log-probabilities. Subscripted values in the Average column denote absolute accuracy changes, in percentage points (pp), relative to the single-stream setting. Language Gap (Lang. Gap, $\downarrow$) denotes the best–worst accuracy difference across distractor-language conditions, also in pp.

2.3 Diagnostic Error Taxonomy

Each item has one target-grounded correct option and three incorrect options:

- **Target Misreasoning (Mis)**: an incorrect answer grounded in the target stream but reflecting faulty reasoning.
- **Distractor Interference (Int)**: an incorrect answer supported by the distractor stream.
- **Ungrounded Inference (Ung)**: an incorrect answer unsupported by either stream.

3 Experimental Setup

3.1 Models

We evaluate six LALMs: two closed-source models, GPT-4o mini Audio (Hurst et al., 2024) and Gemini-2.0-Flash (Team et al., 2023), and four open-weight models, Qwen2-Audio (Chu et al., 2024), MERaLiON-2 (MERaLiON Team, 2024), Audio-Flamingo-3 (Goel et al., 2025), and Qwen2.5-Omni (Xu et al., 2025). We follow each model’s reported decoding configuration where applicable.

3.2 Evaluation Settings and Metrics

We evaluate three settings: single, where models receive only the single English target stream; two-stage separation, where mixtures are first separated by ClearerVoice-Studio (Zhao et al., 2025) and

then evaluated with separate streams; and cocktail party, where models receive the mono (diotic) mixture (You et al., 2026) and must answer from the male English target speech while ignoring the female multilingual distractor. For SNR analysis, we additionally evaluate mixture-based settings at target-to-distractor SNRs of -10, -5, 0, +5, and +10 dB. We evaluate three fixed option permutations per question across all models, distractor languages, and SNR settings to reduce option-order bias (Lin et al., 2025). We report accuracy for all models and additional ECE for open-weight models with token-level log-probabilities. Decoding configuration, full prompts, separation quality, and details on metrics are provided in Appendix B.

4 Results and Analysis

4.1 Overall Performance under Multilingual Interference

Table 2 reports the main 0 dB results. Single-stream performance is substantially higher than cocktail party performance for every model (e.g., Gemini-2.0-Flash: 0.955 $\rightarrow$ 0.242), showing that single-stream understanding does not reliably transfer to source-grounded reasoning under concurrent multilingual speech.

Separation-based inference generally improves over cocktail party inference but does not restore

single-stream accuracy, suggesting that reducing acoustic overlap is insufficient when source attribution remains unresolved. Gemini-2.0-Flash is a notable exception: separation nearly recovers single-stream accuracy (0.952), yet its cocktail party accuracy is the lowest among all models, suggesting a strong reliance on explicit separate streams. For open-weight models, ECE also rises sharply under interference (e.g., 0.030 $\rightarrow$ 0.141 for Audio-Flamingo-3), which indicates that interference degrades calibration in addition to accuracy.

These differences are not readily explained by measured text-embedding similarity alone (0.796–0.842; Table 4) and may also reflect phonetic, temporal, and model-specific factors.

4.2 Target-Dominance Effects across SNRs

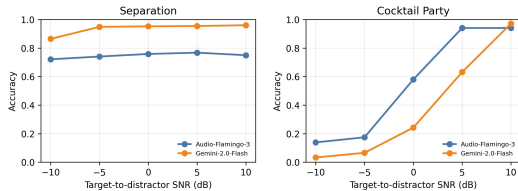


Figure 2: Average accuracy across target-to-distractor SNRs for two models under separation-based (left) and cocktail party (right) settings.

Figure 2 illustrates SNR-dependent robustness for both Audio-Flamingo-3 (open-weight) and Gemini-2.0-Flash (closed-source). In the separation-based setting, both models remain relatively stable across SNRs, suggesting that explicit stream separation effectively mitigates acoustic overlap. In contrast, end-to-end cocktail party inference is highly SNR-sensitive: both models degrade sharply under negative SNRs and recover as the target becomes acoustically dominant. This suggests that, without explicit separation, source grounding is strongly influenced by acoustic dominance. At high positive SNRs, cocktail-party accuracy occasionally exceeds single-stream accuracy by a small margin; we do not interpret these differences as systematic gains, as they may reflect finite-sample and decoding variability.

4.3 Evidence-Source Error Analysis

Table 3 shows that the dominant failure mode shifts from target misreasoning in the single setting to distractor-grounded interference under cocktail party inference. The Int ratio reaches 0.769 for Qwen2-Audio, 0.869 for Audio-Flamingo-3, and

Model	Condition	Mis ↓	Int ↓	Ung ↓
Qwen2-Audio	Single	0.875	–	0.125
	Cocktail Party	0.180	0.769	0.051
	w/ English	0.152	0.827	0.021
	w/ Spanish	0.154	0.806	0.040
	w/ Korean	0.358	0.504	0.138
	w/ Chinese	0.120	0.846	0.035
Audio-Flamingo-3	Single	0.982	–	0.018
	Cocktail Party	0.099	0.869	0.032
	w/ English	0.081	0.906	0.013
	w/ Spanish	0.092	0.886	0.022
	w/ Korean	0.183	0.741	0.076
	w/ Chinese	0.062	0.911	0.027
Gemini-2.0-Flash	Single	0.963	–	0.037
	Cocktail Party	0.072	0.918	0.009
	w/ English	0.098	0.893	0.008
	w/ Spanish	0.067	0.926	0.007
	w/ Korean	0.052	0.939	0.008
	w/ Chinese	0.078	0.907	0.015

Table 3: Error-type distribution in the single and cocktail party (0 dB) settings. Mis: target misreasoning; Int: distractor interference; Ung: ungrounded inference. Single has no distractor stream, so Int is not applicable.

0.918 for Gemini-2.0-Flash, while Ung remains below 0.052. Errors therefore arise primarily from following the competing stream rather than unsupported inference.

For open-weight models, Korean distractors consistently yield lower interference and higher accuracy than other languages. This pattern is not readily explained by text-embedding similarity alone, as Korean similarity (0.815) is comparable to other conditions (Table 4). Longer duration, potentially reduced temporal overlap, acoustic-phonetic properties, and higher separated-target SI-SDR may jointly contribute to this disparity. This model-dependent variation indicates that selective-attention reliability is not uniform across distractor languages. Further breakdowns and stream-selection diagnostics are provided in Appendix C.

5 Conclusion

We introduce MUSI, a multilingual MCQ testbed for evaluating source-grounded LALMs under cocktail-party interference. Across six models, concurrent speech degrades performance through distractor-grounded source confusion and SNR sensitivity, with additional variation across distractor languages. These results highlight selective auditory attention as an important capability for reliable LALMs.

6 Limitations

Scale and ecological validity. MUSI comprises 200 TTS-synthesized cases (50 per domain), enabling controlled evaluation but limiting subgroup analysis. The synthetic two-speaker mixtures also omit natural conversational phenomena such as spontaneous overlap, interruptions, variable turn timing, spatial cues, channel variation, and richer prosody.

Scope of conditions. Targets are restricted to English dialogues and two-speaker mono mixtures with a single off-the-shelf separator (Zhao et al., 2025). We use a fixed male-target/female-distractor configuration and do not counterbalance speaker gender or evaluate same-gender mixtures. This setting intentionally limits additional speaker-identity variation while focusing on multilingual semantic interference. Future work should evaluate non-English targets, bidirectional language pairs, alternative separators, and counterbalanced or same-gender speaker configurations to determine how speaker similarity interacts with selective auditory attention.

Scope of disparity analysis. Our analysis concerns language-conditioned model performance, not demographic fairness, and should not be interpreted as reflecting intrinsic properties of languages or speaker communities.

Residual confounds. Despite reporting length, similarity, and duration statistics (Appendix A), uncontrolled factors such as entity overlap, phonetic similarity, TTS voice characteristics, and temporal alignment may influence cross-lingual performance.

Methodological asymmetry. Open-weight models select streams via confidence comparison, whereas closed-source models receive both streams sequentially in one prompt. This procedural difference, along with the restriction of ECE analysis to open-weight models, limits cross-group comparability.

Evaluation format. The MCQ format enables controlled source attribution but does not capture free-form generation, where models may blend evidence across streams. Extending this testbed to open-ended responses remains future work.

References

Marvin Borsdorf, Haizhou Li, and Tanja Schultz. 2021. Target language extraction at multilingual cocktail parties. In *2021 IEEE Automatic Speech Recognition*

and Understanding Workshop (ASRU), pages 717–724. IEEE.

Adelbert W Bronkhorst. 2015. The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, 77(5):1465–1487.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2026. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398.

Hao Cheng, Erjia Xiao, Jing Shao, Yichi Wang, Le Yang, Chao Shen, Philip Torr, Jindong Gu, and Renjing Xu. 2025. Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models. *arXiv preprint arXiv:2501.13772*.

Edward Collin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *Journal of the acoustical society of America*, 25:975–979.

Barry R Chiswick and Paul W Miller. 2005. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of multilingual and multicultural development*, 26(1):1–11.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 31–35. IEEE.

Guanyu Hou, Jiaming He, Yinhang Zhou, Ji Guo, Yitong Qiao, Rui Zhang, and Wenbo Jiang. 2025. Evaluating robustness of large audio language models to audio injection: An empirical study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25671–25687.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

- Heejoon Koo, Yoon Tae Kim, Miika Toikkanen, and June-Woo Kim. 2026a. Mitigating stethoscope-induced shortcuts in respiratory sound classification under federated domain generalization with causality-inspired interventions. *arXiv preprint arXiv:2605.29862*.
- Heejoon Koo, Miika Toikkanen, Yoon Tae Kim, Soo Yong Kim, and June-Woo Kim. 2026b. Empowering multimodal respiratory sound classification with counterfactual adversarial debiasing for out-of-distribution robustness. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 14967–14971. IEEE.
- Kai Li, Can Shen, Yile Liu, Jirui Han, Kelong Zheng, Xuechao Zou, Zhe Wang, Shun Zhang, Xingjian Du, Hanjun Luo, and 1 others. 2025. Audiobench: Benchmarking the multifaceted trustworthiness of audio large language models. *arXiv preprint arXiv:2505.16211*.
- Yu-Xiang Lin, Chen-An Li, Sheng-Lun Wei, Po-Chun Chen, Hsin-Hsi Chen, and Hung-yi Lee. 2025. Hearing the order: Investigating selection bias in large audio-language models. *arXiv preprint arXiv:2510.00628*.
- Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.
- Josh H McDermott. 2009. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027.
- MERaLiON Team. 2024. [Meralion-audiollm: Bridging audio and language with large language models](#). Preprint, arXiv:2412.09818.
- Thai-Binh Nguyen, Katerina Zmolikova, Pingchuan Ma, Ngoc Quan Pham, Christian Fuegen, and Alexander Waibel. 2026. A cocktail-party benchmark: Multimodal dataset and comparative evaluation results. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19502–19506. IEEE.
- Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Zeren Luo, Jingyi Zheng, Wenhan Dong, Xinlei He, Xuechao Wang, Yingjie Xue, and 1 others. 2025. Jalmbench: Benchmarking jailbreak vulnerabilities in audio language models. *arXiv preprint arXiv:2505.17568*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, and 1 others. 2025. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *arXiv preprint arXiv:2505.15406*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 2025. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, and 1 others. 2020. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *Proc. CHiME 2020*, pages 1–7.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung-yi Lee. 2025. Sakura: On the multi-hop reasoning of large audio-language models based on speech and audio information. In *Proc. Interspeech 2025*, pages 1788–1792.
- Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Airbench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.
- Yuhuan You, Lai Wei, Xihong Wu, and Tianshu Qu. 2026. The world is not mono: Enabling spatial understanding in large audio-language models. *arXiv preprint arXiv:2601.02954*.
- Shengkui Zhao, Zexu Pan, and Bin Ma. 2025. Clearvoice-studio: Bridging advanced speech processing research and practical deployment. *arXiv preprint arXiv:2506.19398*.

A Additional Dataset Details

A.1 Construction

We construct each item through a staged generation-and-revision pipeline: domain-specific topic and target dialogue generation, same-domain distractor generation, and diagnostic answer option construction. Text dialogues and options were produced through manual writing and LLM-assisted drafting, then manually revised. Each script is carefully streamlined to cap each stream at 30 s. All audio is sampled at 16 kHz and presented as mono (diotic) mixtures, as evaluated LALMs lack the ability to process stereo (dichotic) mixtures (You et al., 2026).

A.2 Quality Control

Each item was independently validated by two annotators for target-only answerability, unique target grounding of the correct option, distractor-only support for interference options, and absence of support for ungrounded options. Annotators also checked that each question required evidence integration across target turns and did not rely on superficial lexical cues. Items were revised or removed unless all options could be traced to explicit evidence, and disagreements were resolved through author adjudication.

For Spanish, Korean, and Chinese distractors, translations were checked against the English distractor for entity preservation, event-sequence consistency, numerical consistency, and measurement-unit consistency. We used LLM-assisted back-translation as a secondary screen and manually reviewed flagged cases. We further checked for option leakage by verifying that correct, distractor-grounded, and ungrounded options were supported only by their intended evidence sources. Items with ambiguous grounding, lexical shortcuts, or cross-option leakage were revised or removed.

A.3 Statistics

Table 4 reports compact dataset statistics used to assess potential length, semantic-overlap, and duration confounds across distractor languages. Script length and question length are stable by construction, while option length remains closely matched across distractor-language conditions. Target–distractor similarity, measured by multilingual-e5-large (Wang et al., 2024), is broadly comparable across languages, although non-English distractors show slightly lower similarity than English, which may reflect cross-lingual embedding alignment as well as residual linguistic differences.

Stream	Language	Script Length	Question Length	Option Length	Target–Distractor Similarity $\uparrow$	Duration
Target	English	45.8 $\pm$ 4.6 words	13.4 $\pm$ 2.2 words	12.1 $\pm$ 2.1 words	–	21.4 $\pm$ 2.3 s
Distractor	English	45.1 $\pm$ 4.9 words	–	11.6 $\pm$ 2.0 words	0.842	21.7 $\pm$ 2.8 s
	Spanish	49.6 $\pm$ 4.9 words	–	11.2 $\pm$ 2.0 words	0.809	24.7 $\pm$ 2.7 s
	Korean	158.4 $\pm$ 18.8 chars	–	11.4 $\pm$ 2.0 words	0.815	24.0 $\pm$ 2.9 s
	Chinese	81.8 $\pm$ 16.8 chars	–	11.3 $\pm$ 2.0 words	0.796	22.1 $\pm$ 4.3 s

Table 4: Dataset statistics including script length, target–distractor embedding similarity, and duration. Length is measured in words for English/Spanish and characters for Korean/Chinese.

A.4 Separation Quality

For the separation-based setting, we assess separated target quality against the corresponding single target audio using scale-invariant signal-to-distortion ratio (SI-SDR) and target automatic speech recognition word error rate (ASR WER). WER is computed between the reference target transcript and the Whisper-v3-large transcription of the separated stream (Radford et al., 2023).

Table 5 shows that separation quality improves with target dominance: average SI-SDR rises from 2.95 dB at -10 dB to 16.16 dB at $+10$ dB, while average WER drops from 0.251 to 0.066. The -10 dB condition shows the clearest language dependence: Korean distractors retain the best separated target quality, whereas Spanish and Chinese distractors yield substantially lower SI-SDR and higher WER. These differences narrow at 0 dB and above and indicate that language-dependent separation difficulty is most pronounced when the target is acoustically disadvantaged.

Distractor	-10 dB		-5 dB		0 dB		+5 dB		+10 dB	
	SI ↑	WER ↓	SI ↑	WER ↓	SI ↑	WER ↓	SI ↑	WER ↓	SI ↑	WER ↓
English	2.48	0.221	11.76	0.105	14.22	0.083	15.70	0.076	16.15	0.069
Spanish	1.46	0.372	12.21	0.180	14.66	0.148	15.76	0.121	16.15	0.088
Korean	7.56	0.147	12.94	0.072	14.73	0.062	15.78	0.058	16.19	0.051
Chinese	0.28	0.264	10.96	0.084	14.39	0.058	15.73	0.049	16.13	0.056
Average	2.95	0.251	11.97	0.110	14.50	0.088	15.74	0.076	16.16	0.066

Table 5: Separated target-stream quality across target-to-distractor SNRs. SI denotes SI-SDR.

A.5 A Multilingual Dataset Example from Healthcare & Medicine

Example Item

Target stream (English). Patient Room 402 displays symptomatic hypotension requiring an immediate vasopressor infusion of twenty micrograms per kilogram. Central venous pressure readings currently exceed eighteen millimeters of mercury, indicating a significant risk of fluid volume overload. The care team shifts to restrictive fluid management while maintaining a baseline titration of only five micrograms.

English distractor. Recovery Suite B patient reports intense localized pain, prompting consideration of an immediate ten-milligram intravenous morphine dose. After the initial analgesic administration, respiratory monitoring detects a decline to eight breaths per minute. Cancel the opioid order immediately and prepare the naloxone protocol to stabilize the patient's breathing.

Question. Which observed clinical indicator led to the rejection of the proposed twenty-microgram vasopressor dosage?

Answer options.

- (A) Respiratory rate measurements declining to eight breaths per minute after analgesic exposure.
- (B) Central venous pressure readings exceeding eighteen millimeters, indicating fluid overload risk.
- (C) Reports of intense localized pain prompting opioid consideration.
- (D) Electrolyte disturbances requiring counter-regulatory hormone administration.

Correct answer. B

Diagnostic interpretation. Option B is grounded in the target stream. Options A and C are distractor-grounded, whereas Option D is unsupported by either stream.

Korean distractor. 회복실 B의 환자가 극심한 국소 통증을 호소하여 정맥주사용 모르핀 10밀리그램의 즉시 투여가 제안되었습니다. 초기 진통제 투여 직후 호흡수가 분당 8회로 감소한 것이 감지되었습니다. 마약성 진통제 처방을 즉시 취소하고 환자의 호흡을 안정시키기 위한 나록손 프로토콜을 준비하십시오.

Question. Which observed clinical indicator led to the rejection of the proposed twenty-microgram vasopressor dosage?

Answer options.

- (A) Reports of intense localized pain prompting opioid consideration.
- (B) Respiratory rate measurements declining to eight breaths per minute after analgesic exposure.
- (C) Central venous pressure readings exceeding eighteen millimeters, indicating fluid overload risk.
- (D) Electrolyte disturbances requiring counter-regulatory hormone administration.

Correct answer. C

Diagnostic interpretation. Option C is grounded in the target stream. Options A and B are distractor-grounded, whereas Option D is unsupported by either stream.

Spanish distractor. El paciente de la Sala de Recuperación B refiere dolor intenso y localizado, por lo que se propone administrar de inmediato diez miligramos de morfina por vía intravenosa. Tras el intento inicial de analgesia, los sensores detectan una disminución de la frecuencia respiratoria a ocho respiraciones por minuto. Cancele de inmediato la orden de opioides y prepare el protocolo de naloxona para estabilizar la respiración del paciente.

Question. Which observed clinical indicator led to the rejection of the proposed twenty-microgram vasopressor dosage?

Answer options.

- (A) Electrolyte disturbances requiring counter-regulatory hormone administration.
- (B) Reports of intense localized pain prompting opioid consideration.
- (C) Central venous pressure readings exceeding eighteen millimeters, indicating fluid overload risk.
- (D) Respiratory rate measurements declining to eight breaths per minute after analgesic exposure.

Correct answer. C

Diagnostic interpretation. Option C is grounded in the target stream. Options B and D are distractor-grounded, whereas Option A is unsupported by either stream.

Chinese distractor. 恢复室B的患者报告剧烈局部疼痛，建议立即静脉注射10毫克吗啡。初次尝试镇痛给药后，呼吸频率监测显示其呼吸降至每分钟8次。请立即取消阿片类镇痛药医嘱，并准备纳洛酮处置方案，以稳定患者的呼吸状态。

Question. Which observed clinical indicator led to the rejection of the proposed twenty-microgram vasopressor dosage?

Answer options.

- (A) Central venous pressure readings exceeding eighteen millimeters, indicating fluid overload risk.
- (B) Respiratory rate measurements declining to eight breaths per minute after analgesic exposure.
- (C) Electrolyte disturbances requiring counter-regulatory hormone administration.
- (D) Reports of intense localized pain prompting opioid consideration.

Correct answer. A

Diagnostic interpretation. Option A is grounded in the target stream. Options B and D are distractor-grounded, whereas Option C is unsupported by either stream.

B Additional Experimental Details

B.1 Decoding Parameters

We follow each model’s reported generation configurations, which are summarized in Table 6.

Model	do_sample	Temp.	Top- p	Top- k	Max tokens
Qwen2-Audio	True	0.7	0.5	20	–
Qwen2.5-Omni	False	1.0	1.0	50	–
MERaLiON-2	–	–	–	–	256
Audio-Flamingo-3	True	0.7	0.9	–	256

Table 6: Reported decoding parameters for open-weight LALMs. “–” indicates that the parameter is not explicitly specified in the model card or example used as reference.

B.2 Source-grounding Prompts

Source-Grounding Prompt for Cocktail Party Evaluation

You are an audio-language model being evaluated under a cocktail party test. You will receive:

- One audio stream explicitly labeled as TARGET, spoken by a male English speaker.
- Zero or more audio streams explicitly labeled as DISTRACTOR.
- A multiple-choice question with four options.

Instructions.

- Base your answer exclusively on the audio labeled TARGET.
- Focus only on the male English speech in the TARGET audio.
- Completely ignore all audio labeled DISTRACTOR, regardless of language, speaker gender, or salience.
- Do not use information from any unlabeled or distractor audio.
- The correct answer requires integrating multiple pieces of evidence from the TARGET audio.
- Select exactly one option.

Output rules.

- Respond with a single capital letter only: A, B, C, or D.
- Do not provide explanations, reasoning steps, or additional text.

B.3 Separation-based Inference

The separation-based setting isolates acoustic overlap from source attribution: if separation restores single-level accuracy, the bottleneck is acoustic; if not, the model itself struggles to ground its reasoning in the correct stream. For open-weight models, both separated streams are processed independently, and the final response is selected from the stream with the higher normalized option confidence. This setup further diagnoses whether the model assigns greater confidence to the distractor stream than to the target

stream. For closed-source models, where token-level confidence is unavailable, both separated streams are provided sequentially in one prompt, and the model must generate the answer.

B.4 Confidence and Calibration

For open-weight models, confidence is the normalized probability of the selected option after applying a softmax over the log-probabilities of $\{A, B, C, D\}$. ECE is computed with $M = 10$ equally spaced bins:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (1)$$

where N is the number of examples, B_m is bin m , and acc and conf denote empirical accuracy and mean confidence within the bin.

C Additional Results and Analyses

C.1 Confidence-based Stream Selection

Table 7 shows that separation-based inference remains limited by source attribution. Although accuracy is high when the target stream is selected, open-weight models often assign higher confidence to the distractor stream. At 0 dB, target-selection ratios are 59.0% for Qwen2-Audio, 78.5% for MERaLiON-2, 78.1% for Audio-Flamingo-3, and 68.9% for Qwen2.5-Omni.

When the distractor stream is selected, accuracy drops sharply while ECE becomes large, indicating confident wrong-stream predictions. In particular, this misselection persists even at high SNRs, where separation quality is strong (e.g., SI-SDR > 16 dB at +10 dB; Table 5), suggesting that model-level source preference contributes beyond separation quality alone. Korean distractors generally yield higher target-selection ratios than the other distractor languages, consistent with the main accuracy results, although this pattern may reflect multiple factors including phonetic overlap, duration, and separation quality.

C.2 SNR-wise Error-Type Distributions

Table 8 shows that error types vary systematically with SNR. In the cocktail party setting, negative SNRs are dominated by distractor-grounded interference errors, indicating that models tend to follow the acoustically dominant competing stream. As the target becomes louder, the interference ratio decreases and target-misreasoning errors become more frequent. Ungrounded errors remain consistently below 0.15 across all SNR conditions, suggesting that failures are predominantly source-confusion errors rather than unsupported inferences. In the separation setting, error distributions are more stable across SNRs, but distractor-grounded errors remain common, further supporting the presence of a source-attribution bottleneck.

C.3 Domain- and SNR-wise Performance

Table 9 provides domain- and SNR-wise performance breakdowns. The overall trend is consistent across domains: separation-based inference remains comparatively stable, whereas cocktail party inference is highly sensitive to target-to-distractor SNR. Negative SNRs substantially reduce accuracy, while positive SNRs improve performance as the target becomes acoustically dominant. This pattern indicates that the observed vulnerability is not restricted to a single domain, but reflects a broader limitation in source-grounded reasoning under concurrent speech. Among domains, Construction & Plant tends to yield the lowest cocktail party accuracy at severe SNRs (e.g., 0.097 for Audio-Flamingo-3 at −10 dB), possibly reflecting the confusability of its procedural and numerical distractor content.

Model	Condition	Distractor	SNR				
			-10 dB	-5 dB	0 dB	+5 dB	+10 dB
Qwen2-Audio	Separation	Overall	100.0%; 0.507 / 0.445	100.0%; 0.516 / 0.451	100.0%; 0.529 / 0.441	100.0%; 0.544 / 0.425	100.0%; 0.537 / 0.429
	Target Selected	Overall	59.7%; 0.813 / 0.144	58.9%; 0.867 / 0.103	59.0%; 0.884 / 0.089	61.3%; 0.879 / 0.094	60.4%; 0.879 / 0.092
		w/ English	54.8%; 0.790 / 0.179	55.2%; 0.870 / 0.105	53.8%; 0.895 / 0.088	59.2%; 0.893 / 0.094	56.8%; 0.886 / 0.087
		w/ Spanish	52.5%; 0.787 / 0.172	53.3%; 0.856 / 0.118	53.8%; 0.882 / 0.098	55.2%; 0.858 / 0.115	55.0%; 0.882 / 0.099
		w/ Korean	78.3%; 0.872 / 0.088	76.8%; 0.868 / 0.096	80.2%; 0.863 / 0.100	79.0%; 0.859 / 0.105	80.3%; 0.859 / 0.105
		w/ Chinese	53.2%; 0.774 / 0.181	50.3%; 0.874 / 0.106	48.3%; 0.910 / 0.078	51.7%; 0.916 / 0.069	49.5%; 0.899 / 0.080
	Distractor Selected	Overall	40.3%; 0.054 / 0.896	41.1%; 0.012 / 0.950	41.0%; 0.016 / 0.949	38.8%; 0.014 / 0.950	39.6%; 0.017 / 0.946
		w/ English	45.2%; 0.041 / 0.913	44.8%; 0.000 / 0.971	46.2%; 0.000 / 0.973	40.8%; 0.008 / 0.968	43.2%; 0.004 / 0.965
		w/ Spanish	47.5%; 0.042 / 0.912	46.7%; 0.000 / 0.967	46.2%; 0.004 / 0.967	44.8%; 0.000 / 0.973	45.0%; 0.007 / 0.964
		w/ Korean	21.7%; 0.069 / 0.853	23.2%; 0.079 / 0.825	19.8%; 0.076 / 0.838	21.0%; 0.063 / 0.834	19.7%; 0.085 / 0.816
		w/ Chinese	46.8%; 0.071 / 0.884	49.7%; 0.003 / 0.974	51.7%; 0.019 / 0.957	48.3%; 0.010 / 0.964	50.5%; 0.010 / 0.964
MERaLiON-2	Separation	Overall	100.0%; 0.683 / 0.277	100.0%; 0.685 / 0.281	100.0%; 0.693 / 0.273	100.0%; 0.687 / 0.281	100.0%; 0.688 / 0.278
	Target Selected	Overall	76.7%; 0.853 / 0.114	77.0%; 0.883 / 0.088	78.5%; 0.878 / 0.093	78.2%; 0.875 / 0.097	77.8%; 0.878 / 0.093
		w/ English	67.7%; 0.852 / 0.120	67.2%; 0.888 / 0.098	67.5%; 0.874 / 0.107	66.8%; 0.888 / 0.097	66.3%; 0.872 / 0.116
		w/ Spanish	76.0%; 0.862 / 0.104	73.8%; 0.883 / 0.089	77.2%; 0.875 / 0.098	76.0%; 0.882 / 0.097	75.0%; 0.889 / 0.084
		w/ Korean	84.5%; 0.858 / 0.107	83.8%; 0.883 / 0.082	86.2%; 0.884 / 0.084	86.8%; 0.856 / 0.111	86.2%; 0.857 / 0.110
		w/ Chinese	78.7%; 0.839 / 0.124	83.0%; 0.878 / 0.093	83.2%; 0.878 / 0.089	83.0%; 0.880 / 0.090	83.5%; 0.894 / 0.081
	Distractor Selected	Overall	23.3%; 0.125 / 0.815	23.0%; 0.024 / 0.925	21.5%; 0.016 / 0.929	21.8%; 0.011 / 0.938	22.2%; 0.022 / 0.925
		w/ English	32.3%; 0.082 / 0.881	32.8%; 0.020 / 0.954	32.5%; 0.005 / 0.963	33.2%; 0.005 / 0.970	33.7%; 0.010 / 0.960
		w/ Spanish	24.0%; 0.139 / 0.816	26.2%; 0.000 / 0.953	22.8%; 0.015 / 0.941	24.0%; 0.000 / 0.951	25.0%; 0.013 / 0.937
		w/ Korean	15.5%; 0.032 / 0.882	16.2%; 0.031 / 0.881	13.8%; 0.048 / 0.847	13.2%; 0.051 / 0.878	13.8%; 0.060 / 0.849
		w/ Chinese	21.3%; 0.242 / 0.672	17.0%; 0.059 / 0.879	16.8%; 0.010 / 0.916	17.0%; 0.010 / 0.905	16.5%; 0.030 / 0.899
Audio-Flamingo-3	Separation	Overall	100.0%; 0.721 / 0.194	100.0%; 0.740 / 0.203	100.0%; 0.758 / 0.187	100.0%; 0.767 / 0.180	100.0%; 0.750 / 0.196
	Target Selected	Overall	73.8%; 0.938 / 0.017	76.4%; 0.967 / 0.020	78.1%; 0.970 / 0.018	79.1%; 0.970 / 0.021	77.3%; 0.969 / 0.023
		w/ English	71.8%; 0.947 / 0.018	75.0%; 0.973 / 0.023	76.8%; 0.976 / 0.024	77.7%; 0.974 / 0.019	73.3%; 0.968 / 0.026
		w/ Spanish	69.7%; 0.923 / 0.024	71.7%; 0.953 / 0.025	73.8%; 0.955 / 0.019	74.8%; 0.958 / 0.018	73.8%; 0.962 / 0.030
		w/ Korean	84.5%; 0.955 / 0.024	84.2%; 0.956 / 0.021	85.3%; 0.965 / 0.018	88.3%; 0.962 / 0.015	86.5%; 0.965 / 0.017
		w/ Chinese	69.2%; 0.925 / 0.027	74.8%; 0.987 / 0.038	76.5%; 0.983 / 0.027	75.5%; 0.987 / 0.033	75.7%; 0.982 / 0.028
	Distractor Selected	Overall	26.2%; 0.108 / 0.779	23.6%; 0.005 / 0.918	21.9%; 0.004 / 0.918	20.9%; 0.002 / 0.923	22.7%; 0.000 / 0.923
		w/ English	28.2%; 0.112 / 0.785	25.0%; 0.007 / 0.930	23.2%; 0.000 / 0.928	22.3%; 0.000 / 0.932	26.7%; 0.000 / 0.934
		w/ Spanish	30.3%; 0.093 / 0.794	28.3%; 0.000 / 0.931	26.2%; 0.000 / 0.934	25.2%; 0.007 / 0.926	26.2%; 0.000 / 0.931
		w/ Korean	15.5%; 0.022 / 0.871	15.8%; 0.021 / 0.872	14.7%; 0.023 / 0.850	11.7%; 0.000 / 0.871	13.5%; 0.000 / 0.884
		w/ Chinese	30.8%; 0.162 / 0.712	25.2%; 0.000 / 0.921	23.5%; 0.000 / 0.931	24.5%; 0.000 / 0.938	24.3%; 0.000 / 0.923
Qwen2.5-Omni	Separation	Overall	100.0%; 0.475 / 0.238	100.0%; 0.517 / 0.239	100.0%; 0.517 / 0.236	100.0%; 0.522 / 0.230	100.0%; 0.543 / 0.211
	Target Selected	Overall	65.0%; 0.695 / 0.058	68.0%; 0.748 / 0.060	68.9%; 0.744 / 0.043	69.2%; 0.748 / 0.066	71.5%; 0.754 / 0.053
		w/ English	44.2%; 0.694 / 0.104	43.8%; 0.768 / 0.084	47.2%; 0.756 / 0.059	49.7%; 0.768 / 0.076	50.5%; 0.762 / 0.076
		w/ Spanish	63.3%; 0.697 / 0.067	66.2%; 0.738 / 0.071	66.5%; 0.734 / 0.046	68.8%; 0.741 / 0.071	69.8%; 0.733 / 0.053
		w/ Korean	79.5%; 0.755 / 0.050	81.5%; 0.753 / 0.056	81.0%; 0.778 / 0.053	78.8%; 0.770 / 0.055	84.0%; 0.778 / 0.047
		w/ Chinese	72.8%; 0.627 / 0.068	80.7%; 0.740 / 0.062	81.0%; 0.710 / 0.065	79.3%; 0.721 / 0.074	81.7%; 0.743 / 0.069
	Distractor Selected	Overall	35.0%; 0.068 / 0.636	32.0%; 0.025 / 0.719	31.1%; 0.016 / 0.733	30.8%; 0.015 / 0.728	28.5%; 0.015 / 0.728
		w/ English	55.8%; 0.048 / 0.741	56.2%; 0.015 / 0.795	52.8%; 0.003 / 0.818	50.3%; 0.000 / 0.808	49.5%; 0.003 / 0.803
		w/ Spanish	36.7%; 0.068 / 0.622	33.8%; 0.044 / 0.699	33.5%; 0.035 / 0.714	31.2%; 0.027 / 0.712	30.2%; 0.006 / 0.737
		w/ Korean	20.5%; 0.041 / 0.586	18.5%; 0.009 / 0.655	19.0%; 0.026 / 0.613	21.2%; 0.016 / 0.673	16.0%; 0.021 / 0.655
		w/ Chinese	27.2%; 0.129 / 0.478	19.3%; 0.034 / 0.598	19.0%; 0.009 / 0.652	20.7%; 0.032 / 0.613	18.3%; 0.055 / 0.572

Table 7: SNR-wise confidence-based stream selection in the Separation setting. Each SNR cell reports Selection Ratio; Accuracy / ECE, where Selection Ratio indicates the proportion of items for which the given stream was chosen.

Model	Condition	Distractor	SNR				
			-10 dB	-5 dB	0 dB	+5 dB	+10 dB
Qwen2-Audio	Single	Overall	0.875 / - / 0.125				
	Separation	Overall	0.192 / 0.768 / 0.041	0.166 / 0.804 / 0.030	0.147 / 0.823 / 0.030	0.168 / 0.794 / 0.038	0.162 / 0.808 / 0.030
		w/ English	0.182 / 0.799 / 0.018	0.128 / 0.862 / 0.010	0.100 / 0.894 / 0.006	0.132 / 0.865 / 0.004	0.114 / 0.879 / 0.007
		w/ Spanish	0.144 / 0.832 / 0.024	0.147 / 0.837 / 0.015	0.134 / 0.850 / 0.016	0.155 / 0.826 / 0.019	0.127 / 0.850 / 0.023
		w/ Korean	0.359 / 0.519 / 0.122	0.370 / 0.513 / 0.116	0.398 / 0.472 / 0.131	0.384 / 0.459 / 0.157	0.432 / 0.455 / 0.114
		w/ Chinese	0.159 / 0.805 / 0.036	0.104 / 0.881 / 0.015	0.070 / 0.918 / 0.012	0.086 / 0.895 / 0.019	0.094 / 0.894 / 0.012
	Cocktail Party	Overall	0.061 / 0.894 / 0.045	0.083 / 0.856 / 0.061	0.214 / 0.716 / 0.070	0.594 / 0.302 / 0.104	0.736 / 0.141 / 0.122
		w/ English	0.023 / 0.962 / 0.016	0.048 / 0.929 / 0.024	0.278 / 0.678 / 0.044	0.603 / 0.348 / 0.050	0.748 / 0.153 / 0.099
		w/ Spanish	0.043 / 0.931 / 0.025	0.055 / 0.900 / 0.046	0.184 / 0.725 / 0.091	0.650 / 0.243 / 0.107	0.742 / 0.113 / 0.144
		w/ Korean	0.147 / 0.723 / 0.130	0.220 / 0.613 / 0.168	0.505 / 0.347 / 0.147	0.741 / 0.103 / 0.155	0.757 / 0.126 / 0.117
		w/ Chinese	0.052 / 0.918 / 0.030	0.044 / 0.919 / 0.037	0.081 / 0.880 / 0.038	0.455 / 0.434 / 0.111	0.698 / 0.170 / 0.132
MERaLiON-2	Single	Overall	0.822 / - / 0.178				
	Separation	Overall	0.287 / 0.657 / 0.057	0.235 / 0.713 / 0.052	0.262 / 0.696 / 0.042	0.279 / 0.682 / 0.039	0.260 / 0.705 / 0.035
		w/ English	0.197 / 0.769 / 0.034	0.168 / 0.824 / 0.008	0.204 / 0.788 / 0.008	0.169 / 0.815 / 0.016	0.199 / 0.781 / 0.020
		w/ Spanish	0.230 / 0.711 / 0.059	0.187 / 0.766 / 0.048	0.228 / 0.731 / 0.041	0.232 / 0.732 / 0.035	0.202 / 0.758 / 0.040
		w/ Korean	0.389 / 0.562 / 0.049	0.327 / 0.562 / 0.111	0.360 / 0.561 / 0.079	0.427 / 0.513 / 0.060	0.414 / 0.520 / 0.066
		w/ Chinese	0.376 / 0.532 / 0.092	0.312 / 0.624 / 0.064	0.304 / 0.634 / 0.062	0.366 / 0.578 / 0.056	0.282 / 0.698 / 0.020
	Cocktail Party	Overall	0.076 / 0.858 / 0.066	0.108 / 0.816 / 0.075	0.284 / 0.587 / 0.129	0.602 / 0.262 / 0.136	0.650 / 0.243 / 0.106
		w/ English	0.039 / 0.935 / 0.026	0.036 / 0.939 / 0.025	0.197 / 0.743 / 0.060	0.581 / 0.323 / 0.097	0.659 / 0.228 / 0.114
		w/ Spanish	0.046 / 0.909 / 0.044	0.084 / 0.852 / 0.064	0.268 / 0.583 / 0.149	0.576 / 0.288 / 0.136	0.674 / 0.232 / 0.095
		w/ Korean	0.118 / 0.771 / 0.111	0.187 / 0.694 / 0.119	0.357 / 0.469 / 0.174	0.667 / 0.179 / 0.154	0.658 / 0.234 / 0.108
		w/ Chinese	0.112 / 0.793 / 0.095	0.149 / 0.743 / 0.108	0.353 / 0.488 / 0.158	0.588 / 0.252 / 0.160	0.612 / 0.282 / 0.107
Audio-Flamingo-3	Single	Overall	0.982 / - / 0.018				
	Separation	Overall	0.125 / 0.837 / 0.037	0.101 / 0.886 / 0.013	0.093 / 0.891 / 0.016	0.108 / 0.876 / 0.016	0.100 / 0.882 / 0.018
		w/ English	0.104 / 0.884 / 0.012	0.062 / 0.925 / 0.012	0.067 / 0.920 / 0.013	0.075 / 0.918 / 0.007	0.075 / 0.914 / 0.011
		w/ Spanish	0.157 / 0.807 / 0.036	0.100 / 0.900 / 0.000	0.096 / 0.893 / 0.011	0.107 / 0.893 / 0.000	0.092 / 0.897 / 0.011
		w/ Korean	0.202 / 0.746 / 0.053	0.235 / 0.722 / 0.043	0.183 / 0.779 / 0.038	0.267 / 0.678 / 0.056	0.232 / 0.717 / 0.051
		w/ Chinese	0.065 / 0.882 / 0.054	0.045 / 0.949 / 0.006	0.054 / 0.940 / 0.007	0.046 / 0.935 / 0.020	0.052 / 0.935 / 0.013
	Cocktail Party	Overall	0.028 / 0.937 / 0.035	0.034 / 0.938 / 0.027	0.134 / 0.807 / 0.060	0.860 / 0.098 / 0.042	0.888 / 0.035 / 0.077
		w/ English	0.024 / 0.961 / 0.015	0.027 / 0.966 / 0.008	0.102 / 0.879 / 0.019	0.800 / 0.167 / 0.033	0.949 / 0.000 / 0.051
		w/ Spanish	0.023 / 0.954 / 0.023	0.021 / 0.963 / 0.015	0.095 / 0.857 / 0.048	0.875 / 0.100 / 0.025	0.765 / 0.118 / 0.118
		w/ Korean	0.061 / 0.853 / 0.086	0.082 / 0.847 / 0.070	0.324 / 0.527 / 0.149	0.870 / 0.043 / 0.087	0.878 / 0.024 / 0.098
		w/ Chinese	0.008 / 0.970 / 0.023	0.016 / 0.961 / 0.023	0.108 / 0.827 / 0.065	0.889 / 0.111 / 0.000	0.966 / 0.000 / 0.034
Qwen2.5-Omni	Single	Overall	0.924 / - / 0.076				
	Separation	Overall	0.318 / 0.610 / 0.071	0.321 / 0.626 / 0.053	0.315 / 0.634 / 0.051	0.310 / 0.633 / 0.058	0.347 / 0.598 / 0.056
		w/ English	0.185 / 0.797 / 0.018	0.148 / 0.830 / 0.023	0.143 / 0.834 / 0.023	0.167 / 0.817 / 0.016	0.171 / 0.804 / 0.024
		w/ Spanish	0.278 / 0.622 / 0.100	0.302 / 0.644 / 0.054	0.283 / 0.653 / 0.063	0.311 / 0.616 / 0.073	0.346 / 0.596 / 0.058
		w/ Korean	0.409 / 0.502 / 0.089	0.472 / 0.433 / 0.095	0.447 / 0.470 / 0.082	0.372 / 0.534 / 0.094	0.461 / 0.432 / 0.107
		w/ Chinese	0.466 / 0.436 / 0.098	0.483 / 0.454 / 0.063	0.458 / 0.449 / 0.051	0.458 / 0.474 / 0.067	0.526 / 0.417 / 0.057
	Cocktail Party	Overall	0.084 / 0.843 / 0.073	0.111 / 0.803 / 0.086	0.312 / 0.589 / 0.099	0.721 / 0.189 / 0.090	0.824 / 0.091 / 0.085
		w/ English	0.024 / 0.968 / 0.009	0.050 / 0.941 / 0.009	0.153 / 0.798 / 0.049	0.656 / 0.281 / 0.063	0.813 / 0.123 / 0.065
		w/ Spanish	0.053 / 0.907 / 0.039	0.076 / 0.874 / 0.050	0.312 / 0.605 / 0.083	0.680 / 0.195 / 0.125	0.847 / 0.098 / 0.055
		w/ Korean	0.185 / 0.625 / 0.189	0.254 / 0.493 / 0.252	0.621 / 0.169 / 0.210	0.779 / 0.093 / 0.128	0.811 / 0.016 / 0.173
		w/ Chinese	0.088 / 0.843 / 0.069	0.095 / 0.836 / 0.069	0.281 / 0.623 / 0.096	0.790 / 0.161 / 0.048	0.823 / 0.114 / 0.063
GPT-4o mini Audio	Single	Overall	0.759 / - / 0.241				
	Separation	Overall	0.258 / 0.612 / 0.130	0.265 / 0.607 / 0.128	0.284 / 0.595 / 0.121	0.280 / 0.602 / 0.118	0.257 / 0.642 / 0.101
		w/ English	0.268 / 0.632 / 0.099	0.280 / 0.640 / 0.080	0.287 / 0.631 / 0.082	0.286 / 0.623 / 0.092	0.277 / 0.649 / 0.074
		w/ Spanish	0.226 / 0.611 / 0.163	0.229 / 0.595 / 0.176	0.244 / 0.614 / 0.142	0.237 / 0.595 / 0.168	0.251 / 0.622 / 0.127
		w/ Korean	0.278 / 0.597 / 0.125	0.286 / 0.607 / 0.107	0.335 / 0.539 / 0.127	0.332 / 0.574 / 0.094	0.252 / 0.654 / 0.094
		w/ Chinese	0.261 / 0.606 / 0.133	0.265 / 0.583 / 0.152	0.270 / 0.593 / 0.137	0.273 / 0.611 / 0.116	0.246 / 0.642 / 0.112
	Cocktail Party	Overall	0.158 / 0.767 / 0.075	0.227 / 0.669 / 0.104	0.326 / 0.552 / 0.121	0.432 / 0.432 / 0.136	0.538 / 0.333 / 0.130
		w/ English	0.142 / 0.816 / 0.042	0.199 / 0.739 / 0.062	0.250 / 0.645 / 0.105	0.374 / 0.527 / 0.099	0.437 / 0.479 / 0.084
		w/ Spanish	0.153 / 0.749 / 0.098	0.218 / 0.648 / 0.134	0.250 / 0.615 / 0.135	0.401 / 0.451 / 0.148	0.530 / 0.308 / 0.162
		w/ Korean	0.161 / 0.745 / 0.093	0.264 / 0.639 / 0.097	0.440 / 0.440 / 0.120	0.496 / 0.359 / 0.145	0.681 / 0.181 / 0.138
		w/ Chinese	0.177 / 0.756 / 0.067	0.233 / 0.644 / 0.122	0.411 / 0.463 / 0.126	0.509 / 0.311 / 0.179	0.580 / 0.260 / 0.160
Gemini-2.0-Flash	Single	Overall	0.963 / - / 0.037				
	Separation	Overall	0.273 / 0.724 / 0.003	0.460 / 0.540 / 0.000	0.452 / 0.530 / 0.017	0.536 / 0.445 / 0.018	0.594 / 0.396 / 0.010
		w/ English	0.226 / 0.774 / 0.000	0.129 / 0.871 / 0.000	0.262 / 0.738 / 0.000	0.226 / 0.774 / 0.000	0.286 / 0.714 / 0.000
		w/ Spanish	0.354 / 0.631 / 0.015	0.955 / 0.045 / 0.000	0.750 / 0.250 / 0.000	0.846 / 0.077 / 0.077	0.952 / 0.000 / 0.048
		w/ Korean	0.385 / 0.615 / 0.000	0.739 / 0.261 / 0.000	0.562 / 0.438 / 0.000	0.833 / 0.167 / 0.000	0.833 / 0.167 / 0.000
		w/ Chinese	0.236 / 0.764 / 0.000	0.647 / 0.353 / 0.000	0.786 / 0.071 / 0.143	0.789 / 0.211 / 0.000	0.929 / 0.071 / 0.000
	Cocktail Party	Overall	0.007 / 0.983 / 0.010	0.015 / 0.979 / 0.006	0.040 / 0.949 / 0.011	0.101 / 0.884 / 0.016	0.899 / 0.101 / 0.000
		w/ English	0.009 / 0.988 / 0.004	0.028 / 0.963 / 0.010	0.134 / 0.830 / 0.036	0.741 / 0.259 / 0.000	1.000 / 0.000 / 0.000
		w/ Spanish	0.005 / 0.986 / 0.008	0.009 / 0.988 / 0.003	0.016 / 0.981 / 0.003	0.049 / 0.944 / 0.007	0.760 / 0.240 / 0.000
		w/ Korean	0.007 / 0.985 / 0.008	0.005 / 0.990 / 0.005	0.020 / 0.973 / 0.007	0.084 / 0.899 / 0.018	0.875 / 0.125 / 0.000
		w/ Chinese	0.007 / 0.974 / 0.019	0.019 / 0.973 / 0.007	0.052 / 0.934 / 0.014	0.183 / 0.774 / 0.043	1.000 / 0.000 / 0.000

Table 8: SNR-wise error-type distribution (Mis / Int / Ung) across Single, Separation, and Cocktail Party settings. Single has no distractor stream and is therefore shown once rather than by SNR.

Model	Domain	Single	Cocktail Party (SNR)				
			-10 dB	-5 dB	0 dB	+5 dB	+10 dB
Qwen2-Audio	Aviation	0.687 / 0.197	0.123 / 0.715	0.160 / 0.656	0.445 / 0.371	0.685 / 0.166	0.737 / 0.150
	Healthcare	0.847 / 0.090	0.143 / 0.730	0.147 / 0.707	0.483 / 0.375	0.825 / 0.093	0.890 / 0.061
	Finance	0.833 / 0.093	0.123 / 0.737	0.162 / 0.687	0.502 / 0.345	0.805 / 0.095	0.870 / 0.062
	Construction	0.727 / 0.166	0.092 / 0.749	0.138 / 0.689	0.435 / 0.376	0.708 / 0.161	0.795 / 0.120
	Average	0.773 / 0.132	0.120 / 0.731	0.152 / 0.685	0.466 / 0.366	0.756 / 0.124	0.823 / 0.093
MERaLiON-2	Aviation	0.727 / 0.191	0.280 / 0.513	0.377 / 0.423	0.628 / 0.241	0.792 / 0.119	0.823 / 0.098
	Healthcare	0.760 / 0.188	0.210 / 0.620	0.287 / 0.553	0.555 / 0.327	0.787 / 0.147	0.833 / 0.121
	Finance	0.813 / 0.143	0.215 / 0.615	0.302 / 0.520	0.665 / 0.215	0.858 / 0.076	0.840 / 0.102
	Construction	0.727 / 0.217	0.203 / 0.635	0.267 / 0.576	0.557 / 0.313	0.755 / 0.168	0.783 / 0.149
	Average	0.757 / 0.176	0.227 / 0.596	0.308 / 0.515	0.601 / 0.274	0.798 / 0.127	0.820 / 0.116
Audio-Flamingo-3	Aviation	0.867 / 0.038	0.137 / 0.635	0.172 / 0.571	0.600 / 0.092	0.913 / 0.032	0.925 / 0.023
	Healthcare	0.953 / 0.068	0.145 / 0.639	0.180 / 0.589	0.543 / 0.195	0.968 / 0.044	0.978 / 0.049
	Finance	0.947 / 0.043	0.175 / 0.590	0.193 / 0.552	0.622 / 0.121	0.968 / 0.036	0.967 / 0.031
	Construction	0.867 / 0.024	0.097 / 0.652	0.152 / 0.595	0.555 / 0.165	0.912 / 0.021	0.892 / 0.027
	Average	0.908 / 0.030	0.138 / 0.626	0.174 / 0.576	0.580 / 0.141	0.940 / 0.026	0.940 / 0.026
Qwen2.5-Omni	Aviation	0.600 / 0.123	0.062 / 0.544	0.103 / 0.467	0.325 / 0.213	0.653 / 0.064	0.738 / 0.055
	Healthcare	0.647 / 0.099	0.053 / 0.599	0.122 / 0.485	0.387 / 0.212	0.750 / 0.078	0.798 / 0.071
	Finance	0.727 / 0.048	0.105 / 0.499	0.115 / 0.459	0.327 / 0.233	0.662 / 0.027	0.743 / 0.046
	Construction	0.627 / 0.087	0.065 / 0.578	0.095 / 0.505	0.365 / 0.217	0.637 / 0.098	0.715 / 0.083
	Average	0.650 / 0.061	0.071 / 0.555	0.109 / 0.479	0.351 / 0.216	0.675 / 0.042	0.749 / 0.051
GPT-4o mini Audio	Aviation	0.680 / -	0.415 / -	0.527 / -	0.633 / -	0.708 / -	0.765 / -
	Healthcare	0.800 / -	0.397 / -	0.498 / -	0.675 / -	0.785 / -	0.812 / -
	Finance	0.860 / -	0.430 / -	0.527 / -	0.687 / -	0.792 / -	0.853 / -
	Construction	0.747 / -	0.365 / -	0.423 / -	0.550 / -	0.712 / -	0.773 / -
	Average	0.772 / -	0.402 / -	0.494 / -	0.636 / -	0.749 / -	0.801 / -
Gemini-2.0-Flash	Aviation	0.940 / -	0.027 / -	0.088 / -	0.292 / -	0.715 / -	0.977 / -
	Healthcare	0.960 / -	0.035 / -	0.058 / -	0.233 / -	0.608 / -	0.987 / -
	Finance	0.960 / -	0.045 / -	0.065 / -	0.238 / -	0.623 / -	0.980 / -
	Construction	0.960 / -	0.023 / -	0.050 / -	0.205 / -	0.578 / -	0.942 / -
	Average	0.955 / -	0.033 / -	0.065 / -	0.242 / -	0.631 / -	0.971 / -

Table 9: Domain-wise single and Cocktail Party performance across target-to-distractor SNRs. Each cell reports Accuracy / ECE. Single is SNR-independent and is shown once per domain.

Early Warning for Target-Speaker Extraction: Reliability under Distribution and Failure-Definition Shift

Shubham Sharma and Manik Gupta

Department of Computer Science and Information Systems
Birla Institute of Technology and Science Pilani, Hyderabad Campus
Hyderabad, India
{p20240435,manik}@hyderabad.bits-pilani.ac.in

Abstract

Target-speaker extraction (TSE) systems are typically evaluated only after the complete output has been produced, limiting the ability to identify potentially unreliable outputs during generation. We investigate whether failure risk can instead be estimated causally from only the first three seconds of the emitted TSE waveform.

We develop an early-warning monitor using 11,946 outputs spanning five conversational corpora and six TSE systems. Because independent failure labels are unavailable for this discovery cohort, we define a proxy adverse outcome from seven complete-output indicators of degraded or atypical TSE behaviour. After excluding monitoring features that overlap with the proxy construction, the monitor predicts this eventual outcome from the early waveform with an AUROC of 0.8062 and an AUPRC of 0.6606. At a 15% review budget, it identifies 54.47% of proxy-adverse outputs with 66.29% precision.

To test whether this early-warning signal transfers beyond the discovery setting, we freeze the monitor and evaluate it unchanged on 2,000 unseen outputs with independently computed SI-SDRi. When failure is defined as $\text{SI-SDRi} \leq 0$ dB, performance falls to near chance (AUROC 0.5354; AUPRC 0.0700). Yet, on the same external outputs, the frozen predictions remain strongly associated with the original proxy outcome (AUROC 0.8092; AUPRC 0.8481). The proxy and SI-SDRi failure definitions agree only weakly ($\kappa = 0.028$, $\text{MCC} = 0.064$), while substantial feature-distribution shift is observed between the discovery and external cohorts (45 of 63 features have $\text{PSI} \geq 0.25$).

These results show that short-prefix monitoring can provide useful early risk prioritization within the setting in which the monitor is developed, but that such signals do not automatically transfer to independently defined failure outcomes. Reliable deployment therefore

requires explicit attention to both distribution shift and the definition of failure.

Keywords: target-speaker extraction; reliability monitoring; early failure detection; distribution shift; out-of-domain evaluation; selective prediction.

1 Introduction

Target-speaker extraction (TSE) aims to recover the speech of a designated speaker from a multi-speaker mixture using an enrollment utterance or speaker representation as the target cue (Delcroix et al., 2018; Zhang et al., 2025). Although modern TSE systems perform well on controlled benchmarks, assessing their reliability is more difficult in real conversational audio, where overlapping speech, reverberation, noise, intermittent target activity, and enrollment–mixture mismatch are common. Under such conditions, a TSE system may suppress the intended speaker, retain an interfering speaker, or distort the temporal structure of the extracted speech.

Existing evaluation protocols largely assess TSE outputs after extraction is complete, using aggregate measures of intelligibility, speaker similarity, timing, and perceptual quality. These measures are valuable for comparing systems, but they do not address an important runtime question: can we identify an output that is likely to become unreliable while it is still being produced? An early causal monitor could provide such a warning and support abstention, escalation, or downstream review without modifying the extraction waveform itself.

Developing such a monitor raises two main challenges. The first is how to define an unreliable output. Independent failure labels are not available for all outputs in our discovery cohort. We therefore construct a proxy adverse label from seven complete-output indicators of degraded or

atypical TSE behaviour: silence, clipping, energy mismatch, spectral flatness, temporal instability, output–mixture inconsistency, and cross-system disagreement. Because some candidate monitoring features may overlap with the indicators used to construct this proxy, prediction could be artificially inflated by target circularity. We therefore explicitly test whether useful early prediction remains after removing these overlapping features.

The second challenge is whether a monitor developed in one setting remains reliable in another. Its performance may change across conversational corpora or TSE architectures, particularly when the underlying data distribution changes. More fundamentally, performance may also depend on what is considered a failure. We therefore examine the monitoring approach across multiple corpora and TSE systems and subsequently test a frozen monitor on unseen outputs using an independently computed reconstruction-based definition of failure.

Our central question is therefore: *can the first three seconds of a TSE output provide an early warning that the completed output will be labelled as adverse?* We investigate this using a single-stream causal monitor that observes only the waveform emitted by the evaluated TSE system. The monitor has no access to future audio, does not generate or compare alternative outputs, and does not modify or replace the emitted waveform. It simply assigns a risk score from the acoustic evidence available in the early output. We then examine how well this risk ranking generalizes across corpora, TSE systems, and failure definitions.

We make four main contributions. First, we formulate TSE reliability monitoring as a single-stream causal risk-prediction problem that uses only the first three seconds of one emitted waveform. Second, we explicitly audit target circularity and show that a predefined reduced non-overlap feature set retains substantial predictive signal under trial-disjoint evaluation, achieving an AUROC of 0.8062 and an AUPRC of 0.6606. Third, we evaluate the practical utility of the resulting risk ranking under limited review budgets and examine how consistently the same policy performs across five conversational corpora and six TSE systems.

Fourth, we freeze the trained monitor and apply it unchanged to 2,000 unseen outputs from four TSE systems. These outputs provide an independently computed reconstruction measure, SI-SDRi, which indicates whether TSE improves

the target signal relative to the original mixture. When failure is defined as $\text{SI-SDRi} \leq 0$ dB, the frozen monitor performs close to chance (AUROC 0.5354; AUPRC 0.0700). Yet, on these same external outputs, its predictions remain strong against the original proxy-defined outcome (AUROC 0.8092; AUPRC 0.8481). Together with substantial differences between the discovery and external feature distributions and weak agreement between the two failure labels, these findings show that external reliability depends not only on how the data change, but also on how failure is defined.

Overall, our results support causal early warning as a useful mechanism for prioritizing potentially unreliable TSE outputs within the setting in which the monitor is developed. At the same time, they show that strong within-setting performance should not be interpreted as evidence of universal failure detection: reliable transport to new data therefore requires explicit attention to both distribution shift and independently grounded definitions of failure.

2 Related Work

Target-speaker extraction. Target-speaker extraction conditions a separation model on an enrollment utterance or speaker representation to recover a designated speaker from a mixture. SpeakerBeam established enrollment-conditioned extraction for single-channel speech and was later extended to target-speaker recognition (Delcroix et al., 2018, 2019). Subsequent work strengthened speaker conditioning through improved embedding architectures and multi-level speaker representations (Desplanques et al., 2020; Wang et al., 2023; Zhang et al., 2025). Other methods explicitly address speaker-confusion errors arising from ambiguous target-speaker representations (Zhao et al., 2022; Liu et al., 2023), while band-split recurrent modelling provides an effective time–frequency reconstruction architecture for source separation (Luo and Yu, 2023). These studies primarily improve or repair the extraction process itself. In contrast, we investigate whether an output produced by a fixed TSE system can be identified as high risk from early causal evidence, without generating alternatives or modifying the emitted waveform.

Confidence estimation and selective prediction.

Selective prediction allows a model to abstain or defer when its estimated risk is high, inducing

an explicit trade-off between prediction coverage and reliability (Geifman and El-Yaniv, 2019). In speech recognition, confidence-estimation methods identify potentially incorrect words or utterances and support downstream review, filtering, and data selection (Oneață et al., 2021). Confidence estimation can also degrade under acoustic or linguistic domain mismatch, motivating domain-aware evaluation and calibration (Li et al., 2022). Our setting differs from token- or word-level ASR confidence estimation because the monitored object is a continuously emitted extraction waveform and the decision is made from a fixed three-second causal prefix.

Responsible evaluation and cross-domain reliability. Responsible speech-system evaluation requires more than reporting average performance. Subgroup analyses must account for confounding factors, speaker heterogeneity, and incomplete demographic evidence to avoid unsupported fairness conclusions (Liu et al., 2022). We therefore audit intervention precision, failure coverage, false-intervention burden, target circularity, metadata coverage, and cross-domain consistency. Because the available demographic metadata are selective, we do not claim population-level fairness; instead, we report where the evidence supports reliability conclusions and where it remains insufficient.

A further distinction is required between within-cohort cross-validation and frozen external transport. A monitor can rank examples successfully within its development distribution while failing after changes in acoustic conditions, extraction architecture, or outcome definition. We therefore separate trial-disjoint internal evaluation from a later frozen-model study using independently computed system-specific reconstruction labels, and analyse the feature-distribution shift associated with any loss of transportability.

3 Causal Monitoring Framework

3.1 Causal Monitoring Setup

We use the beginning of a TSE output to estimate the risk that the completed output will be labelled adverse. For each trial, let x denote the mixture waveform, e the target-speaker enrollment utterance, and y the waveform emitted by a TSE system. The reduced monitor observes only an early prefix $y_{1:t}$ and estimates

$$r_t = f_\theta(\phi(y_{1:t})), \quad (1)$$

where $\phi(\cdot)$ extracts acoustic evidence available from the emitted waveform up to time t , and $r_t \in [0, 1]$ is the estimated failure risk. The enrollment utterance is used by the upstream TSE system to produce y , but is not an input to the reduced 63-feature monitor used in the principal analyses.

The causal monitor operates only on the waveform emitted by the evaluated TSE system. Here, “causal” means that it uses only audio available up to time t ; it does not imply causal inference. The monitor neither accesses future samples nor generates, selects, or replaces output waveforms. Its prediction is intended only for warning, abstention, escalation, or downstream review.

As illustrated in Figure 1, the emitted waveform is analysed using one-second causal windows updated every 0.25 seconds. These updates provide progressively more acoustic evidence and are not separate failure decisions. Our main evaluation uses the first 3 seconds of the output. This $t = 3$ s operating point was fixed before evaluation and is not claimed to be optimal. We additionally evaluate prefixes of 1, 2, 3, and 4 s to examine how early useful warning information appears and whether the results depend strongly on the 3-s choice.

Because independent failure labels are unavailable for all discovery outputs, we use the archived binary variable `proxy_adverse_outcome`. It combines seven complete-output indicators: silence, clipping, energy mismatch, spectral flatness, temporal instability, output-mixture inconsistency, and cross-system disagreement. The indicators are equally weighted after component scaling estimated on the predefined 7,812-output development split. An output is labelled proxy-adverse when the resulting score exceeds the fixed development-set 80th-percentile threshold of 0.7830. Neither the combination rule nor the threshold is selected using predictor performance. The proxy therefore represents potentially degraded or atypical TSE behaviour rather than an independently verified physical or perceptual failure label.

Some monitoring features may overlap with the indicators used to construct the proxy and thereby inflate predictive performance. To assess this target circularity, we compare three feature configurations. The *full acoustic set* contains all eligible causal variables; the *target-overlap-only set* contains variables directly overlapping with proxy construction; and the *reduced non-overlap set* excludes the target column, compos-

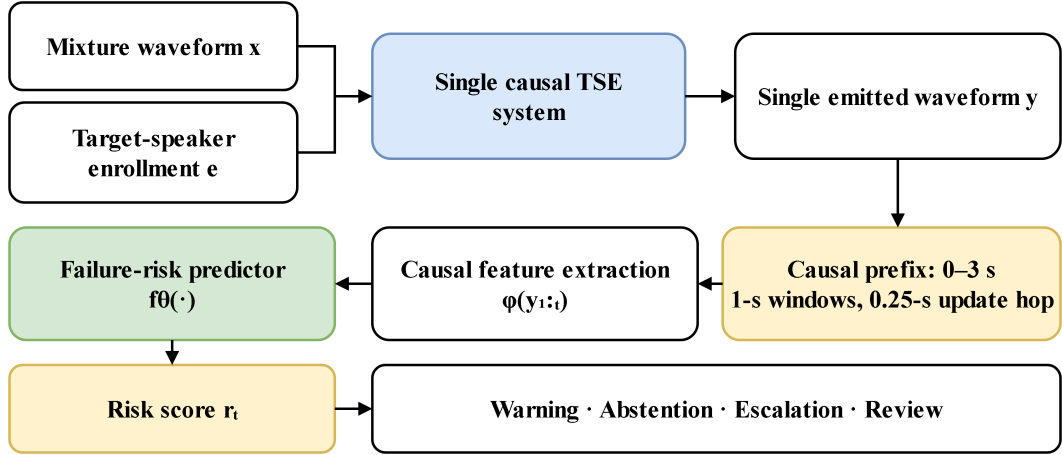


Figure 1: Single-stream causal monitoring framework. A fixed TSE system uses the mixture waveform and target-speaker enrollment utterance to produce one emitted waveform. The reduced monitor then uses only acoustic evidence available within the first three seconds of that waveform to estimate failure risk. It does not access future audio, generate alternative waveforms, select complete utterances, or replace the emitted output.

ite risk score, seven component scores, mixture-derived variables, and all other identified target-overlapping causal variables.

The reduced non-overlap set contains 63 features describing emitted-waveform amplitude, spectral-band ratios, spectral shape, zero-crossing behaviour, and temporal summaries over the causal prefix. It is “reduced” because target-overlapping variables are excluded, not because features were selected using evaluation-label performance. This configuration is used for all operating-policy analyses.

The predictor f_θ is a class-balanced L2-regularized logistic regression model with $C = 1.0$, fitted using the liblinear solver after median imputation, missing-value-indicator augmentation, and standard scaling. The same preprocessing and classifier are used across all three feature sets; only the input representation changes. Logistic regression provides a simple, interpretable basis for comparing the feature configurations.

3.2 Evaluation Protocol

We evaluate 11,946 outputs from 1,991 trials, where each trial is one mixture–target–enrollment instance evaluated by six TSE systems. The data span five conversational corpora: AISHELL-4, AMI, AliMeeting, CHiME6, and DipCo. The systems are USEF_SepFormer_WSJ0_2mix, USEF_TFGridNet_WSJ0_2mix, spk_emb_100, spk_emb_causal_100, tfmap_context_100, and tfmap_context_causal_100. Each trial contributes one output per system, yielding

$1,991 \times 6 = 11,946$ outputs.

Because the six outputs from a trial share the same mixture, target, and enrollment, they are kept in the same fold. We use five-fold trial-disjoint cross-validation repeated ten times and average the ten out-of-fold predictions for each output to obtain its final risk score. Let $z_i \in \{0, 1\}$ denote the proxy-defined discovery outcome.

Although cross-system disagreement contributes to offline target construction, the deployed monitor observes only the emitted waveform. The target, composite score, component scores, and all identified target-overlapping causal variables are excluded from the reduced non-overlap representation. We refer to outputs with $z_i = 1$ as proxy-defined failures; this does not imply independently verified perceptual or task failure. No independent measure of intelligibility, speaker preservation, timing, reconstruction quality, perceptual quality, or human listening quality is available for all 11,946 discovery outputs.

For frozen external evaluation, we use 500 trials and four TSE systems, yielding 2,000 system-output instances. Each has an independently computed system-specific scale-invariant signal-to-distortion ratio improvement (SI-SDRi), which measures whether extraction improves the target signal relative to the original mixture. External failure is defined as $\text{SI-SDRi} \leq 0$ dB, yielding 129 failures: 27 for spk_emb_100, 79 for spk_emb_causal_100, 9 for tfmap_context_100, and 14 for tfmap_context_causal_100. The saved re-

duced monitor is applied without refitting, feature reselection, calibration, or threshold adjustment. We also evaluate the same predictions against the original proxy-defined outcome on these 2,000 outputs, helping distinguish the effects of distribution shift and failure definition.

An intervention budget b denotes the fraction of outputs that can be reviewed or escalated. Outputs are ranked by predicted risk, and the highest-risk fraction b is selected. Internal analyses use averaged trial-disjoint out-of-fold scores, whereas external analysis uses scores from the unchanged frozen monitor. Let S_b denote the selected set and $\ell_i \in \{0, 1\}$ the relevant cohort-specific failure label: proxy-defined internally and SI-SDRi-defined externally. Intervention precision and failure coverage are

$$\begin{aligned} \text{Prec}(b) &= \frac{\sum_{i \in S_b} \ell_i}{|S_b|}, \\ \text{Cov}(b) &= \frac{\sum_{i \in S_b} \ell_i}{\sum_i \ell_i}. \end{aligned} \quad (2)$$

Precision is the proportion of selected outputs that are failures, while failure coverage is the proportion of all failures captured within the budget. This budgeted-selection perspective is related to selective classification and reject-option methods (El-Yaniv and Wiener, 2010). Selective classification has also been studied for deep neural networks (Geifman and El-Yaniv, 2017).

We define coverage lift as $\text{Lift}(b) = \text{Cov}(b)/b$, since random selection has expected failure coverage b . Thus, a lift of $3.63\times$ at a 15% budget indicates 3.63 times the failure coverage expected under random selection. We define the false-intervention rate (FIR), equivalent to the false-positive rate when intervention is treated as the positive decision, as

$$\text{FIR}(b) = \frac{\sum_{i \in S_b} (1 - \ell_i)}{\sum_i (1 - \ell_i)}. \quad (3)$$

The FIR is therefore the proportion of negative outputs selected for intervention.

We report representative budgets of 5%, 15%, and 30%; the cross-corpus and cross-system analysis uses the predefined set $\{15\%, 20\%, 30\%\}$. All ranking rules and budgets are fixed before subgroup comparisons and applied unchanged across corpora and TSE systems.

4 Results

4.1 Early Proxy-Adverse-Outcome Prediction

We first ask whether the first three seconds of the emitted TSE waveform contain sufficient information to predict whether the completed output will be labelled proxy-adverse. All predictions are out-of-fold and trial-disjoint, so outputs derived from the same underlying trial do not appear in both training and evaluation partitions. Of the 11,946 outputs, 2,181 are labelled proxy-adverse, corresponding to a prevalence of 18.26%. Because the expected AUPRC of a random ranking equals the positive-class prevalence, the random AUPRC baseline is 0.1826.

To assess the effect of feature–target overlap, all three feature configurations use the same class-balanced logistic-regression classifier, preprocessing pipeline, proxy target, repeated trial-disjoint folds, and evaluation procedure; only the input feature set changes. As shown in Table 1, the full acoustic feature set achieves an AUROC of 0.8991 and an AUPRC of 0.7613. Restricting the input to features identified as overlapping with proxy construction yields an AUROC of 0.8851 and an AUPRC of 0.7316, indicating that these features alone contain substantial information about the automatically constructed outcome. After removing all identified overlapping features, the reduced non-overlap model achieves an AUROC of 0.8062 and an AUPRC of 0.6606, compared with a random AUPRC baseline of 0.1826. Thus, substantial predictive signal remains after removal of the identified feature–target overlap.

Bootstrap comparisons show that the full model exceeds the reduced model by 0.0929 AUROC (95% CI: [0.0810, 0.1053]) and 0.1007 AUPRC (95% CI: [0.0862, 0.1163]). Both confidence intervals exclude zero, showing a clear contribution from the identified target-overlapping features while confirming that the reduced model retains useful ranking ability after their removal.

As a post-hoc sensitivity analysis, we separately evaluate full-prefix models using 1, 2, 3, and 4 seconds of observed audio. Their holdout AUROCs are 0.828, 0.832, 0.873, and 0.863, respectively, with corresponding AUPRCs of 0.358, 0.434, 0.498, and 0.495. These results indicate that predictive information is present before three seconds and that extending the observed prefix beyond three seconds does not improve performance

Feature set	AUROC	95% CI	AUPRC	95% CI
Full acoustic	0.8991	[0.8900, 0.9078]	0.7613	[0.7421, 0.7796]
Target-overlap only	0.8851	[0.8750, 0.8945]	0.7316	[0.7108, 0.7506]
Reduced non-overlap	0.8062	[0.7914, 0.8215]	0.6606	[0.6359, 0.6848]

Table 1: Trial-disjoint prediction performance at the predefined three-second causal decision point. Confidence intervals are obtained using 3,000 trial-cluster bootstrap repetitions. The reduced non-overlap model is used for all subsequent operating-policy analyses.

Budget	Precision	Coverage	Lift	FIR
5%	93.14%	25.54%	5.10×	0.42%
15%	66.29%	54.47%	3.63×	6.19%
30%	43.83%	72.03%	2.40×	20.61%

Table 2: Representative budget-constrained operating points for the reduced non-overlap model. FIR denotes the false-intervention rate.

in this sensitivity analysis. These models constitute a separate sensitivity experiment and should not be directly compared with the principal trial-disjoint reduced-model estimates in Table 1.

4.2 Budget-Constrained Utility

We next evaluate whether the reduced non-overlap model can prioritize proxy-adverse outputs when only a limited fraction of outputs can be reviewed or escalated. Outputs are ranked by their out-of-fold risk scores, and the highest-risk fraction is selected for intervention. The budget denotes the fraction of outputs that can be reviewed or escalated; for example, a 15% budget corresponds to selecting the highest-risk 15% of outputs. These budgets represent illustrative capacity constraints rather than deployment recommendations.

Table 2 reports three representative operating points. At a 5% budget, the monitor achieves 93.14% precision and captures 25.54% of proxy-defined failures, corresponding to a $5.10\times$ lift over random selection and a false-intervention rate of 0.42%. At a 15% budget, precision is 66.29% and failure coverage is 54.47%, corresponding to a $3.63\times$ lift. Increasing the budget to 30% raises failure coverage to 72.03%, while precision decreases to 43.83%, lift decreases to $2.40\times$, and the false-intervention rate increases to 20.61%.

4.3 Within-Cohort Cross-Corpus and Cross-System Analysis

We next apply the same predefined 15%, 20%, and 30% review policies across the five discovery corpora and six TSE systems, without corpus-

or system-specific retraining or threshold adjustment. The monitor captures more proxy-defined failures than expected under random selection in every corpus and system. At the 15% budget, the lowest corpus-level lift is $1.71\times$ on CHiME6, while the lowest system-level lift is $2.59\times$ for `spk_emb_100`.

Performance nevertheless varies across corpora and systems. At the 15% budget, AISHELL-4 achieves 64.71% precision and 45.83% failure coverage, whereas AliMeeting achieves 15.70% failure coverage. At the 20% budget, DipCo achieves 19.62% coverage, and at the 30% budget, AliMeeting precision is 17.07%. The comparatively large lift observed for AISHELL-4 should be interpreted in light of its low proxy-adverse prevalence (1.67%), and therefore jointly with precision and coverage. Overall, the same risk-ranking policy performs above random selection across all discovery corpora and systems, although its utility varies substantially across these groups.

4.4 Frozen Multi-System External Validation

We next apply the frozen reduced monitor to 2,000 unseen outputs from 500 trials and four TSE systems. External failure is independently defined as $\text{SI-SDR}_i \leq 0$ dB, yielding 129 failures and a prevalence of 6.45%. No refitting, feature reselection, calibration, threshold search, or hyperparameter adjustment is performed.

Against the independently defined SI-SDR_i outcome, the frozen monitor achieves a pooled AUROC of 0.5354 (95% CI: [0.4776, 0.5914]) and an AUPRC of 0.0700 (95% CI: [0.0552, 0.0977]), compared with a random AUPRC baseline of 0.0645. At a 15% review budget, the monitor selects 300 outputs and identifies 19 of the 129 SI-SDR_i-defined failures, corresponding to 6.33% precision, 14.73% failure coverage, and a $0.982\times$ lift over random selection. Per-system AUROC ranges from 0.5172 to 0.5494; because the system-specific failure counts are small, we treat these estimates as descriptive and base the external con-

Evaluation	Outcome	Outputs	Failures	AUROC	AUPRC	Random AUPRC
Internal OOF	Proxy-defined	11,946	2,181	0.8062 [0.7914, 0.8215]	0.6606 [0.6359, 0.6848]	0.1826
Frozen external	SI-SDRi ≤ 0 dB	2,000	129	0.5354 [0.4776, 0.5914]	0.0700 [0.0552, 0.0977]	0.0645

Table 3: Internal trial-disjoint and frozen external evaluation of the reduced causal monitor. The evaluations use different outcome definitions and therefore address different validity questions; their performance values should not be interpreted as a controlled comparison of the same prediction task.

clusion primarily on the pooled analysis. Overall, the frozen monitor does not reliably discriminate SI-SDRi-defined failures in this external cohort.

To examine whether this result reflects only a change in the input distribution, we also evaluate the unchanged predictions against the original proxy-defined outcome on the same 2,000 external outputs. Under this outcome definition, the monitor achieves an AUROC of 0.8092 and an AUPRC of 0.8481. Agreement between the proxy-defined and SI-SDRi-defined failure labels is weak ($\kappa = 0.028$, $MCC = 0.064$). Thus, the near-chance performance against SI-SDRi cannot be attributed to distribution shift alone; differences in the definition of failure are also associated with the observed loss of discrimination.

4.5 Distribution-Shift and Metadata Audits

A post-evaluation audit reveals substantial differences between the reduced monitoring features in the discovery and external cohorts. Of the 63 features, 24 have an absolute standardized mean difference of at least 0.50, 45 have a population stability index of at least 0.25, and 40 have a Kolmogorov–Smirnov statistic of at least 0.20. The median predicted risk increases from 0.3092 in the discovery cohort to 0.7448 in the external cohort. No individual feature achieves an oriented external AUROC of 0.65 or greater; the highest is 0.6070. These findings confirm substantial feature-distribution shift between the two cohorts. Together with the external proxy analysis above, they indicate that the external SI-SDRi result occurs in the presence of both feature shift and a change in outcome definition; the present analysis does not isolate their individual contributions.

We also assess the extent to which the available metadata support demographic analyses. Authentic language or gender metadata are available for only 355 of the 1,991 discovery trials, corresponding to 17.83% trial coverage. This subset is selective and cannot be assumed to represent the full discovery cohort. The available evidence therefore

does not support population-level claims about demographic fairness or disparity.

Finally, the discovery and external evaluations address different notions of adverse outcome. The discovery analysis uses the proxy-defined label because independent measures of intelligibility, target-speaker preservation, timing, reconstruction quality, perceptual quality, or human listening quality are unavailable for all 11,946 outputs. The external cohort provides independently computed SI-SDRi for 2,000 outputs from four systems, but SI-SDRi represents a single reconstruction-based definition of failure. The two evaluations should therefore be interpreted as complementary tests of within-setting prediction and transport to an independently defined outcome, rather than as evaluations of an identical prediction task.

5 Discussion

The results separate two questions that can otherwise be conflated. First, early causal acoustic evidence contains useful predictive information within the proxy-defined discovery setting. After removing features identified as overlapping with proxy construction, the reduced model retains an AUROC of 0.8062 and an AUPRC of 0.6606. The monitor also provides useful budget-constrained ranking within this cohort, including a $3.63\times$ lift over random selection at the representative 15% review budget.

Second, performance within the proxy-defined discovery setting does not establish reliability for independently defined failure outcomes on unseen data. Although the same risk-ranking policy performs above random selection across all five discovery corpora and six discovery systems, the frozen monitor performs near chance against SI-SDRi-defined failure on the external four-system cohort (AUROC 0.5354; AUPRC 0.0700). The external cohort exhibits substantial feature-distribution shift. However, on the same external outputs, the unchanged monitor remains strongly predictive of the original proxy-defined outcome

(AUROC 0.8092; AUPRC 0.8481), while agreement between the proxy and SI-SDRi failure labels is weak. Together, these findings indicate that the near-chance SI-SDRi performance cannot be attributed to distribution shift alone; the change in failure definition is also associated with the loss of discrimination. The present analysis does not disentangle the individual contributions of these two factors.

Calibration may alter score interpretation and operating thresholds, but it cannot recover ranking information when discrimination against the target outcome is near chance. Reliable transport may therefore require more robust representations, independently grounded training targets, or adaptation to the intended deployment setting. These possibilities remain to be evaluated.

Overall, causal early warning provides a useful mechanism for prioritizing proxy-adverse outputs within the setting in which the monitor is developed, but reliable monitoring across new systems, data distributions, and independently defined failure outcomes remains unresolved. The monitor should therefore be treated as an auxiliary prioritization and review tool rather than as a deployment-ready failure detector.

6 Limitations

First, the discovery outcome is defined automatically using seven equal-weight proxy diagnostics rather than independent perceptual or task-level labels. This proxy is not assumed to be an optimal definition of failure, and alternative weightings or outcome definitions may produce different results.

Second, the external evaluation is limited to 500 trials from four systems and uses SI-SDRi as a single reconstruction-oriented failure criterion. SI-SDRi does not capture other important aspects of TSE performance, such as intelligibility, target-speaker preservation, timing, perceptual quality, or human preference. Moreover, because the internal and external evaluations use different failure definitions, their performance values are not directly comparable. The external cohort also exhibits substantial feature shift. Although evaluation against the original proxy suggests that both data shift and the change in failure definition are relevant, the present study does not disentangle their individual effects. We also do not establish whether more complex nonlinear models would improve external transport.

Third, authentic demographic metadata are available for only 355 of the 1,991 discovery trials. The resulting analysis is therefore a limited demographic audit and cannot establish population-level fairness or disparity.

Finally, the study evaluates early warning and review prioritization, not the effectiveness of subsequent interventions. It does not establish that acting on an alert prevents failure, improves the emitted waveform or standard TSE metrics, or leads to better outcomes for human listeners.

7 Ethics and Responsible Use

A failure-risk monitor may reduce the need to inspect every output, but it also introduces operational trade-offs. False alerts can lead to unnecessary review or abstention, whereas missed alerts may allow unreliable outputs to pass without intervention. The operating point should therefore be selected according to the relative costs of these errors and the review capacity available in the intended deployment setting.

The monitor should not be interpreted as a fairness certificate, a substitute for human evaluation, or evidence of safety in high-consequence settings. Its use should be supported by independent outcome labels, representative data, and domain-specific cost assumptions. In particular, the frozen external results show that performance can degrade substantially under changes in the data distribution and failure definition, and that recalibration alone should not be assumed to restore reliable performance. Deployment should therefore be preceded by independently labelled, prospective validation on the intended systems, populations, and recording conditions.

8 Conclusion

We presented a single-stream causal early-warning framework that uses only the first three seconds of a TSE system’s emitted waveform. Under trial-disjoint evaluation against the proxy-defined discovery outcome, the reduced non-overlap monitor achieved an AUROC of 0.8062 and an AUPRC of 0.6606. At a 15% intervention budget, it identified 54.47% of proxy-defined failures with 66.29% precision, corresponding to a $3.63\times$ lift over random prioritization.

The unchanged monitor did not reliably discriminate independently defined SI-SDRi failures on 2,000 external outputs from four systems. Under

this evaluation, AUROC was 0.5354 and AUPRC was 0.0700, compared with a random AUPRC baseline of 0.0645. The external cohort also exhibited substantial feature-distribution shift. Importantly, on the same external outputs, the unchanged monitor retained strong discrimination for the original proxy-defined outcome (AUROC 0.8092; AUPRC 0.8481). Together, these findings indicate that the near-chance SI-SDRi performance cannot be attributed to distribution shift alone; the change in failure definition is also associated with the loss of discrimination. The present analysis does not disentangle the individual contributions of these factors.

Overall, our findings support causal early warning as a useful within-cohort prioritization mechanism, while highlighting the limitations of treating such a monitor as a universal, deployment-ready failure detector. Future work should focus on learning shift-robust causal representations and on prospective validation against independently defined outcomes that capture perceptual quality, speaker preservation, timing, and intelligibility.

Generative AI Assistance

Generative language tools assisted language editing and organization. The authors verified all experiments, analyses, references, and final text.

References

- Marc Delcroix, Shinji Watanabe, Tsubasa Ochiai, Keisuke Kinoshita, Shigeki Karita, Atsunori Ogawa, and Tomohiro Nakatani. 2019. [End-to-end Speaker-Beam for single channel target speech recognition](#). In *Proceedings of Interspeech*, pages 451–455.
- Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. 2018. [Single channel target speaker extraction and recognition with SpeakerBeam](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5554–5558.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuyne. 2020. [ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification](#). In *Proceedings of Interspeech*, pages 3830–3834.
- Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.
- Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159.
- Qiujia Li, Yu Zhang, David Qiu, Yanzhang He, Lianliang Cao, and Philip C. Woodland. 2022. [Improving confidence estimation on out-of-domain data for end-to-end speech recognition](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6537–6541.
- Kai Liu, Ziqing Du, Xucheng Wan, and Huan Zhou. 2023. [X-SepFormer: End-to-end speaker extraction network with explicit optimization on speaker confusion](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Zhe Liu, Irina-Elena Veliche, and Fuchun Peng. 2022. [Model-based approach for measuring the fairness in ASR](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6532–6536.
- Yi Luo and Jianwei Yu. 2023. [Music source separation with band-split RNN](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901.
- Dan Oneață, Alexandru Caranica, Adriana Stan, and Horia Cucu. 2021. [An evaluation of word-level confidence estimation for end-to-end automatic speech recognition](#). In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 258–265.
- Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023. [WeSpeaker: A research and production oriented speaker embedding learning toolkit](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Ke Zhang, Junjie Li, Shuai Wang, Yangjie Wei, Yi Wang, Yannan Wang, and Haizhou Li. 2025. [Multi-level speaker representation for target speaker extraction](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Zifeng Zhao, Dongchao Yang, Rongzhi Gu, Haoran Zhang, and Yuexian Zou. 2022. [Target confusion in end-to-end speaker extraction: Analysis and approaches](#). In *Proceedings of Interspeech*, pages 5333–5337.

Evaluating Bias in Phoneme-Based Automatic Speech Recognition Systems: An Analysis of IPA Transcription Models

Maneesha Rani Saha *
University of Utah
maneesha.saha@utah.edu

Catherine Bao *
University of Utah
u1459030@utah.edu

Neal Patwari
University of Utah
neal.patwari@utah.edu

Abstract

As automatic speech recognition (ASR) systems shift toward multilingual support and low-resource language modeling, phoneme-based layers serve as a critical language-agnostic foundation. However, most evaluations of ASR’s demographic biases related to race, age, gender, and accent focus on standard grapheme-based ASR systems with comparatively little emphasis on phoneme-based systems. In this study, we evaluate the performance of WhisperIPA and ZIPA, two state-of-the-art open-source systems that generate International Phonetic Alphabet (IPA) transcriptions. Our evaluation includes existing multilingual speech corpora and demographically annotated English-language corpora, comparing model-generated IPA transcriptions against grapheme-to-phoneme (G2P) systems using both standard phoneme error rate (PER) and a proposed Soft PER metric that tolerates linguistically similar phoneme substitutions. Our analysis examines how performance varies across language, gender, accent, ethnicity, and age, revealing persistent disparities even after accounting for acceptable phonemic variation. These findings, while limited, provide insight into potential sources of bias and inform the development of more inclusive and linguistically robust phoneme-based ASR systems. Our code and data are publicly available.¹

1 Introduction

Automatic Speech Recognition (ASR) systems are a foundational component of modern agent-based technology. Current ASR systems typically rely on grapheme representations: graphemes are written units of language (e.g., letters or characters), while phonemes are units of sound that represent how words are pronounced. Existing research shows

that grapheme-based ASR systems exhibit systematic disparities across demographic groups, including race, gender, age, accent, and language background (Tatman and Kasten, 2017; Tatman, 2017; DiChristofano et al., 2023; Jahan et al., 2025; El-Ghazaly et al., 2025; Scott et al., 2025; Serditova et al., 2025; Cunningham et al., 2025). ASR evaluation currently mostly focuses on word error rates (WER) (Feng et al., 2021).

While a majority of industry models remain grapheme-based, there has been a growing trend toward integrating IPA layers for multilingual capability (Li et al., 2020; Xu et al., 2021). The International Phonetic Alphabet (IPA) captures fine-grained phonetic details and cross-linguistic sound variations. Multilingual training on IPA tokens reduces phonetic token error rate by 60–70% in high-resource languages and 14–42% in low-resource languages (Želasko et al., 2022), and including IPA layers enables more effective generalization by providing a more direct mapping between the acoustic signal and its phonological categories. This makes recent work in phonetic (Bharadwaj et al., 2026; Goriely and Buttery, 2025) and IPA models (Fang et al., 2020; Taguchi et al., 2023; Zhu et al., 2024; Lee et al., 2025; Zhu et al., 2025), integrated as standalone or intermediate layers, valuable for multilingual and low-resource ASR systems.

Despite this growing importance, bias research has focused almost exclusively on grapheme-based, WER-measured systems; comparatively little has been done on phoneme representations. This gap matters in both directions: if disparities are already present at the phoneme level, they may propagate through downstream components of integrated ASR systems; conversely, if phonetic representations prove more robust, they may lead to more equitable speech recognition. Phonemic-level demographic bias thus remains largely underexplored, motivating the present study.

Since the same phoneme can be realized differ-

*These authors contributed equally.

¹<https://github.com/Maneesha28/ipa-asr-transcription-bias-evaluation>

ently across accents, dialects, and languages, strict Phonetic Error Rate (PER) may penalize valid pronunciation variation, introducing noise when evaluating demographic disparities. To reduce this effect, we introduce a soft variant of PER that tolerates minor phonemic variation. We evaluate how this PER variant changes phoneme ASR bias analysis. Our contributions include:

1. Introducing and evaluating a Soft PER that tolerates minor phonemic variation rather than strictly penalizing all substitutions. We compare Soft PER with standard PER to determine whether it yields a more robust metric for evaluating demographic bias.
2. Evaluating the biases of IPA-based phoneme ASR models across accent, gender, age, ethnicity and language. Our evaluation spans 11 languages (English, Hindi, Bangla, Panjabi, Tamil, Telugu, Arabic, French, Spanish, Turkish, and Shona). We also analyze different age groups, gender, four ethnicity groups, and English accents across eight accent/dialect groups.

2 Related Work

End-to-end ASR systems have increased accuracy by scaling data and model size, especially with transformer-family architectures (Radford et al., 2023). Large models such as wav2vec 2.0 (Baeovski et al., 2020), Whisper (Radford et al., 2023), and XLS-r (Babu et al., 2022) demonstrate that grapheme-based models can generalize to dozens of languages. Recent work focuses on even larger multilingual models that replace language-specific systems and target lower-resource languages (Meng et al., 2025; Omnilingual ASR team et al., 2025; Ramirez et al., 2024).

Despite these advances, ASR is concentrated in high-resource languages with standardized orthographic transcriptions (Gale et al., 2023; Gorisch and Schmidt, 2024). Grapheme-based models rely on widely adopted writing systems in paired speech-text data. For languages and conversation without a formal writing system, non-standardized orthography, limited textual resources, or code-switching, grapheme models are less reliable or unavailable. Most grapheme mappings do not consider phonetic variation (Fang et al., 2020; Omnilingual ASR team et al., 2025; Ramirez et al., 2024).

By mapping acoustically similar sounds to shared phonetic symbols, these systems can transfer knowledge from high-resource to truly under-resourced languages (Daul et al., 2026). Currently, architectures introduce an explicit phonetic layer that improves robustness and data efficiency in low-resource settings (Lee et al., 2025). Incorporating phonemic supervision can reduce the ultimate WER (Fu et al., 2025; Scott et al., 2025).

Phonemic models are particularly advantageous for orthographic limitations or linguistic variability. Phonemic/phonetic token models outperform orthographic baselines in both data-constrained and cross-lingual domains (Daul et al., 2026). Shared phonetics allow easier cross-lingual transfer and generalize underrepresented languages (Goriely and Buttery, 2025; Zhu et al., 2025; Devi and Sarma, 2026). Phonemic modeling directly captures acoustic distinctions, thus improving robustness to pronunciation variability (Lyu and Lyu, 2008; Solorio and Liu, 2008; Onda et al., 2026).

Rather than replacing grapheme-based systems, phonetic layers are often integrated within them. Multi-task learning frameworks jointly predict phonemes and graphemes. Hybrid approaches have demonstrated improved robustness in noisy, low-resource, and multilingual settings, suggesting that phonetic supervision can complement larger ASR architectures (Fang et al., 2020).

Bias in ASR systems has been extensively studied in grapheme-based systems, with prior research showing significant disparities in WER across demographic groups — such as race, gender, dialect, and accent (Tatman and Kasten, 2017; DiChristofano et al., 2023; Jahan et al., 2025). An ASR system exhibits bias when its performance differs by user demographics (Zhang et al., 2023). Within grapheme-based models, commercial systems consistently have higher error rates for women compared to men (Ngueajio and Washington, 2022; Aryal et al., 2023). Models often have higher error rates for children (Bhardwaj et al., 2022). Non-native and non-urban accents have been associated with higher error rates (Serditova et al., 2025; Torgbi et al., 2025; Wassink et al., 2022), with second-language speech accents showing higher error rates than speakers of non-standard or regional accents (Feng et al., 2021; Zhang et al., 2022). For example, commercial models produce twice as many errors for African American speakers as for white speakers of English (Koenecke et al., 2020; Martin and Tang, 2020).

Standardized evaluation benchmarks and bias audits have been introduced to quantify these performance gaps and assess fairness in widely used ASR models. In contrast, investigations into phonetic or IPA-based ASR models remain limited. As phoneme-level systems become more common, either as full speech transcription models or as intermediate components within larger architectures, it is increasingly important to understand how they may contribute to or mitigate existing biases.

3 Methods

3.1 Datasets

We evaluate phoneme recognition models across two complementary dimensions: cross-lingual robustness and demographic variation.

Cross-lingual Datasets. To assess multilingual phoneme recognition performance, we use IPA-PACK (FLEURS-IPA) (Zhu et al., 2024), MediaSpeech (Kolobov et al., 2021) and WAXAL (Diack et al., 2026). From these multilingual corpora, we select 11 languages for evaluation, including high-resource languages (English, Spanish, and French) and lower-resource languages (Hindi, Bangla, Panjabi, Tamil, Telugu, Arabic, Turkish, and Shona). Languages are selected based on support from both the G2P pipeline and the evaluated IPA models. Table 1 in Appendix A.1 provides more details about the language distributions of these datasets.

Demographic Datasets. To analyze demographic bias and variation in phoneme recognition, we use English speech datasets with speaker metadata: the Corpus of Regional African American Language (CORAAL) (Kendall and Farrington, 2018), the Edinburgh International Accents of English Corpus (EdAcc) (Sanabria et al., 2023), the Sonos Voice Control Bias Assessment Dataset (SVC) (Sekkat et al., 2024), and WAXAL (Diack et al., 2026). These datasets contain annotations for demographic attributes, including age, gender, ethnicity, accent, geographic region, and native-speaker status, with the demography presented in Table 2 in Appendix A.1.

3.2 Models

In this work, we evaluate two IPA-based ASR models which allows us to maximize coverage in cross-lingual evaluation, since they have the largest number of overlapping supported languages.

WhisperIPA. We evaluate WhisperIPA², a transformer-based encoder-decoder model built on the Whisper architecture (Radford et al., 2023). Whisper is originally pre-trained on 680,000 hours of weakly supervised multilingual speech data. The base WhisperIPA variant, with approximately 74M parameters, is further fine-tuned for direct speech-to-IPA transcription on 15,000 labeled synthetic IPA audio samples. The fine-tuning data is derived from Common Voice 21, which covers >70 languages with corresponding phonetic annotations.

ZIPA. We evaluate ZIPA (Zhu et al., 2025), a family of efficient multilingual phone recognition models built on the Zipformer architecture (Yao et al., 2024), that includes both transducer-based and CTC-based variants. In this work, we use the large consistency-regularized CTC model with noisy-student training. The ZIPA-CR-NS³ large variant has approximately 300M parameters and is trained on IPAPack++, a large-scale multilingual speech corpus with 17,132 hours of normalized phone transcriptions across 88 languages. The noisy-student training stage further incorporates approximately 11,000 hours of pseudo-labeled multilingual speech from >4,000 languages, improving cross-linguistic phone recognition performance.

3.3 Experimental Setup

All models are evaluated in a zero-shot setting without additional fine-tuning on the evaluation datasets in order to measure out-of-the-box phoneme recognition performance. After both models generate IPA transcriptions for every audio sample, outputs are post-processed to remove special tokens, normalize whitespace, and ensure consistency with the reference phoneme format. Output post-processing details are described in Appendix A.2.

Reference phoneme sequences are generated from orthographic transcripts using a multilingual grapheme-to-phoneme (G2P) pipeline. We use G2P+ (Goriely and Buttery, 2025) to generate IPA-based phonemic transcriptions across languages. Although generated phonemic references introduce limitations due to potentially incorrect, oversimplified phoneme mappings, incomplete phonetic detail, and inconsistencies across languages, they provide a scalable framework for multilingual evaluation. These generated phoneme sequences serve

²<https://huggingface.co/neurlang/ipa-whisper-base>

³<https://huggingface.co/anyspeech/zipa-large-crctc-ns-800k>

as the ground-truth references for all experiments.

Soft PER We evaluate phoneme recognition accuracy using phoneme error rate (PER). However, because phoneme realizations can vary across accents, dialects, and languages without reflecting true recognition failures, we also introduce and evaluate a Soft PER metric that reduces penalties for linguistically similar substitutions. This evaluation uses a two-tier phoneme mapping to (1) establish broad interchangeable phonemic variation and (2) create language-specific pronunciation overlaps that are not global equivalents. A single equivalence tier is insufficient because many phonetic relationships are not identical between languages.

The first tier groups phonemes into transitive equivalence classes using two allophonic resources. AlloVera (Mortensen et al., 2020) provides language-specific mappings between surface phones and canonical phonemes across 14 languages (e.g., aspirated /p^h/→/p/ where a strongly breathed “p” sound is treated as a standard /p/, devoiced /b̥/→/b/ where a partially unvoiced “b” is mapped to its canonical form, etc.), with English given priority, because English defines the largest proportion of the demographic evaluation data, when a surface phone appears in multiple languages. We then use articulatory feature representations from PHOIBLE (Moran and McCloy, 2019) to identify pairs of canonical phonemes that differ in only one of 34 articulatory features. These pairs are merged into shared equivalence classes. Through this grouping, phonemes may belong to the same class even if they differ in more than one feature, as long as they are connected through phonemes that differ by only one feature at each step. Together, these mappings create 62 equivalence classes spanning 254 phones. Table 4 in Appendix A.3 provides the full mapping for the first tier.

The second tier captures non-transitive cross-language similarities, derived from AlloVera (Mortensen et al., 2020). When the same surface phone maps to different canonical phonemes across languages (e.g., the flap sound /ɾ/ is mapped to /t/ in English words such as “water,” but to /r/ in Spanish words such as “pero”), those canonical phonemes are recorded as a direct pair without being merged into a shared equivalence class. This preserves the individual relations /t/≈/ɾ/ and /r/≈/ɾ/ without implying /t/≈/r/ and keeping language-specific relationships. A total of 90 such language-specific

pairs are extracted from cross-language allophonic overlaps. These pairs are scored as equivalent within their respective language mappings, but are not transitively combined. Table 5 in Appendix A.3 provides the full mapping for the second tier.

For example, Tier 1 places devoiced /b̥/ and canonical /b/ in the same equivalence class, so substituting one for the other incurs no error. In contrast, /d/ and /t/ belong to different Tier 1 classes and would initially be counted as a substitution error. Tier 2 contains this pair as a language-specific similarity in English, Kazakh, Russian, and Vietnamese, so the substitution isn’t counted when it happens in one of these languages. Unlike Tier 1 equivalences, this relation is not extended transitively to other phonemes.

During evaluation, Tier 1 equivalence classes are applied universally across all languages, while Tier 2 similarities are only applied when relevant to the target language. Substitutions within either similarity tier receive a zero penalty when computing Soft PER, allowing the metric to better distinguish genuine recognition errors from acceptable phonetic variation.

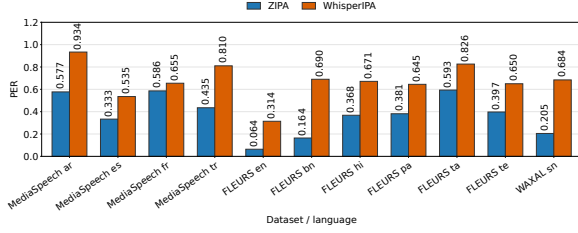
4 Results

4.1 Language Performance Evaluation

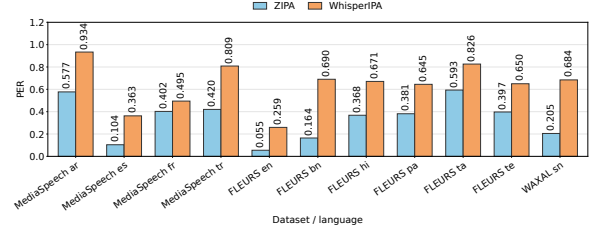
We first evaluate model performance across languages to understand how IPA-based ASR system accuracy varies in a multilingual setting. This analysis compares standard PER and Soft PER for both ZIPA and WhisperIPA across the available language datasets. Standard PER measures exact phoneme-level mismatch, while Soft PER applies the equivalence mappings described in Section 3.3 to reduce penalties for linguistically similar phoneme substitutions.

Figure 1 compares ZIPA and WhisperIPA across the evaluated languages using both standard PER and Soft PER. ZIPA achieves lower error rates than WhisperIPA for every language under both metrics, indicating stronger overall phoneme-recognition performance in this multilingual evaluation. Under standard PER, English has the lowest error rate for both models. For ZIPA, the highest error rates occur for Tamil, French, and Arabic, whereas WhisperIPA produces its highest error rates for Arabic, Tamil, and Turkish. These results demonstrate substantial variation in performance across languages.

Under Soft PER, ZIPA continues to outperform WhisperIPA for every language, although the ef-

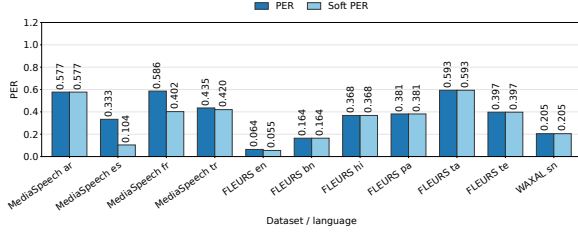


(a) Standard PER across languages for ZIPA and WhisperIPA.

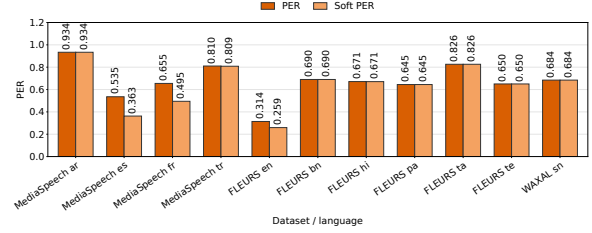


(b) Soft PER across languages for ZIPA and WhisperIPA.

Figure 1: Comparison of standard PER and Soft PER across evaluated languages: (ar=Arabic, es=Spanish, fr=French, tr=Turkish, en=English, bn=Bangla, hi=Hindi, pa=Panjabi, ta=Tamil, te=Telugu, and sn=Shona).



(a) Standard PER and Soft PER comparison for ZIPA.



(b) Standard PER and Soft PER comparison for WhisperIPA.

Figure 2: Effect of Soft PER on language-level evaluation for each model.

fect of the soft-equivalence mapping varies considerably across languages. The largest reductions occur for Spanish and French. For ZIPA, Spanish decreases from 0.333 to 0.104 and French decreases from 0.586 to 0.402. WhisperIPA shows similar reductions, with Spanish decreasing from 0.535 to 0.363 and French decreasing from 0.655 to 0.495. English also improves under Soft PER, particularly for WhisperIPA, while Turkish shows only a small reduction for ZIPA and a negligible change for WhisperIPA. In contrast, the values for Arabic, Bangla, Hindi, Panjabi, Tamil, Telugu, and Shona remain unchanged. Thus, Soft PER affects only a subset of the evaluated languages rather than reducing errors uniformly across the multilingual evaluation.

Figure 2 presents the effect of Soft PER separately for each model. For ZIPA, the reduction is largest for Spanish and French, smaller for Turkish and English, and absent for the remaining languages. WhisperIPA follows a similar pattern, with substantial reductions for Spanish and French, a smaller reduction for English, and almost no change for Turkish. The other languages remain unchanged. Despite these language-specific reductions, WhisperIPA continues to produce higher error rates than ZIPA across all languages.

Key Takeaways: Overall, ZIPA consistently outperforms WhisperIPA across all evaluated lan-

guages. ZIPA achieves an average standard PER of 0.373, compared with 0.674 for WhisperIPA, corresponding to an absolute difference of 0.301. Under Soft PER, the averages decrease to 0.333 and 0.639, respectively, while the absolute difference remains nearly unchanged at 0.305. Thus, Soft PER lowers the aggregate error rates but does not narrow the performance gap between the models. This suggests that ZIPA’s advantage is not explained primarily by the phonemically similar substitutions captured by the current Soft PER mapping.

Considering English, Spanish, and French as high-resource languages and Hindi, Bangla, Panjabi, Tamil, Telugu, Arabic, Turkish, and Shona as low-resource languages, average error rates are higher for the low-resource group. Averaged across both models, standard PER is 0.565 for low-resource languages and 0.415 for high-resource languages, representing a 36.2% relative increase. Under Soft PER, the corresponding averages are 0.563 and 0.280, producing an absolute difference of 0.284 and a relative difference of approximately 101%.

The widening relative gap under Soft PER is driven primarily by the substantial reductions observed for Spanish and French, both of which belong to the high-resource group. English also improves, although by a smaller amount. In comparison, most low-resource languages remain unchanged, with only a small reduction for Turk-

ish. One possible explanation is that the AlloVera-derived equivalence mappings, which prioritize English when mappings overlap across languages, do not adequately represent phonemic variation in many of the low-resource languages. Differences in model architecture, training objectives, and multilingual training data may also contribute to the observed pattern. Overall, Soft PER identifies phonemically similar substitutions in a limited subset of languages but does not change the model ranking or reduce the average performance disparity between the high- and low-resource groups.

4.2 Demographic Performance Evaluation

We next analyze how phoneme recognition performance varies across demographic groups using CORAAL, EdAcc, SVC, and WAXAL, focusing on gender, age, ethnicity, and accent/dialect. We interpret demographic patterns within each dataset rather than across datasets, since dataset characteristics may contribute more to performance variation than the demographic attributes themselves. We report bootstrap 95% confidence intervals for all group means shown as error bars in the figures and test every pairwise group contrast with a Welch’s *t*-test. Full pairwise results are reported by demographic attribute in Tables 6–11 in Appendix A.4.

Gender: Figure 3 presents PER and Soft PER across gender groups in SVC, EdAcc, WAXAL, and CORAAL datasets. In SVC, EdAcc, and CORAAL datasets, male and female speakers exhibit similar error rates, with no consistent direction of disparity across models or metrics. WAXAL shows the only significant gender contrast: PER and Soft PER for ZIPA are significantly higher for female speakers (+0.067, $p < 0.01$; Table 6). In contrast, WhisperIPA shows a larger but not significant descriptive gap in the opposite direction (−0.149), likely reflecting the small male sample ($n = 52$). Soft PER lowers absolute error rates while preserving within-dataset patterns (identical to PER for WAXAL, since none of its substitutions fall under the soft-equivalence mapping). Overall, aside from this single ZIPA effect in WAXAL, the results show limited evidence of systematic gender-based disparity.

Age: Figure 4 presents PER and Soft PER across age groups in SVC, EdAcc, and CORAAL datasets. In SVC, both models show small absolute differences across age groups (Figure 4a), though several are statistically significant: the 42–54 group has

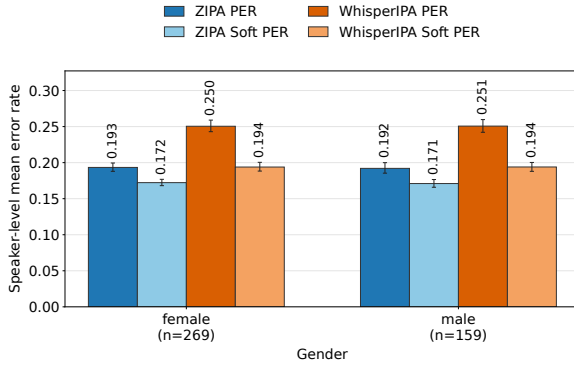
significantly lower PER than the 17–28 group (all four metrics, $p < 0.05$, $p < 0.01$ for ZIPA) and the 9–16 group (ZIPA PER/Soft PER and WhisperIPA Soft PER, $p < 0.05$; Table 8), with no other SVC pair reaching significance. In EdAcc, ZIPA is stable across ages, while WhisperIPA peaks for the 45–59 group (Figure 4b); this should be read cautiously, as the group has only five speakers and none of EdAcc’s three pairwise age contrasts reach significance (Table 9). CORAAL⁴ shows the clearest pattern: the 0–19 group has significantly higher PER than the 20–29, 30–50, and 51+ groups on all four metrics ($p < 0.01$, mostly $p < 0.001$) and higher ZIPA PER/Soft PER than the 0–29 group ($p < 0.01$), though the WhisperIPA contrast with 0–29 is not significant ($p = 0.11$, $p = 0.13$; Table 10). No other pair differs significantly. Soft PER lowers absolute error rates while largely preserving these patterns. Overall, age effects are dataset-dependent: a robust effect for the youngest speakers in CORAAL, a modest but significant effect in SVC’s 42–54 group, and no significant effect in EdAcc despite a large descriptive gap.

Ethnicity: Figure 5 presents PER and Soft PER across ethnicity groups in the EdAcc dataset. “White”, “Asian”, “Black”, and “South Asian” are possible self-selected ethnicity descriptors in EdAcc.⁵ As shown in Figure 5a, WhisperIPA produces substantially higher error rates than ZIPA across all groups, consistent with the model-level differences observed throughout the demographic evaluation.

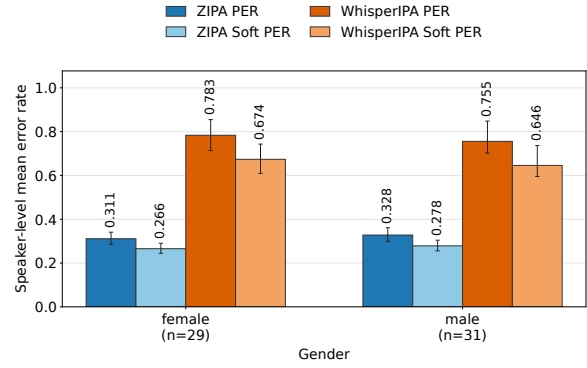
Pairwise Welch’s *t*-tests (Table 7) show “Black” speakers have significantly higher PER than “South Asian” speakers for ZIPA PER, WhisperIPA PER (+0.157, $p < 0.01$), and WhisperIPA Soft PER (+0.144, $p < 0.05$), and higher PER than “White” speakers for ZIPA PER and Soft PER (+0.074, +0.054; $p < 0.05$). No other contrast is significant: “Asian” speakers show an elevated disparity from the overall mean (Figure 5b) but are not significantly different from any individual group, including “Black” speakers. All estimates rest on small samples ($n = 6–23$) and should be interpreted accordingly.

⁴CORAAL component corpora use nonuniform age bins; although the speaker groups are disjoint, the source label 0–29 overlaps in range with 0–19 and 20–29 and may partly capture subcorpus differences. See the [CORAAL User Guide](#).

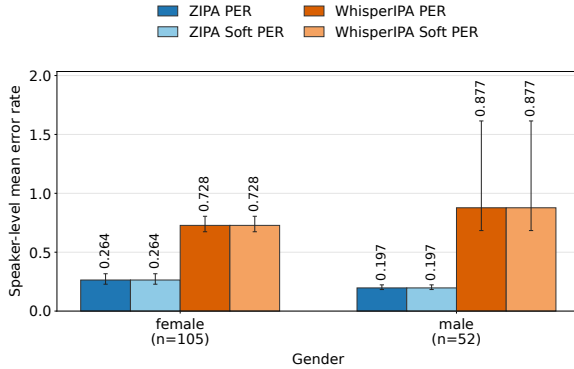
⁵https://groups.inf.ed.ac.uk/EdAcc/EdAcc_statement.pdf



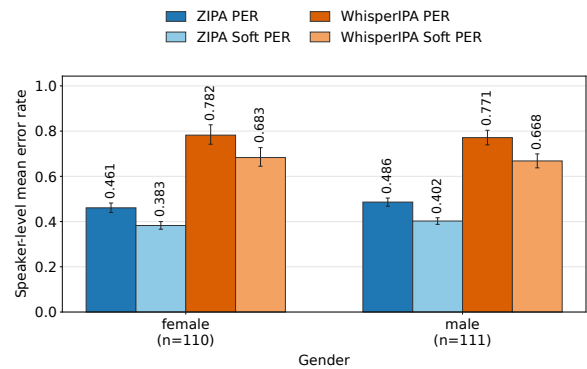
(a) PER and Soft PER by gender for SVC Dataset



(b) PER and Soft PER by gender for EdAcc Dataset



(c) PER and Soft PER by gender for WAXAL Dataset



(d) PER and Soft PER by gender for CORAAL Dataset

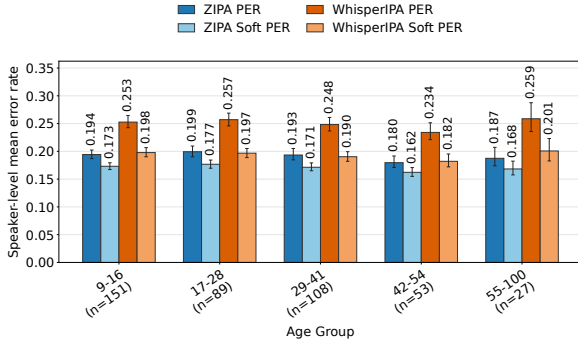
Figure 3: PER and Soft PER comparison by gender across demographic datasets. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.

Accent/Dialect: Figure 6 presents PER and Soft PER across accent and dialect groups in the SVC dataset. In SVC, six regional groups—“USA Southern”, “USA Western”, “USA Mid-Atlantic”, “USA Midland”, “USA Inland-North”, and “USA New England”—represent native speakers of American English of any racial or ethnic group. The “Asian” and “Latino” groups represent non-native speakers from Asia and Latin America who reside in any region of the USA.

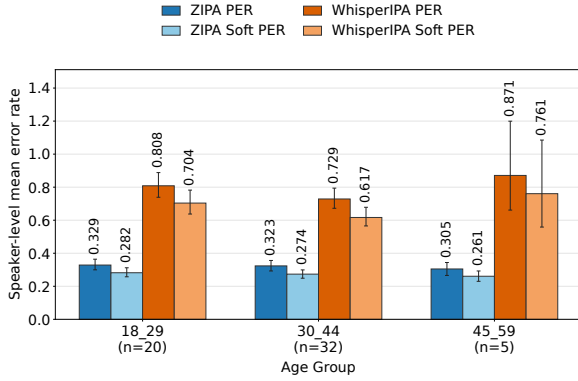
As shown in Figure 6a, most USA regional groups show similar or slightly below-average error for both models, while the “Latino” and “Asian” groups have the highest error, particularly for WhisperIPA — reflecting greater difficulty with accents that differ from SVC’s native regional varieties. The disparity heatmap (Figure 6b) and pairwise Welch’s t -tests confirm this is a robust effect (Table 11): “Latino” speakers have the largest positive disparity from the overall mean (+0.057/+0.037 ZIPA PER/Soft PER, +0.063/+0.034 WhisperIPA), and both “Latino” and “Asian” speakers have significantly higher PER

than nearly every individual USA regional group ($p < 0.001$ in most cases). “Latino” speakers are also significantly higher than “Asian” speakers under ZIPA (+0.033, +0.021; $p < 0.05$), though not under WhisperIPA. The six USA regional groups are largely indistinguishable from one another, with only a handful of pairwise contrasts reaching significance. Soft PER reduces the magnitude of these disparities but preserves the relative pattern, indicating the differences are not fully explained by the minor phonemic variation captured by the soft-equivalence mapping.

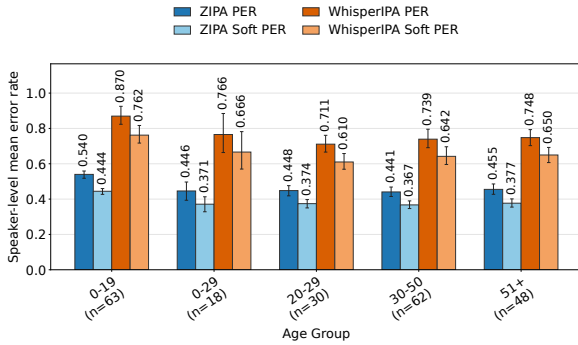
Key Takeaways: Overall, gender shows limited evidence of disparity, aside from a small but significant ZIPA effect in WAXAL. Accent and ethnicity show the most robust effects: in SVC, “Latino” and “Asian” speakers have significantly higher PER than nearly every USA regional group across both models; in EdAcc, “Black” speakers have significantly higher PER than “South Asian” speakers (and, for ZIPA, “White” speakers as well), while “Asian” speakers are not significantly different from any other group despite an elevated disparity from



(a) PER and Soft PER by age group for the SVC dataset.



(b) PER and Soft PER by age group for the EdAcc dataset.



(c) PER and Soft PER by age group for the CORAAL dataset.

Figure 4: PER and Soft PER comparison by age group across datasets. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.

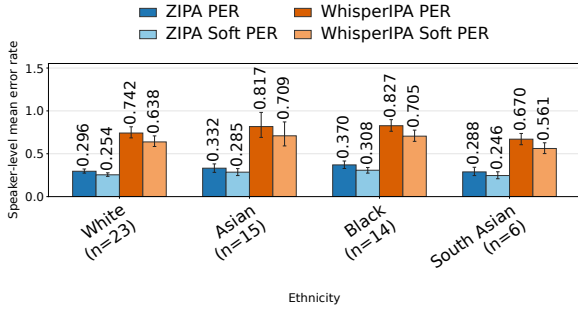
the overall mean. Age effects are more dataset-dependent: a robust elevation for the youngest speakers in CORAAL, a modest but significant effect in SVC’s 42–54 group, and no significant effect in EdAcc despite a large descriptive gap. Several estimates rest on small samples and should be interpreted cautiously. Overall, demographic performance patterns reflect an interaction of model behavior, dataset characteristics, and group representation rather than any single attribute.

PER vs. Soft PER: Soft PER generally reduces absolute error rates while preserving the relative performance patterns observed across demographic groups. If the disparities measured by standard PER were driven primarily by phonologically similar substitutions, including variations associated with accents or dialects, applying the soft-equivalence mapping would be expected to narrow the gaps between groups substantially. Instead, the persistence of similar group-level patterns under both metrics suggests that the observed differences also involve recognition errors not covered by the current mapping. The limited effect of Soft PER may additionally reflect incomplete coverage of demographic and cross-lingual phonemic variation within the equivalence classes. In this evaluation, standard PER and Soft PER therefore lead to broadly similar qualitative conclusions about relative group performance. Standard PER remains useful for comparison with prior work, whereas Soft PER provides a complementary measure for examining how acceptable phonemic variation contributes to the reported error rates.

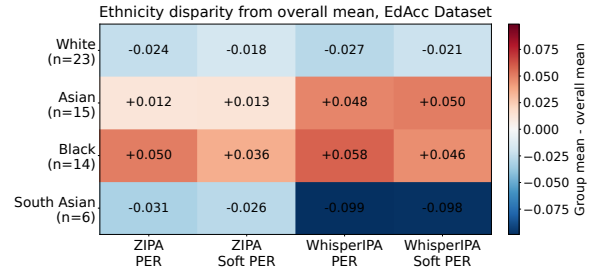
5 Conclusion

In this work, we evaluate bias and performance variation in IPA-based phoneme recognition models across languages and speaker demographics. We compare two open-source IPA transcription systems, ZIPA and WhisperIPA, using both standard PER and a proposed Soft PER metric that reduces penalties for linguistically similar phoneme substitutions. Across the multilingual evaluation, ZIPA consistently achieves lower error rates than WhisperIPA, while both models show substantial variation across languages. This performance gap may be partly attributable to differences in model architecture, training objective, and training data scale. Although Soft PER reduces absolute error rates for several languages, it does not remove cross-lingual performance disparities or change the overall model ranking. This indicates that some standard PER errors reflect acceptable phonemic variation, but that meaningful language-level differences remain even under relaxed matching.

The demographic evaluation also shows that the performance variation is not uniform across speaker identity characteristics. Gender-based differences are small and largely inconsistent across datasets and models, with one exception: ZIPA shows a small but statistically significant gender

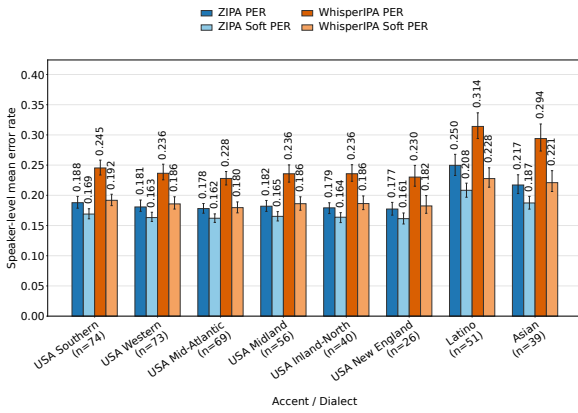


(a) PER and Soft PER by ethnicity group for the EdAcc dataset.

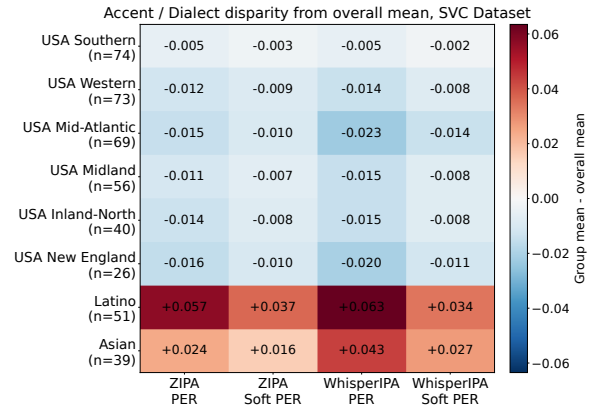


(b) Ethnicity disparity from the overall mean for the EdAcc dataset.

Figure 5: PER, Soft PER, and group-level disparity by ethnicity for the EdAcc dataset. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.



(a) PER and Soft PER by accent/dialect group for the SVC dataset.



(b) Accent/dialect disparity from the overall mean for the SVC dataset.

Figure 6: PER, Soft PER, and group-level disparity by accent/dialect for the SVC dataset. Bars and numerical labels represent speaker-level mean error rates. Error bars indicate 95% BCa bootstrap confidence intervals, with percentile bootstrap intervals used when BCa estimation does not converge or when $n < 10$.

effect in WAXAL. Accent and ethnicity show the most robust effects: in SVC, “Latino” and “Asian” speakers have significantly higher PER than nearly every USA regional group across both models, while in EdAcc, “Black” speakers have significantly higher PER than “South Asian” speakers, and, for ZIPA specifically, than White speakers as well; “Asian” speakers, despite an elevated disparity from the overall mean, are not significantly different from any individual ethnicity group under pairwise testing. Age effects are more dataset-dependent, with a robust elevation for the youngest speakers in CORAAL, a modest but significant effect in SVC’s 42–54 group, and no statistically significant effect in EdAcc despite a large descriptive gap for its smallest age group. Across demographic attributes, Soft PER lowers absolute error rates but generally preserves relative group-level patterns, suggesting that observed disparities are not fully

explained by minor phonemic variation alone.

Overall, our findings show that IPA-based ASR systems exhibit demographic and language-level disparities in measured PER. However, these disparities are measured against G2P references, which may not capture legitimate pronunciation variation and can conflate differences in realized pronunciation with ASR error. Although phoneme-based representations provide a promising foundation for low-resource and multilingual speech technologies, their performance can still vary substantially between languages, accents, and demographic groups. These results highlight the need for more careful evaluation of phoneme-level ASR systems, richer demographic and sociophonetic benchmarks, and evaluation metrics that distinguish genuine recognition errors from acceptable pronunciation variation.

Limitations

The key limitation of phoneme-level evaluation is the reliance on auto-generated IPA ground truth, because of the high cost of specialized manual phonemic annotation. Unlike orthographic transcriptions, phonemic annotations depend on G2P conversion and forced alignment, which can introduce systematic noise and model bias in the ground truth. Because our pipeline converts written transcripts into IPA representations, the resulting “ground truth” often reflects standardized pronunciation, which can encode biases of what constitutes “correct” speech into our analysis (Vythelingum et al., 2017; MacKenzie and Turton, 2020). This issue will naturally increase the error rates for accented speech, dialectal variation, and non-canonical pronunciations, where no single canonical phonemic form may exist (Fatema et al., 2024). Even with a softened PER, it may reflect annotation artifacts rather than true model error. Addressing this limitation may require future evaluation frameworks that utilize expert-validated phonemic annotations.

Our Soft PER hopes to partially mitigate sensitivity to annotation noise. While this acts as a necessary smoothing, it introduces a tradeoff between linguistic fidelity and metric standardization. Our definition of Soft PER may not be directly comparable to standard PER or WER, limiting cross-study comparability. Additionally, the equivalence mapping relies on English-prioritized relations when a surface phone appears across multiple languages. Although this provides a consistent normalization strategy, it may bias the metric toward English phonological interpretations and over-normalize distinctions that remain contrastive in other languages. Future work should explore language-adaptive mappings that better preserve cross-lingual phonological variation.

The analysis is further constrained by the demographic metadata available in the evaluated datasets. These categories do not capture finer sociophonetic variation such as code-switching, intra-speaker variability, or regional micro-dialects. In addition, certain demographic groups are underrepresented, which may limit the statistical robustness and generalizability of the results. Future work could address this by incorporating or creating a dataset with richer sociolinguistic annotations optimized for multilingual analysis to better understand the relationship among different dimensions.

Finally, while IPA provides a more fine-grained

representation of speech than grapheme-based transcription, it is an abstract acoustic representation. There is ongoing debate in linguistics regarding the extent to which phonemic representations can fully capture gradient phonetic variation (Mallaband, 2024; Mitterer et al., 2018). Dialectal variation and allophonic realizations may not be fully represented within standardized IPA inventories, which may further influence both the accuracy of the ground truth and evaluation results (Fatema et al., 2024). Hybrid evaluation frameworks in the future that jointly leverage phonemic representations and acoustic similarity measures could potentially provide a more accurate representation of ASR performance across diverse demographics.

References

- Surakshya Aryal, Mikel K Ngueajio, Saurav Keshari Aryal, and Gloria Washington. 2023. Hey, Siri! why are you biased against women?(student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16154–16155.
- Arun Babu, Changan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. In *Interspeech 2022*, pages 2278–2282.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Shikhar Bharadwaj, Chin-Jou Li, Yoonjae Kim, Kwanghee Choi, Eunjung Yeo, Ryan Soh-Eun Shim, Hanyu Zhou, Brendon Boldt, Karen Rosero Jacome, Calvin Chang, Darsh Agrawal, Keer Xu, Chao-Han Huck Yang, Jian Zhu, Shinji Watanabe, and David R. Mortensen. 2026. *Prism: Benchmarking phone realization in speech models*. *Preprint*, arXiv:2601.14046.
- Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. 2022. Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9):4419.
- Jay L Cunningham, Adinawa Adjagbodjou, Jeffrey Basoah, Jainaba Jawara, Kowe Kadoma, and Aaleyah Lewis. 2025. Toward responsible ASR for african american english speakers: A scoping review of bias and equity in speech technology. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 665–678.

- Massimo Marie Daul, Alessio Tosolini, and Claire Bower. 2026. [Linguistically informed tokenization improves ASR for underresourced languages](#). In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 31–37, Rabat, Morocco. Association for Computational Linguistics.
- Kumari Jayashree Devi and Parismita Sarma. 2026. Automated international phonetic alphabet (ipa) representation for assamese speech. In *ICT for Intelligent Systems*, pages 603–614, Singapore. Springer Nature Singapore.
- Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokor, Fiifi Baffoe Payin Winful, Kafui Kwashie Solaga, Jamal-Deen Abdulai, Akon Obu Ekpezu, and 24 others. 2026. [Waxal: A large-scale multilingual african language speech corpus](#). *Preprint*, arXiv:2602.02734.
- Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Performance disparities between accents in automatic speech recognition (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16200–16201.
- Hend ElGhazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025. Exploring gender disparities in automatic speech recognition technology. *arXiv preprint arXiv:2502.18434*.
- Anjie Fang, Simone Filice, Nut Limsopatham, and Oleg Rokhlenko. 2020. Using phoneme representations to build predictive models robust to asr errors. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 699–708.
- Kanij Fatema, Fazle Dawood Haider, Nirzona Ferdousi Turpa, Tanveer Azmal, Sourav Ahmed, Navid Hasan, Mohammad Akhlaqur Rahman, Biplab Kumar Sarkar, Afrar Jahin, Md. Rezuwan Hassan, Md Foriduzzaman Zihad, Rubayet Sabbir Faruque, Asif Sushmit, Mashrur Imtiaz, Farig Sadeque, and Syed Shahrier Rahman. 2024. [Ipa transcription of bengali texts](#). *Preprint*, arXiv:2403.20084.
- Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.
- Li Fu, Yu Xin, Sunlu Zeng, Lu Fan, Youzheng Wu, and Xiaodong He. 2025. Pac: Pronunciation-aware contextualized large language model-based automatic speech recognition. *arXiv preprint arXiv:2509.12647*.
- Robert C. Gale, Alexandra C. Salem, Gerasimos Fergadiotis, and Steven Bedrick. 2023. [Mixed orthographic/phonemic language modeling: Beyond orthographically restricted transformers \(BORT\)](#). In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 212–225, Toronto, Canada. Association for Computational Linguistics.
- Zébulon Goriely and Paula Buttery. 2025. [IPA CHILDES & G2P+: Feature-rich resources for cross-lingual phonology and phonemic language modeling](#). In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 502–521, Vienna, Austria. Association for Computational Linguistics.
- Jan Gorisch and Thomas Schmidt. 2024. [Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6564–6574, Torino, Italia. ELRA and ICCL.
- Maliha Jahan, Priyam Mazumdar, Thomas Thebaud, Mark Hasegawa-Johnson, Jesús Villalba, Najim Dehak, and Laureano Moro-Velazquez. 2025. Unveiling performance bias in asr systems: A study on gender, age, accent, and more. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Tyler Kendall and Charlie Farrington. 2018. The corpus of regional african american language. *Version*, 6:1.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.
- Rostislav Kolobov, Olga Okhapkina, Olga Omelchishina, Andrey Platunov, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. [Mediaspeech: Multilanguage asr benchmark and dataset](#). *Preprint*, arXiv:2103.16193.
- Jaeyoung Lee, Masato Mimura, and Tatsuya Kawahara. 2025. Leveraging ipa and articulatory features as effective inductive biases for multilingual asr training. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and 1 others. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.
- Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Interspeech*, pages 711–714.

- Laurel MacKenzie and Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of british english. *Linguistics Vanguard*, 6(s1):20180061.
- Laura Jane Mallaband. 2024. The agreement of phonetic transcriptions between paediatric speech and language therapists transcribing a disordered speech sample. *International Journal of Language & Communication Disorders*, 59(5):1981–1995.
- Joshua L Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual "be". In *Interspeech*, pages 626–630.
- Yangyang Meng, Jinpeng Li, Guodong Lin, Yu Pu, Guanbo Wang, Hu Du, Zhiming Shao, Yukai Huang, Ke Li, and Wei-Qiang Zhang. 2025. Dolphin: A large-scale automatic speech recognition model for eastern languages. *arXiv preprint arXiv:2503.20212*.
- Holger Mitterer, Eva Reinisch, and James M McQueen. 2018. Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, 98:77–92.
- Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.
- David R Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan W Black, Florian Metze, and Graham Neubig. 2020. Allovera: A multilingual allophone database. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5329–5336.
- Mikel K Ngueajio and Gloria Washington. 2022. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *International conference on human-computer interaction*, pages 421–440. Springer.
- Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.
- Kentaro Onda, Satoru Fukayama, Daisuke Saito, and Nobuaki Minematsu. 2026. *Advanced modeling of interlanguage speech intelligibility benefit with 11-12 multi-task learning using differentiable k-means for accent-robust discrete token-based asr*. *Preprint*, arXiv:2601.19767.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Francis McCann Ramirez, Luka Chkhetiani, Andrew Ehrenberg, Robert McHardy, Rami Botros, Yash Khare, Andrea Vanzo, Taufiqzaman Peyash, Gabriel Oexle, Michael Liang, Ilya Sklyar, Enver Fakhani, Ahmed Etefy, Daniel McCrystal, Sam Flamini, Domenic Donato, and Takuya Yoshioka. 2024. *Anatomy of industrial scale multilingual asr*. *Preprint*, arXiv:2404.09841.
- Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. *The edinburgh international accents of english corpus: Towards the democratization of english asr*. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Michael Scott, Siyu Liang, Alicia Wassink, and Gina-Anne Levow. 2025. A sociophonetic analysis of racial bias in commercial asr systems using the pacific northwest english corpus. *arXiv preprint arXiv:2510.22495*.
- Chloe Sekkat, Fanny Leroy, Salima Mdhaffar, Blake Perry Smith, Yannick Estève, Joseph Dureau, and Alice Coucke. 2024. *Sonos voice control bias assessment dataset: A methodology for demographic bias assessment in voice assistants*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15056–15075, Torino, Italia. ELRA and ICCL.
- Dana Serditova, Kevin Tang, and Jochen Steffens. 2025. *Automatic Speech Recognition Biases in Newcastle English: an Error Analysis*. In *Interspeech 2025*, pages 3204–3208.
- Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981.
- Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. *Universal Automatic Phonetic Transcription into the International Phonetic Alphabet*. In *Interspeech 2023*, pages 2548–2552.
- Rachael Tatman. 2017. *Gender and dialect bias in YouTube's automatic captions*. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.
- Rachael Tatman and Conner Kasten. 2017. *Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions*. In *Interspeech 2017*, pages 934–938.
- Melissa Torgbi, Andrew Clayman, Jordan J Speight, and Harish Tayyar Madabushi. 2025. Adapting whisper for regional dialects: Enhancing public services for vulnerable populations in the united kingdom. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 29–38.

- Kévin Vythelingum, Yannick Estève, and Olivier Rosee. 2017. Error detection of grapheme-to-phoneme conversion in text-to-speech synthesis using speech signal and lexical context. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 692–697. IEEE.
- Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.
- Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*.
- Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *International Conference on Learning Representations*, volume 2024, pages 44440–44455.
- Yuanyuan Zhang, Aaricia Herygers, Tanvina Patel, Zhengjun Yue, and Odette Scharenborg. 2023. Exploring data augmentation in bias mitigation against non-native-accented speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.
- Yuanyuan Zhang, Yixuan Zhang, Bence Halpern, Tanvina Patel, and Odette Scharenborg. 2022. [Mitigating bias against non-native accents](#). In *Interspeech 2022*, pages 3168–3172.
- Jian Zhu, Farhan Samir, Eleanor Chodroff, and David R Mortensen. 2025. Zipa: A family of efficient models for multilingual phone recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19568–19585.
- Jian Zhu, Changbing Yang, Farhan Samir, and Jahu-rul Islam. 2024. [The taste of ipa: Towards open-vocabulary keyword spotting and forced alignment in any language](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 750–772, Mexico City, Mexico. Association for Computational Linguistics.
- Piotr Żelasko, Siyuan Feng, Laureano Moro Velázquez, Ali Abavisani, Saurabhchand Bhati, Odette Scharenborg, Mark Hasegawa-Johnson, and Najim Dehak. 2022. [Discovering phonetic inventories with crosslingual automatic speech recognition](#). *Computer Speech & Language*, 74:101358.

A Appendix

A.1 Dataset Details

Table 1 summarizes the multilingual speech datasets used for cross-lingual evaluation, including the languages represented in each dataset and the dataset split used for analysis. Together, these datasets support evaluation across both high-resource and lower-resource languages.

Table 2 provides a detailed summary of the English-language datasets and the speaker-level demographic metadata available in each dataset. These metadata fields, including age, gender, ethnicity, accent, and dialectal region, are used to evaluate whether IPA-based ASR models exhibit systematic performance variation across demographic groups.

A.2 IPA Transcription Post-Processing

To ensure that the evaluation focused on phonemic correspondence rather than superficial transcription differences, we apply the same normalization procedure to both the G2P-generated ground-truth IPA transcriptions and the model-generated outputs before computing PER and Soft PER as described in Table 3. The normalization rules are designed to reduce variation caused by stress notation, IPA symbol variants, tokenization conventions, and formatting artifacts, while preserving the underlying phone sequence as much as possible.

After applying these transformations, the normalized phone strings were re-tokenized into whitespace-separated phone sequences. This produced a consistent representation across reference and generated outputs, reducing the impact of orthographic IPA variants and tokenization artifacts on downstream evaluation.

A.3 Soft PER Details

Table 4 lists the transitive phonetic equivalence classes used in the first tier of Soft PER. These classes group phones that are treated as interchangeable during evaluation, allowing the metric to avoid penalizing substitutions that correspond to linguistically similar or allophonic variants. The Tier 1 mapping is applied universally across languages and is intended to capture broad phonemic variation that should not necessarily be counted as a recognition error.

The Tier 1 equivalence classes are constructed from two sources of phonetic similarity. First,

language-specific allophonic mappings are derived from AlloVera, which links surface phones to canonical phonemes across multiple languages. Second, articulatory feature representations from PHOIBLE are used to identify canonical phonemes that differ by only one articulatory feature. These phones are merged into shared transitive classes, producing a mapping that spans 62 equivalence classes and 254 phones. During Soft PER computation, substitutions within the same Tier 1 class receive zero penalty, allowing the metric to better distinguish substantial recognition errors from acceptable phonetic variation.

Table 5 lists the language-specific non-transitive similarity pairs used in the second tier of Soft PER. Unlike the Tier 1 classes, these pairs are not merged into broader equivalence classes. Instead, they capture cases where the same surface phone corresponds to different canonical phonemes across languages. This allows Soft PER to recognize language-specific pronunciation overlap without incorrectly making all connected phonemes globally interchangeable.

The Tier 2 mapping is also derived from AlloVera. When a surface phone maps to different canonical phonemes in different languages, the corresponding canonical phonemes are recorded as a direct similarity pair. During evaluation, these pairs are applied only when relevant to the target language. Substitutions covered by Tier 2 receive zero penalty, but the relation remains non-transitive; for example, if two separate language-specific mappings relate one phone to two different phones, Soft PER does not automatically treat those two phones as equivalent to each other. This design preserves language-specific phonological distinctions while still reducing penalties for acceptable cross-lingual or allophonic variation.

A.4 Pairwise Welch’s *t*-test Details

Tables 6–11 report pairwise comparisons of speaker-level PER between demographic groups within each dataset, for the demographic attributes examined in Section 4.2: gender, age, ethnicity, and accent/dialect. For every pair of groups within a demographic attribute, we compute the difference in group-mean PER (row minus column) and test the difference using a two-sided Welch’s *t*-test. This is reported separately for ZIPA and WhisperIPA, under both standard PER and Soft PER.

In each table, the reported value is the point-estimate difference in speaker-level mean PER be-

Table 1: Dataset Metadata and Speech Characteristics

Dataset	Languages / Varieties	Subset [†]	Collection Period	Speech Style	Recording Environment
Fleurs_IPA	English, Hindi, Bangla, Panjabi, Tamil, Telugu	Test	2020–2022	Read speech (utterance-level)	Unedited original recordings
MediaSpeech	Arabic, French, Spanish, Turkish	N/A	2020	Read and spontaneous speech	YouTube broadcasts/channels
Waxal	Shona	Train	Jan. 2021–Mar. 2024	Natural, spontaneous speech	Natural environments
EdAcc	English	Test	2023	Naturalistic, spontaneous interactions	Zoom recordings
SVC	English	Train	2024	Read speech	Clean, close-field environments
CORAAL	African American English	Train	2023	Sociolinguistic interviews	Unedited original recordings

[†] Dataset partition used for bias evaluation.

Table 2: Dataset Languages and Demographic Attributes

Dataset	Language(s)	Demographic Attributes
WAXAL	Shona	gender
EdAcc	English	gender, year_of_birth, ethnicity, education, l1, l2, accent, english_since, countries_lived, lang_home, lang_friends
SVC Bias Assessment	English	age_group, gender, dialectal_region
CORAAL	African American English	city, se_group, age_group, gender

tween the row groups and column groups. A positive value indicates that the row group has a higher (worse) mean PER than the column group.

Group-level means and their 95% confidence intervals, computed using bias-corrected and accelerated (BCa) bootstrap with 10,000 resamples (percentile intervals when a group contains fewer than 10 speakers or when the BCa procedure does not converge), are the same estimates used for the error bars in the demographic performance figures in Section 4.2.

Table 3: Normalization rules applied to both ground-truth IPA transcriptions and model-generated outputs prior to evaluation.

Category	Normalization behavior
Stress marks	Primary and secondary stress marks, such as ' and ˌ, were removed.
Length marks	Length markers, including ː, ˑ, and the ASCII colon :, were removed.
Rhotics	Rhotic consonants such as ɹ, ɻ, ɹ̥, and ɹ̥̄ were mapped to r. Rhotic vowels such as ə̣ and ɜ̣ were expanded to ə ɹ and ɛ ɹ, respectively. Rhoticity marks, such as ə̣ˀ, were also normalized to a vowel followed by r.
Diphthongs	Joined diphthong symbols were tokenized as separate phone sequences. For example, eɪ was normalized to e ɪ.
Affricates	Joined and tie-barred affricates were tokenized as separate phones. For example, tʃ, tʃ̄, and tʃ̄̄ were normalized to t ʃ.
Token separators	Whitespace, SentencePiece markers such as _, hyphens, slashes, and similar separators were collapsed into phone-token boundaries.
Combining marks	Separated combining marks were attached to the preceding phone. For example, a ˜ was normalized to ā.
Symbol choices	Common spelling variants were standardized. For example, ASCII g was normalized to IPA g, and dark-/l variants such as ɫ̥, ɫ̥̄, and separated l ɫ were normalized to ɫ.

Table 4: Phonetic Transitive Equivalence Classes (Tier 1).

Index	Representations
0	b', p'
1	b, b̥, b̥̄
2	dʒ, dʒ̄
3	d, d̥
4	e, e̥
5	kʷ, k̥ʷ
6	k, k̥
7	o, o̥, ō̥
8	pf, pf̄
9	ts, ts̄
10	tɕ, tɕ̄
11	tʃ, tʃ̄
12	tʷ, t̥ʷ
13	t, t̥
14	v, v̥
15	xʲ, x̥ʲ
16	ɲ, ɲ̥
17	ɑ, ɑ̄
18	ɛ, ɛ̄
19	g, g̥
20	rʲ, r̥ʲ
21	ɻ, ɻ̥

Table 5: Language Specific Non-Transitive Pairs (Tier 2).

Phoneme 1	Phoneme 2	Language	Phoneme 1	Phoneme 2	Language
a	a:	jpn, tur	a	æ	deu, eng
a	ɑ	cmn, eng	a	ɔ	eng, jav
a	ə	eng, rus	a	ɛ	cmn, eng
a	u	eng, rus	a:	ɑ	deu, eng
b	b ^j	jpn, rus	b	p	amh, eng, kaz, rus, spa
b	v	spa, tur	b ^j	p ^j	rus
d	d ₃	ita, tgl	d	t	eng, kaz, rus, vie
d	ð	eng, spa	d	ʈ	spa, tgl
d ^j	t ^j	rus	d ₃	ts	ita, tgl
d ₃	ʒ	amh, eng, ita, tgl	e	e:	deu, eng, jpn, tur
e	i	eng, tgl	e	æ	eng, tur
e	ɛ	deu, eng	e	ɪ	eng, tgl
e	u	cmn, eng	h	r	fra, kaz
h	x	deu, tur	h	ç	deu, jpn
i	i:	deu, eng, jpn, tur	i	ɛ	eng, tgl
i	ɪ	deu, eng	j	y	eng, jpn
j	ç	deu, rus, spa	j	fi	cmn, fra
j	ɪ	eng	j	ʃ	deu, spa
k	q	eng, jav	k	x	deu, tgl
k	g	amh, deu, eng, kaz, rus	k	g ^j	rus
m	m ^j	jpn, rus	m	n	eng, ita
m	ŋ	eng, ita, tgl	m	ɲ	eng, fra, ita
n	ŋ	eng, spa, tgl	n	ɲ	eng, fra, ita
o	o:	deu, eng	o	u	eng, tgl
o	ɔ	eng	o	ʊ	eng, tgl
p	p ^j	jpn, rus	p	q	jpn, vie
q	r	fra, kaz	q	s	eng, jpn
r	x	deu, fra, kaz, spa	r	ɹ	eng, spa
r	ʈ	jpn, spa, tgl	r	ʃ	eng, kaz
r	ʒ	eng, kaz	s	z	amh, deu, eng, kaz
s	ʃ	eng, tgl	s	θ	eng, spa
t	t ^j	spa, tgl	ts	tʃ	spa, tgl
ts	z	ita, jpn	u	u:	deu, eng, jpn, tur
u	w	eng, ita	u	i	jpn, rus
u	ʊ	deu, eng	w	ʊ	eng
x	g	rus, spa	y	y:	deu, fra
ø	ø:	deu, fra	ø	œ	deu, fra
ŋ	ɲ	eng, ita	ɔ	ɔ	amh, eng
ə	ɛ	amh, eng	g	u	tgl, tur
i	ɪ	amh, eng	i	ɲ	rus, vie
i	ʊ	amh, eng	ɹ	ʈ	eng, tgl
f	ʒ	eng, kaz	ʃ	ʒ	eng, spa

Table 6: Gender bias analysis across all four demographic datasets. Values represent the difference in speaker-level mean PER between the row and column groups (row — column). Statistically significant difference in mean between two groups via Welch’s *t*-test is shown as *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

Dataset	Groups	ZIPA PER	ZIPA Soft PER	Whisper PER	Whisper Soft PER
SVC	Female vs Male	+0.001	+0.001	-0.001	+0.000
EdAcc	Female vs Male	-0.017	-0.012	+0.028	+0.028
WAXAL	Female vs Male	+0.067**	+0.067**	-0.149	-0.149
CORAAL	Female vs Male	-0.025	-0.019	+0.011	+0.015

Table 7: Ethnicity bias analysis on the EdAcc dataset. Values are the difference in speaker-level mean PER between the row and column groups (row – column). Statistically significant difference in mean between two groups via Welch’s t -test is shown as: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		Black	South Asian	White
ZIPA PER	Asian	-0.038	+0.044	+0.036
ZIPA Soft PER		-0.023	+0.039	+0.031
Whisper PER		-0.010	+0.147	+0.075
Whisper Soft PER		+0.004	+0.148	+0.071
ZIPA PER	Black		+0.082*	+0.074*
ZIPA Soft PER			+0.062	+0.054*
Whisper PER			+0.157**	+0.085
Whisper Soft PER			+0.144*	+0.067
ZIPA PER	South Asian			-0.008
ZIPA Soft PER				-0.008
Whisper PER				-0.072
Whisper Soft PER				-0.077

Table 8: Age-group bias analysis on the SVC dataset. Values are the difference in speaker-level mean PER between the row and column groups (row – column). Statistically significant difference in mean between two groups via Welch’s t -test is shown as: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		17-28	29-41	42-54	55-100
ZIPA PER	9-16	-0.005	+0.001	+0.014*	+0.007
ZIPA Soft PER		-0.004	+0.002	+0.011**	+0.005
Whisper PER		-0.004	+0.005	+0.019	-0.006
Whisper Soft PER		+0.001	+0.008	+0.016*	-0.003
ZIPA PER	17-28		+0.006	+0.019**	+0.012
ZIPA Soft PER			+0.006	+0.015**	+0.009
Whisper PER			+0.009	+0.023*	-0.002
Whisper Soft PER			+0.007	+0.015*	-0.004
ZIPA PER	29-41			+0.013	+0.006
ZIPA Soft PER				+0.009	+0.003
Whisper PER				+0.014	-0.011
Whisper Soft PER				+0.008	-0.011
ZIPA PER	42-54				-0.007
ZIPA Soft PER					-0.006
Whisper PER					-0.025
Whisper Soft PER					-0.019

Table 9: Age-group bias analysis on the EdAcc dataset. Values are the difference in speaker-level mean PER between the row and column groups (row – column). Statistically significant difference in mean between two groups via Welch’s t -test is shown as: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		30-44	45-59
ZIPA PER	18-29	+0.006	+0.024
ZIPA Soft PER		+0.008	+0.021
Whisper PER		+0.079	-0.063
Whisper Soft PER		+0.087	-0.057
ZIPA PER	30-44		+0.018
ZIPA Soft PER			+0.013
Whisper PER			-0.142
Whisper Soft PER			-0.144

Table 10: Age-group bias analysis on the CORAAL dataset. Values are the difference in speaker-level mean PER between the row and column groups (row – column). Statistically significant difference in mean between two groups via Welch’s t -test is shown as: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		0-29	20-29	30-50	51+
ZIPA PER	0-19	+0.094**	+0.092***	+0.099***	+0.085***
ZIPA Soft PER		+0.073**	+0.070***	+0.077***	+0.067***
Whisper PER		+0.104	+0.159***	+0.131***	+0.122***
Whisper Soft PER		+0.096	+0.152***	+0.120**	+0.112**
ZIPA PER	0-29		-0.002	+0.005	-0.009
ZIPA Soft PER			-0.003	+0.004	-0.006
Whisper PER			+0.055	+0.027	+0.018
Whisper Soft PER			+0.056	+0.024	+0.016
ZIPA PER	20-29			+0.007	-0.007
ZIPA Soft PER				+0.007	-0.003
Whisper PER				-0.028	-0.037
Whisper Soft PER				-0.032	-0.040
ZIPA PER	30-50				-0.014
ZIPA Soft PER					-0.010
Whisper PER					-0.009
Whisper Soft PER					-0.008

Table 11: Accent/dialect bias analysis on the SVC dataset. Region names abbreviated (Inland-N.=USA Inland-North, Mid-Atl.=USA Mid-Atlantic, New Eng.=USA New England). Values are the difference in speaker-level mean PER between the row and column groups (row – column). Statistically significant difference in mean between two groups via Welch’s t -test is shown as: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Positive values indicate higher PER for the first-listed group.

		Latino	Inland-N.	Mid-Atl.	Midland	New Eng.	Southern	Western
ZIPA PER	Asian	-0.033**	+0.038***	+0.039***	+0.035***	+0.040***	+0.029**	+0.036***
ZIPA Soft PER		-0.021*	+0.023***	+0.025***	+0.022***	+0.026***	+0.018**	+0.024***
Whisper PER		-0.020	+0.058***	+0.066***	+0.058***	+0.064***	+0.049***	+0.058***
Whisper Soft PER		-0.007	+0.035**	+0.041***	+0.035**	+0.039**	+0.029**	+0.035**
ZIPA PER	Latino		+0.071***	+0.072***	+0.068***	+0.073***	+0.062***	+0.069***
ZIPA Soft PER			+0.044***	+0.046***	+0.043***	+0.047***	+0.039***	+0.045***
Whisper PER			+0.078***	+0.086***	+0.078***	+0.084***	+0.069***	+0.078***
Whisper Soft PER			+0.042***	+0.048***	+0.042***	+0.046***	+0.036***	+0.042***
ZIPA PER	Inland-N.			+0.001	-0.003	+0.002	-0.009	-0.002
ZIPA Soft PER				+0.002	-0.001	+0.003	-0.005*	+0.001
Whisper PER				+0.008	+0.000	+0.006	-0.009	+0.000
Whisper Soft PER				+0.006	+0.000	+0.004	-0.006	+0.000
ZIPA PER	Mid-Atl.				-0.004	+0.001	-0.010	-0.003
ZIPA Soft PER					-0.003	+0.001	-0.007*	-0.001
Whisper PER					-0.008	-0.002	-0.017*	-0.008
Whisper Soft PER					-0.006	-0.002	-0.012	-0.006
ZIPA PER	Midland					+0.005	-0.006	+0.001
ZIPA Soft PER						+0.004	-0.004	+0.002
Whisper PER						+0.006	-0.009	+0.000
Whisper Soft PER						+0.004	-0.006	+0.000
ZIPA PER	New Eng.						-0.011	-0.004
ZIPA Soft PER							-0.008*	-0.002
Whisper PER							-0.015	-0.006
Whisper Soft PER							-0.010	-0.004
ZIPA PER	Southern							+0.007
ZIPA Soft PER								+0.006
Whisper PER								+0.009
Whisper Soft PER								+0.006

Lost in Compression: Calibration Bias and the Accent-WER Gap in Speech LLMs

Angad Ripudaman Singh Bajwa

Independent Researcher

angadisthere@gmail.com

Abstract

GPTQ and similar calibration-based quantization methods tune their rounding to a small calibration sample, in practice usually convenient, native-accented read speech. Edge devices disproportionately run quantized models, so the speakers most likely to depend on the compressed model are often the ones calibration represented least. Compression is known to amplify bias in text LLMs and across languages in multilingual speech models, but whether it does so within a single language, between native and non-native accents, in a modern speech LLM, remains untested. We quantize the LLM backbone of Qwen2-Audio-7B-Instruct with GPTQ, hold the recipe fixed and vary only which EdAcc speakers fill the calibration set (native-only, balanced, L2-only), measuring the native-vs-L2 word error rate (WER) gap at matched overall WER. We expected native-only calibration to widen the gap and balanced calibration to close it. Instead, homogeneous calibration, native-only or L2-only, narrows the gap while balanced calibration does almost nothing. It recurs when we quantize the audio encoder, weakens at 3-bit and collapses at 2-bit.

1 Introduction

Post-training quantization (PTQ) methods such as GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) fit their scaling and channel-protection decisions to a small calibration sample, usually a few hundred examples. Whoever assembles that sample tends to reach for whatever is convenient, which usually skews toward majority, native-accented, read speech. That matters because resource-constrained and edge devices disproportionately run compressed models, so the group most likely to receive the compressed model is often the group calibration represented least.

Compression amplifying bias is already established for text LLMs (Marcuzzi et al., 2026; Rath and Maliakkal, 2026) and for language identity in

multilingual, multitask speech-model compression (Ferraz, 2024). Calibration-set composition is likewise known to affect PTQ and pruning accuracy in general (Williams and Aletras, 2024). What nobody has tested is a demographic axis within a single language, native versus non-native (L2) accent, in a speech LLM rather than a classical ASR system. Even if such a gap moves, it is worth asking why, whether it is the calibration population itself or just degradation from compression in general.

This paper does three things. First, we quantize Qwen2-Audio-7B-Instruct’s LLM backbone to 4-bit, 3-bit and 2-bit with GPTQ and track the native-vs-L2 WER gap on EdAcc (Sanabria et al., 2023), a corpus of accented conversational English. Second, holding the quantizer and the calibration size fixed, we swap only which speakers’ audio fills the calibration set (native-only, balanced, L2-only), so that matched overall WER across conditions serves as an aggregate-WER control: if overall WER does not move across conditions, whatever happens to the gap can be attributed to calibration composition rather than to one condition being compressed harder. Third, we characterise a small, directionally consistent effect at 4-bit and show that the same manipulation on the mechanistically distinct audio encoder points the same way, which we read as convergent evidence rather than a strict replication. The direction of the effect is the opposite of what we expected. That reversal is the central finding.

2 Related Work

The closest prior work is Ferraz (2024), who show that multitask, multilingual Whisper compression produces disparity across languages. We differ in modality (a speech LLM, not classical ASR), axis (demographic accent within one language, not language identity) and design (an explicit calibration-composition swap at matched overall WER, not a

compression-language correlation). [Marcuzzi et al. \(2026\)](#) and [Rath and Maliakkal \(2026\)](#) show quantization reshapes bias and alignment in text LLMs differently across precision levels, part of why we sweep bit-widths rather than testing one. [Williams and Aletras \(2024\)](#) shows calibration-data choice affects PTQ and pruning accuracy generally, supporting the mechanism we test, though without a fairness axis.

3 Method

3.1 Model and quantization targets

Qwen2-Audio-7B-Instruct ([Chu et al., 2024](#)) has 8.40B parameters, which by our count split into an audio encoder (`audio_tower`, 7.6%), a small projector (0.1%), an LLM decoder (`language_model`, 84.7%) and an output head (`lm_head`, 7.6%). Our primary condition quantizes only `language_model` (the backbone) and freezes everything else at FP16. A mirror-image condition instead quantizes `audio_tower` and the projector, the encoder condition, freezing the LLM decoder and `lm_head`. Because the frozen components still run during calibration, the backbone Hessian is computed from real accent-conditioned post-encoder activations rather than any hand-rolled embedding capture, and the same is true in reverse for the encoder condition. We run GPTQ ([Frantar et al., 2023](#)) through the `llm-compressor` library ([AI and vLLM Project, 2024](#)) with symmetric int weights and a group size of 128, at W4/W3/W2 for the backbone and W4 for the encoder. The library has no named preset for 3-bit or 2-bit, so we built the scheme by hand, keeping the same shape as the W4 preset and changing only the bit-width. We also use the library’s non-sequential calibration pipeline instead of its default error-propagating one, because the default pipeline’s graph-tracing calibration path is not compatible with this model’s forward pass.

3.2 Data and grouping

We evaluate on the validation split of EdAcc ([Sanabria et al., 2023](#)), the corpus’s own dev portion, 9,848 utterances from 62 speakers speaking conversational English with a range of accents. Every speaker carries a first-language field (11), either one of 8 English varieties (Southern British English, Indian English, Jamaican English and so on) or one of 21 non-English languages (Spanish, Mandarin, Arabic and others). The two never over-

lap, so every speaker falls unambiguously on one side. We call a speaker native if their first language is one of the English varieties (23 speakers, 3,706 utterances) and L2 if English came later (39 speakers, 6,142 utterances). These labels are about language background, not skill, so an L2 speaker here is not assumed to be less fluent, only accented by a first language other than English. This native/L2 split, and the gap in transcription accuracy between the two groups, is the central quantity the paper measures.

Before scoring we drop 191 utterances that would otherwise distort the gap. Of these, 62 carry EdAcc’s `IGNORE_TIME_SEGMENT_IN_SCORING` marker and 129 have a reference transcript that is only a non-speech tag such as `<LAUGH>`. Scoring either kind would count as a spurious error, and if such tags occur more for one group than the other that noise would land unevenly and distort the native/L2 gap itself rather than adding harmless noise equally to both sides.

3.3 Calibration conditions

For each condition, native, balanced and L2, we select 128 calibration utterances, sampling round-robin across speakers within the eligible group so that no single speaker dominates the calibration Hessian. The native and L2 conditions draw all 128 from their group. Balanced splits them 64/64. Everything else about the GPTQ setup is identical across conditions. Because overall WER is matched across the three W4 (or W3) conditions, that match acts as an aggregate-WER control, so any resulting gap movement cannot be attributed to one condition being compressed more severely than another. At W4 this is not a degradation control, since overall WER sits at or below FP16 (0.331–0.332 vs 0.336), so the conditions are matched to each other rather than degraded from full precision. Matched aggregate WER need not mean matched per-group behaviour, so we also report group-wise WER below.

3.4 Metrics

We report corpus-level WER per group, the total edit distance over total reference words rather than a mean of per-utterance rates. Two quantities carry the analysis. The **gap** is $WER(L2) - WER(native)$, how much worse the model transcribes L2 speakers than native speakers for a given model. The Δgap is $gap(quantized) - gap$

gap(FP16), how much quantization changes that gap relative to full precision, so a negative Δgap means quantization narrowed the native-vs-L2 gap.

Confidence intervals come from a speaker-clustered bootstrap (2,000 resamples) that draws speakers with replacement rather than utterances, since utterance counts per speaker range from 27 to 434. The bootstrap is stratified by group. Each replicate draws the 23 native and the 39 L2 speakers with replacement separately, preserving each group’s size. Corpus-level edit distance and reference-word counts are then summed over the drawn speakers (a speaker drawn k times contributes k times). For Δgap and the direct condition contrasts we pair the resampling, applying the same speaker draw to both files, since both are scored on identical utterances. Pairing cancels speaker-specific noise common to both and gives a tighter estimate than treating them as independent.

The intervals do the inferential work here. The effects are small in absolute terms, the gap and its movement both on the order of a single WER point, comparable to the wobble a different draw of 62 speakers could produce on its own. A bare point estimate like $\Delta\text{gap} = -0.010$ says nothing about how much it would shift under a different speaker sample, so throughout we treat an effect as real only when its 95% interval excludes zero rather than reading into the sign or size of a point estimate alone.

For reproducibility, GPTQ runs through the `llm-compressor` library (AI and vLLM Project, 2024) and WER is scored with `jiwer`. We report word error rate rather than the related MER or WIL measures (Morris et al., 2004). Decoding is greedy and the inference prompt is “Transcribe the English audio to text. Output only the transcription, with no punctuation or capitalization.”

4 Results

The FP16 baseline (Table 1) has overall WER 0.336 (native 0.330, L2 0.341), a small native-vs-L2 gap of +0.011 whose 95% CI $[-0.038, +0.062]$ already crosses zero given only 23 native speakers. We therefore rely on the tighter paired Δgap rather than each condition’s raw gap CI for the headline comparisons. At W4 overall WER is tightly matched across the three backbone conditions (0.331–0.332) and the three encoder conditions (0.333–0.339), so the aggregate-WER control holds and any movement in the gap is not just one

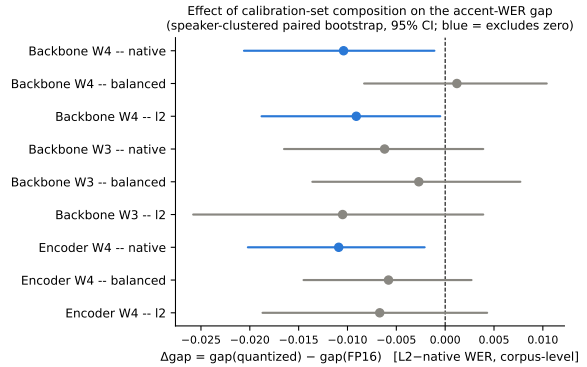


Figure 1: Δgap by calibration condition, target and bit-width. Blue intervals exclude zero.

condition being compressed harder. Figure 1 plots Δgap with its paired CI for every condition. Table 1 is the paper’s primary evidence. It reports the corpus-level gap and how much quantization moves it. Table 2 is diagnostic rather than a second headline result. It decomposes that same movement into each group’s own WER, native and L2 scored separately, to check what is driving the gap rather than just that it moved.

Two results stand out. At W4, native-only calibration narrows the gap relative to FP16 in both the backbone (-0.0104 , CI $[-0.0206, -0.0011]$) and the encoder (-0.0109 , CI $[-0.0202, -0.0021]$), the opposite of what we expected if the field’s default calibration population were what widens disparity. L2-only calibration moves the same way in the backbone (-0.0091 , CI $[-0.0188, -0.0005]$), with a similar but non-significant estimate in the encoder. Balanced calibration is the condition closest to no effect in both targets ($+0.0012$ and -0.0058 , both CIs crossing zero), which is not the fix we expected balancing to be. The pattern is not an artifact of one component, appearing in the same direction and roughly the same magnitude across two mechanistically different targets, a decoder stack and an acoustic encoder. Because those are different mechanisms rather than the same experiment re-run, we read the encoder result as convergent evidence for the backbone finding, not an independent replication. We treat the W4 Δgap and the direct contrasts as primary analyses. The per-group breakdown and the macro-weighted check below are secondary evidence that the effect is not an artifact of corpus-level weighting or leveling-down.

The narrowing is not leveling-down. A gap can close two ways, the worse-off group improving or

Target	Condition	Bits	Overall WER	Gap (L2–native)	Δ gap vs. FP16 [95% CI]
–	FP16	16	0.336	+0.011	–
Backbone	native	4	0.332	+0.000	–0.0104 [–0.0206, –0.0011]
Backbone	balanced	4	0.331	+0.012	+0.0012 [–0.0083, +0.0104]
Backbone	L2	4	0.332	+0.002	–0.0091 [–0.0188, –0.0005]
Backbone	native	3	0.339	+0.005	–0.0062 [–0.0165, +0.0039]
Backbone	balanced	3	0.347	+0.008	–0.0027 [–0.0136, +0.0077]
Backbone	L2	3	0.341	+0.000	–0.0105 [–0.0258, +0.0039]
Backbone	native	2	1.913	+0.131	+0.1167 [–0.1516, +0.3708]
Backbone	balanced	2	2.515	+0.173	+0.1587 [–0.2668, +0.5670]
Backbone	L2	2	1.801	+0.031	+0.0129 [–0.2349, +0.2601]
Encoder	native	4	0.333	+0.000	–0.0109 [–0.0202, –0.0021]
Encoder	balanced	4	0.335	+0.005	–0.0058 [–0.0145, +0.0027]
Encoder	L2	4	0.339	+0.004	–0.0067 [–0.0187, +0.0043]

Table 1: All conditions. Gap and Δ gap are corpus-level WER(L2)–WER(native); CIs are a paired, speaker-clustered bootstrap (2,000 resamples). W2 rows are reported for completeness but reflect uninterpretable transcription collapse, not a meaningful accent-specific effect.

the better-off group being dragged down to meet it. Table 1 shows only the net effect, gap(condition) minus gap(FP16), and cannot tell these two apart. Table 2 can. It gives each group’s own WER and its paired bootstrap change from FP16. At W4 the gap closes the good way: L2 WER falls (backbone native –0.0088, backbone L2 –0.0080, encoder native –0.0072, all 95% CIs excluding zero) while native WER does not significantly rise (all three Δ Native CIs cross zero). Homogeneous calibration improves the worse-off group rather than equalising by regression. The one place both groups worsen together is W3 balanced, consistent with the shakier W3 aggregate-WER control.

Direct contrasts between conditions. Testing each condition against FP16 leaves the key comparison implicit, so we also contrast the conditions directly as gap(A) minus gap(B) under the same paired bootstrap. At W4 backbone each homogeneous condition differs significantly from balanced. Native calibration narrows the gap by 0.0119 more than balanced (CI [–0.0199, –0.0042]) and L2 calibration by 0.0102 more than balanced (CI [–0.0182, –0.0030]), while native and L2 do not differ from each other (0.0015, CI crossing zero). What moves the gap is calibration homogeneity, not which population it is homogeneous in. At W3 and on the encoder no direct contrast clears zero, consistent with the effect weakening away from the W4 backbone. The direction is also unchanged under equal-weight-per-speaker (macro) weighting, where the FP16 macro gap (+0.0237) is about twice the corpus gap (+0.0110), so the per-speaker disparity is larger than the word-weighted headline suggests.

Robustness across calibration draws. Everything above rests on one calibration draw. We re-drew the 128-utterance set under two more seeds and re-ran GPTQ at W4 backbone. The narrowing direction holds: both homogeneous-vs-FP16 Δ gaps stay negative in all three draws (means –0.0077 for native, –0.0060 for L2). Its strength does not hold. Each is significant in only one draw of three, and the direct homogeneity-vs-balanced contrast, significant in two draws, reverses sign in the third. With only 23 native speakers the movement is about the size of the draw-to-draw wobble it rides on. We read this as directional consistency without draw-level robustness, which is the main reason we frame the paper as a case study.

At W3 the same direction holds (all three point estimates stay negative) but no condition’s Δ gap CI excludes zero, and the control is shakier (overall WER 0.339–0.347, balanced the worst). At W2 all three conditions fall apart: overall WER climbs past 1.0, meaning more edits than reference words, and the transcripts are incoherent, partly non-English output unrelated to the audio (a reference of ‘BYE’ decoded as multilingual noise). That uniform collapse has nothing to do with calibration composition, so we read nothing causal into its gap numbers.

5 Discussion

We frame this as a case study rather than a general causal claim. Within this one model, quantizer and recipe, calibration-set composition moves the accent-WER gap at matched overall WER, in the same direction across two different quantized components. What we do not claim is that the direction,

Target	Cond	Bits	Native WER	L2 WER	Δ Native vs. FP16 [95% CI]	Δ L2 vs. FP16 [95% CI]
–	FP16	16	0.3297	0.3407	–	–
Backbone	native	4	0.3316	0.3319	+0.0019 [−0.0056, +0.0091]	−0.0088 [−0.0159, −0.0020]
Backbone	balanced	4	0.3232	0.3354	−0.0065 [−0.0125, −0.0009]	−0.0053 [−0.0133, +0.0016]
Backbone	l2	4	0.3307	0.3327	+0.0010 [−0.0038, +0.0066]	−0.0080 [−0.0160, −0.0015]
Backbone	native	3	0.3357	0.3406	+0.0060 [−0.0021, +0.0152]	−0.0001 [−0.0075, +0.0060]
Backbone	balanced	3	0.3425	0.3505	+0.0128 [+0.0051, +0.0197]	+0.0098 [+0.0022, +0.0172]
Backbone	l2	3	0.3406	0.3409	+0.0109 [−0.0015, +0.0244]	+0.0002 [−0.0085, +0.0079]
Encoder	native	4	0.3334	0.3335	+0.0037 [−0.0029, +0.0102]	−0.0072 [−0.0144, −0.0016]
Encoder	balanced	4	0.3323	0.3376	+0.0026 [−0.0038, +0.0087]	−0.0031 [−0.0099, +0.0030]
Encoder	l2	4	0.3366	0.3409	+0.0069 [−0.0022, +0.0170]	+0.0002 [−0.0066, +0.0060]

Table 2: Group-wise WER by condition, with each group’s own paired speaker-clustered bootstrap change from FP16 (2,000 resamples). A gap that narrows via rising native WER (Δ Native > 0) is leveling-down. A gap that narrows via falling L2 WER (Δ L2 < 0) is a genuine L2 improvement. At W4, Δ L2 is negative with a CI excluding zero in three conditions while the paired Δ Native CI crosses zero in all of them.

or a clean composition-only attribution, generalises beyond this setting, or that we have identified the mechanism. Our best guess is that a homogeneous calibration set, whichever population it draws from, produces lower-variance per-layer activation statistics and therefore more consistent GPTQ rounding, while a mixed set splits the difference between two distributions and specialises in neither. We present that as a hypothesis to test, not something we have shown.

The multi-seed check sharpens this caution. Across three draws the narrowing direction is consistent, but its magnitude is on the order of the variation a fresh draw introduces. The test that most directly encodes “homogeneity beats balancing” reverses in one draw of three. The practical takeaway is not to calibrate on a single group. It is a caution in both directions, that calibrating on the majority is not a neutral choice and that balancing is not a dependable fix.

The most motivated next step is covariate-matched calibration sets that hold speaker count, duration and word budget fixed, separating composition from draw-to-draw wobble. A mechanistic probe of per-layer activation variance would test the lower-variance hypothesis. Extending past one model, quantizer and dataset remains the broader open question.

Limitations

Our evidence comes from a single setting, one model (Qwen2-Audio-7B-Instruct), one quantizer (GPTQ) and one calibration size of 128 examples. The narrowing direction is stable across three calibration draws, but individual-seed significance is not and the homogeneity-vs-balanced

contrast reverses in one draw, so the effect may not generalise. EdAcc’s 62 speakers are a convenience sample where native/L2 grouping may confound with recording channel, session and style, and with only 23 native speakers the per-condition CIs are wide, so headline claims use the paired Δ gap. Two design points remain. The calibration utterances are drawn from the same split used for evaluation, but the overlap is benign, since removing each condition’s 128 calibration utterances moves the W4 backbone gap by at most 0.0021. Speaker-level overlap remains by design, and a fully speaker-disjoint split is future work. Because only 23 native speakers exist, the conditions are also not matched on covariates beyond native/L2 composition (unique speakers 23/62/39, duration 733/930/911 s, words 2209/2438/2370), so covariate-matched calibration sets are the most important next step.

Ethics Statement

We study accent-based transcription disparity under quantization to surface risk, not to prescribe fixes. Balanced calibration is not a validated mitigation here.

References

- Red Hat AI and vLLM Project. 2024. [LLM Compressor](#).
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. [Qwen2-Audio technical report](#). *Preprint*, arXiv:2407.10759.
- Thomas Palmeira Ferraz. 2024. [Efficient compression](#)

of multitask multilingual speech models. *Preprint*, arXiv:2405.00966. Master’s thesis.

- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*.
- Federico Marcuzzi, Xuefei Ning, Roy Schwartz, and Iryna Gurevych. 2026. [How quantization shapes bias in large language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 363–404, Rabat, Morocco. Association for Computational Linguistics.
- Andrew Morris, Viktoria Maier, and Phil Green. 2004. [From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition](#). In *Proc. Interspeech*.
- Plawan Kumar Rath and Rahul Maliakkal. 2026. Quantization undoes alignment: Bias emergence in compressed LLMs across models and precision levels. In *2026 IEEE Cloud Summit*. IEEE.
- Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. [The Edinburgh international accents of English corpus: Towards the democratization of English ASR](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Miles Williams and Nikolaos Aletras. 2024. [On the impact of calibration data in post-training quantization and pruning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.

Pruning for Efficiency, Paying in Fairness: Demographic Disparities in Pruned Speech-LLMs

Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis, Ravi Shekhar

School of Computer Science and Electronic Engineering, University of Essex, UK

{karthik.kolluri, mkampo, r.shekhar}@essex.ac.uk

Abstract

Speech-LLMs are expensive to run, making compression important for real-world deployment. However, compressed models are usually selected using aggregate word error rate (WER), which can hide how pruning affects different demographic groups. In this work, we systematically study the effect of audio encoder pruning on SLAM-ASR for different demographic groups. Using the Fair-Speech and Common Voice datasets, we found that the pruning does not affect all demographic groups equally; the gap between best- and worst-performing groups increases in fold. These disparities appear across all three encoder scales, but only the largest model initially hides them behind aggregate WER. LoRA adaptation improves WER for every group, but benefits groups already performing well more strongly and widens for certain groups. On Common Voice English, Danish, and Dutch, accent gaps persist but do not clearly widen, showing that the fairness effects of pruning vary across datasets and must be measured directly. Our findings suggest that for pruned models, deployment decisions should include per-group WER, with the worst-performing group’s error rate as an explicit criterion.

1 Introduction

Automatic speech recognition (ASR) transcribes spoken language, and it is used in voice assistants, dictation tools, and captioning services (Koenecke et al., 2020). The field’s current dominant architecture is the Speech-LLM: a pretrained speech encoder coupled to a large language model through a lightweight projector. These stacks are generally large, making them computationally expensive, and model compression has become a routine step between training and deploying.

One of the widely used compression techniques is pruning. Usually, a pruned model is validated on aggregate word error rate (WER). However, ASR

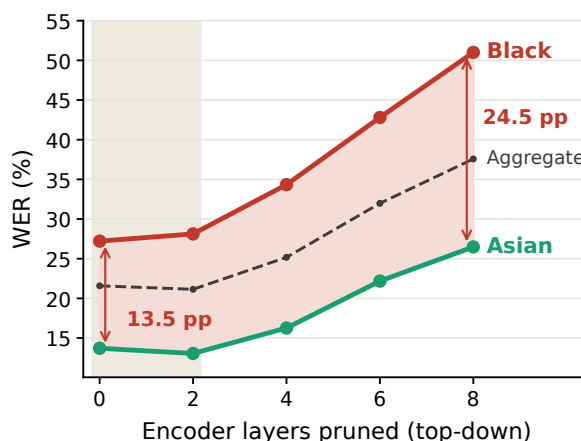


Figure 1: **Compression widens the racial gap, and the aggregate WER hides it.** WER for Black and Asian speakers (Fair-Speech, Whisper large) under top-down pruning; the wedge between the curves is their gap, which nearly doubles (13.5 → 24.5 pp). Within the grey band, aggregate WER *improves* (21.6 → 21.1%), yet Black speakers are the only group whose word error rate rises significantly (+0.9 pp, $p=.009$).

systems are already performing unevenly across demographic groups (Koenecke et al., 2020). That means an aggregated WER hides group disparities. In that case, when we apply pruning to those models, the question will be: what does pruning do to that unevenness? ¹

We study this effect systematically in the SLAM-ASR pipeline (Ma et al., 2024; Zhang et al., 2026a,b): a frozen Whisper encoder (Radford et al., 2022), a trainable projector, and a frozen Qwen2.5-3B decoder (Qwen et al., 2025) connected together as an ASR pipeline. We experiment on three Whisper scales (12, 24, and 32 encoder layers), prune each encoder top-down two layers at a time, and retrain the projector at every depth, so each configuration is a deployable model. This pipeline is demographically blind: no group label enters train-

¹Code is available at <https://github.com/KarthikKolluriKB/SpeechLLM-Pruning-Fairness>

ing, pruning, or LoRA adaptation. Our primary corpus is Fair-Speech (Veliche et al., 2024), which offers the richest demographic schema, and we add Common Voice English, Dutch, and Danish for cross-lingual study.

As shown in Fig. 1, in a Speech-LLM built on Whisper Large, removing the top two encoder layers improves aggregate WER from 21.6 to 21.1 ($p=.038$). The same setting raises WER for Black speakers by 0.9 points ($p=.009$). Here, the aggregate WER hides the subgroup degradation: the total improves while a group gets worse. Motivated by this, in this work we address three research questions:

- **RQ1:** Does encoder pruning amplify demographic disparities, and does aggregate WER hide it?
- **RQ2:** Does the effect depend on encoder scale and language resource level?
- **RQ3:** Does LoRA adaptation recover performance equally across groups?

Our results answer these questions as follows. Pruning amplifies the racial disparity on the largest model (RQ1). The gap between Black and Asian speakers grows from 13.5 to 24.5 percentage points (pp) by eight removed layers, even though the first prune improves aggregate WER. The disparity exists at every encoder scale: on all three models, the worst-performing group starts with roughly twice the WER of the best-performing. Concealment, however, appears only at the largest scale (RQ2). On the small and medium models, the first prune raises aggregate WER by 10.8 and 2.3 pp, respectively. So the damage is already visible in the aggregate. Across languages, the accent gaps on Common Voice English and Dutch persist under pruning but do not amplify. Finally, LoRA improves aggregate WER at every depth but widens the Black-to-Asian ratio at every matched depth (RQ3): it recovers accuracy for every group, but better for the groups the model already performs well on and worse for the groups it performs worse on.

Our findings carry a direct practical message: aggregate WER cannot validate a pruned speech model as fair. Deployment decisions for compressed speech models should therefore include per-group WER, with explicit attention to the worst-performing group. To our knowledge, this is the

first systematic study of how structural encoder pruning in Speech-LLM affects demographic disparity across pruning depth, model scale, and LoRA adaptation.

2 Related work

2.1 Bias in ASR

Performance disparities across speaker groups are well documented in ASR. Koenecke et al. (2020) found that five commercial systems transcribed speech from Black speakers with an average word error rate (WER) of 35%, against 19% for white speakers, and traced the gap to the underlying acoustic models, since it persisted on a matched subset of identical phrases. Subsequent work investigated the linguistic mechanisms behind this disparity, linking errors to morpho-syntactic features of African American English such as habitual “be” (Martin and Tang, 2020) and to ethnicity-related dialectal variation more broadly (Wassink et al., 2022). Disparities along gender and dialect lines were reported earlier (Tatman, 2017; Tatman and Kasten, 2017), and Feng et al. (2021) systematically quantified bias across gender, age, regional accent, and non-native speech; we adopt this per-group WER evaluation. These imbalances persist in large pretrained models: evaluations of Whisper (Radford et al., 2022) report higher accuracy for North American than for British or Australian accents, and for native than for non-native speech, with WER associated with speaker sex, first-language typology, and second-language proficiency (Graham and Roll, 2024). Performance gaps for child speakers likewise persist across Whisper scales (Attia et al., 2023). To measure such gaps systematically, corpora with speaker demographics have emerged: Common Voice collects optional self-reported accent, gender, and age at scale (Ardila et al., 2019), the Edinburgh corpus targets accent diversity (Sanabria et al., 2023), and Fair-Speech provides the richest schema of the three, covering age, gender, ethnicity, socio-economic status, and geographic variation (Veliche et al., 2024). We justify our choice among these in Section 3.

2.2 Model Compression, Scale, and Bias

How model compression affects bias is still an open question, with outcomes that vary by method, domain, and demographic axis. In computer vision, Hooker et al. (2020) demonstrated that pruning and quantization can amplify bias: minimal changes

in overall accuracy may hide substantial errors affecting a small subset of under-represented examples, an observation that supports evaluating worst-group rather than average performance (Sagawa et al., 2019). Tran et al. (2022) attributed such disparities to differences in gradient norms and decision-boundary distance across groups, while Iofinova et al. (2023) found that, although moderate pruning need not increase bias, disparities for specific protected attributes emerge at high sparsity.

In LLMs, compression’s effect on fairness depends on the base model and method (Ramesh et al., 2023), but two direct comparisons find pruning riskier than quantization, for protected groups (Xu et al., 2024) and for trustworthiness (Hong et al., 2024), and compression’s effect on social bias varies with model scale (Gonçalves and Strubell, 2023).

In speech, evidence on compression and bias is scarce. Lin et al. (2024b) examined self-supervised models under pruning and distillation, finding, in embedding-level probes, that distillation increased gender bias while decreasing age bias, whereas row pruning reduced bias. For speech-LLMs, bias studies either examine response content (Lin et al., 2024c,a) or benchmark WER disparities across architectures: Ginjala et al. (2026) evaluated nine models spanning CTC, encoder-decoder, and LLM-decoder designs and reported that the audio encoder, more than language model scale, governs fairness.

We build directly on Kolluri et al. (2026), who pruned Whisper encoder layers in the speech-LLM (Ma et al., 2024) and showed that removing a few layers costs little aggregate WER and that LoRA (Hu et al., 2021) recovers the unpruned baseline’s WER; that study reports aggregate WER only. How adaptation affects fairness is itself unexplored: Ding et al. (2024) found no consistent subgroup effect of LoRA relative to full fine-tuning; low-rank adaptation can propagate non-demographic biases inherited from noisy pretraining data (Chang et al., 2026); and fine-tuning aimed at under-represented speakers lowers their WER (Shekoufandeh et al., 2025), suggesting repair depends on where adaptation is aimed. No prior work examines whether encoder pruning shifts demographic disparity in a speech-LLM’s transcriptions, or whether LoRA’s recovery is shared evenly across groups; we address both.

3 Experimental Setup

3.1 System and Pruning

We follow the SLAM-ASR architecture (Ma et al., 2024): a Whisper encoder (Radford et al., 2022) feeds a small trainable projector (ConcatLinear), which maps audio representations into a frozen Qwen2.5-3B decoder (Qwen et al., 2025). We hold the LLM fixed so that encoder capacity is the only varied component; both the encoder and the LLM stay frozen, and only the projector and LoRA adapters are trained. To treat encoder capacity as an experimental variable, we run the full procedure with three Whisper variants, Small (12 layers), Medium (24), and Large-v2 (32), each paired with the same projector and LLM. As in Kolluri et al. (2026), we prune each variant top-down, removing two layers at a time, and report pruning as the absolute number of layers removed ($L-2$, $L-4$, $\dots$). The projector is retrained at every depth, so each configuration is a fully retrained, deployable system rather than a fixed model probed after pruning.

Procedure. Concretely, the full protocol at each encoder scale (small, medium, and large-v2) proceeds as follows.

1. Initialise the system with the unpruned encoder and train the projector, recording the result as the $L-0$ baseline.
2. Remove the top two encoder layers to obtain depth $L-k$, for $k \in \{2, 4, 6, 8\}$.
3. Retrain the projector from scratch at that depth, so that each configuration constitutes a deployable system rather than a mismatched stack probed after pruning.
4. Select the configuration on aggregate WER alone; no demographic label enters training, pruning, or selection.
5. Decode the evaluation corpora and compute WER separately for each demographic group meeting the analysability threshold of Section 3.4.

The models never see demographic labels: training and model selection use aggregate WER alone. In practice, this is also how pruned models are judged, since compression work reports aggregate WER and stops there. Our question is what this number hides: does pruning damage the WER of individual demographic groups even when the aggregate looks fine? To answer it, we take each

pruned model after training and evaluate it separately for every demographic group on the evaluation corpora.

3.2 Datasets

Table 1 summarizes the corpora used for training and evaluation.

Training data: The projector and LoRA adapters are trained on Common Voice 22 (Ardila et al., 2019), a crowdsourced multilingual read-speech corpus, following the same training configuration as Kolluri et al. (2026). We use the official train and test splits without modification, so speakers are disjoint across splits by construction, and the counts reported in Table 1 are those remaining after the duration filter described below. Our primary system uses the English subset (100 hours); for the cross-lingual analysis, we additionally train separate systems on the Danish (4.2 hours) and Dutch (54 hours) subsets, which span an order-of-magnitude range in training data.

Evaluation data and demographic axes: We aim to evaluate as many demographic axes as possible. To this end, we selected two datasets: Fair-Speech (Veliche et al., 2024) and Common Voice 22 (Ardila et al., 2019). Fair-Speech (Veliche et al., 2024) is a comprehensive evaluation benchmark built specifically to expose demographic disparities in speech recognition, and it carries the richest bias annotation of our corpora. Across 26,417 utterances, it labels *ethnicity* (seven analyzable groups), *socio-economic status*, gender, age, and first language. Fair-Speech is used exclusively for evaluation and contributes no training data. Common Voice 22 (Ardila et al., 2019) (CV-22) is a large multilingual crowdsourced corpus of read speech in which contributors may optionally self-report speaker attributes. We use this metadata to measure disparities by *accent*, *gender*, and *age* on the English, Danish, and Dutch splits, which allowed us to compare results across languages. Because these attributes are optional, group-level results on CV-22 are computed over the subset of test utterances carrying the relevant annotation, which is smaller than the full test split. The dataset statistics and supported demographic axes are presented in Table 1 and Table 2 respectively.

Preprocessing. Audio is resampled to 16 kHz and converted to 80-channel log-Mel spectrograms following the Whisper pipeline. We retain utterances between 0.5 and 30 seconds. Transcriptions are lowercased with punctuation removed, preserv-

Corpus	Split	Samples
<i>Training (Common Voice 22)</i>		
English (EN)	Train	58,140
Dutch (NL)	Train	43,458
Danish (DA)	Train	3,592
<i>Evaluation</i>		
Common Voice (EN)	Test	16,391
Common Voice (NL)	Test	12,033
Common Voice (DA)	Test	2,684
Fair-Speech	eval	26,417

Table 1: Dataset statistics. Common Voice 22 provides the training and test splits; Fair-Speech is used for evaluation only.

Corpus	Bias axes
CV-22 (EN/DA/NL)	accent, gender, age
Fair-Speech	ethnicity, SES, gender, age, L1

Table 2: Evaluation corpora and the demographic axes each supports.

ing apostrophes for contractions.

3.3 Implementation Details

Model Architecture. Our system follows SLAM-ASR (Ma et al., 2024) and comprises three components: a Whisper speech encoder (Radford et al., 2022), a lightweight ConcatLinear projector, and a Qwen2.5-3B LLM (Qwen et al., 2025). We run it with three Whisper encoder variants, Small (12 layers), Medium (24), and Large-v2 (32), each paired with the same projector and LLM.

Projector. The projector is a two-layer MLP with concatenation-based temporal downsampling: it concatenates five consecutive frames, reducing the sequence length five times, then applies two linear layers with ReLU and dropout 0.1 (hidden dimension 2048), followed by LayerNorm to match the scale of the LLM text embeddings. The projector is reinitialized and retrained from scratch at every pruning depth rather than fine-tuned from the unpruned checkpoint.

LoRA. We apply LoRA (Hu et al., 2021) to the query, key, value, and output projections of all LLM attention layers. To limit overfitting on smaller datasets, we set the rank per resource level: $r=8$, $\alpha=16$ for Danish (low-resource) and $r=16$, $\alpha=32$ for Dutch and English; $r=16$ on Danish overfit in preliminary runs, motivating the reduced

rank. LoRA modules use dropout 0.1, adding roughly 0.8M ($r=8$) and 1.5M ($r=16$) trainable parameters.

Training Details. All models are trained with AdamW (Loshchilov and Hutter, 2019) at learning rate 1×10^{-4} , weight decay 0.01, and gradient clipping 1.0, under a cosine schedule with linear warmup over the first 5% of steps, in bfloat16 on a single NVIDIA A6000 (48 GB). The Whisper encoder and Qwen2.5-3B weights remain frozen; only the projector and, where enabled, the LoRA adapters are updated. For English configurations, we use a batch size of 8 and train for 2 epochs, or 4 with LoRA. Dutch uses a batch size of 8 and Danish a batch size of 4, with more passes over the smaller corpora; per-configuration epoch counts are given in Appendix A. Training stops early on validation WER, so model selection uses aggregate validation performance alone, with no demographic label involved. Every pruning depth within a setting uses the same setup, so depths differ only in encoder size. All runs use a single seed (42), so we rely on the paired bootstrap in Section 3.4 rather than variance across seeds.

Decoding. At inference, the projected audio embeddings are prepended to a fixed plain-text prompt, Transcribe speech to text., applied without a chat template and held constant across all scales, depths, and languages. Transcriptions are generated with beam search (width 2, no sampling, no repetition or length penalty) up to 128 new tokens, identically for base and LoRA-adapted systems. Checkpoint selection during training uses greedy decoding on the validation split with a repetition penalty of 1.5.

Baselines. The unpruned ($L-0$) system at each encoder scale is the reference for every pruned configuration at that scale, so configurations differ only in encoder depth and in the projector retrained at that depth. LoRA comparisons are matched the same way, against the base system at the same scale and depth. We report no external ASR baseline, since our claim concerns how a fixed pipeline behaves under compression rather than how it compares against other systems.

3.4 Evaluation Details

We evaluate every configuration with WER: aggregate WER is computed over the full evaluation set, and a group’s WER over the utterances spoken by members of that group, so every group is scored by the same rule.

Group	n	L-0	L-2	L-4	L-6	L-8
Asian	3,854	13.7	13.1	16.3	22.2	26.5
Native Haw.	969	13.4	13.7	17.8	24.7	31.4
Hispanic	2,811	20.0	19.2	22.5	27.5	33.9
White	5,619	20.6	20.1	22.7	28.4	31.7
Native Am.	4,616	21.5	19.0	22.1	29.0	33.5
MENA	749	23.0	22.3	26.0	34.2	37.3
Black	7,799	27.2	28.1	34.3	42.8	51.0
Aggregate	26,417	21.6	21.1	25.2	32.0	37.6

Table 3: Fair-Speech, ethnicity (Whisper large); n is the number of utterances per group. Black speakers have the highest WER at every depth. At L-2 the aggregate improves while Black speakers degrade, and by L-8 the Black–Asian gap has nearly doubled, from 13.5 to 24.5 pp.

We assess disparity with three measures, fixed before the analysis: the worst-group WER at each depth as our primary measure, the absolute gap $\Delta = \text{WER}_{\text{worst}} - \text{WER}_{\text{best}}$ in percentage points (pp), and the disparity ratio $\rho = \text{WER}_{\text{worst}}/\text{WER}_{\text{best}}$ as a scale-free check; within a group we report relative degradation $\text{WER}_{Lx}/\text{WER}_{L0}$. Since Δ and ρ diverge by construction when all groups degrade together, we report both at every depth, tracking the Black–Asian pair throughout as the worst- and best-served groups. Significance comes from a paired resampling test: we draw 2,000 bootstrap resamples of the utterances, re-score both setups on each, and take p as the fraction in which the WER difference flips direction. We limit our claims to the usable range, the contiguous depths at which aggregate WER stays at or below 40%, and analyze groups only where they have at least 200 utterances and 30 minutes of audio; smaller groups appear in tables for coverage only.

4 Results and Discussions

4.1 Pruning and demographic disparity (RQ1)

We present our results by demographic dimensions, starting with ethnicity. Table 3 presents WER for the seven ethnic groups as we prune Whisper large encoder layers: at L-0 (without pruning), the groups are already far apart, from 13.4% for Native Hawaiian and 13.7% for Asian speakers to 27.2% for Black speakers, twice the Asian rate. The bias already exists before any pruning.

As we prune the top two encoder layers, the aggregate WER shows that the model performs slightly better: it falls from 21.6% to 21.1%

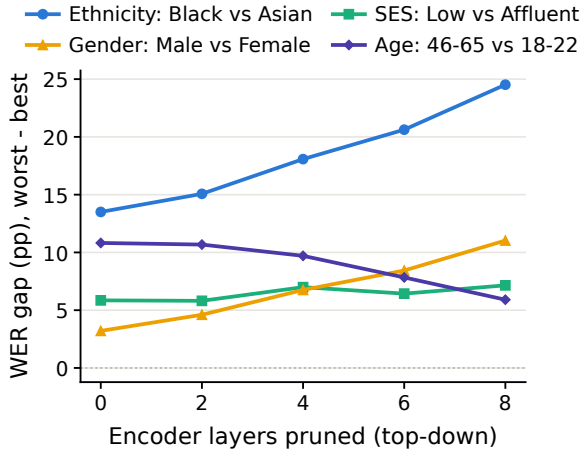


Figure 2: Worst-best WER gap per demographic axis under top-down pruning (Fair-Speech, Whisper large); one line per axis, with the pair named in the legend. The ethnicity and gender gaps widen, the Socioeconomic gap holds, and the age gap narrows by leveling down.

($p=.038$, Table 3). The group columns tell a different story: Black speakers are the only group whose error rate rises among other groups (+0.9 pp, $p=.009$), Native American speakers improve most (−2.5 pp, $p=.001$), and no other group changes significantly ($p \geq .13$). In relative terms, L-2 is closer to baseline on every group (aggregate at 0.98) except the Black group, at 1.03 (Appendix B). A practitioner relying on aggregate WER alone would see a small win, while the group that started worst has become worse.

At deeper prunes, all groups degrade but at different rates. At L-4 the aggregate has risen to 25.2%, a cost any practitioner would notice, and the groups have begun to separate: Asian speakers are at 16.3% while Black speakers are at 34.3%, a gap of 18.0 pp. At L-6 the aggregate is at 32.0%, within the 40% usability bound, while Black speakers are at 42.8% and Asian speakers are at 22.2%. The same configuration is therefore usable by the aggregate criterion and unusable for its worst-performing group. At L-8, Black speakers are at 51.0% and Asian speakers at 26.5%. Over the full sweep, Black speakers lose 23.8 pp and Asian speakers 12.8 pp, so the gap between them widens from 13.5 pp to 24.5 pp. Aggregate WER is misleading in two ways: it reports an improvement at the first prune (L-2) while the worst-performing group degrades, and it stays within the usability bound at depths where the worst-performing group has already exceeded it.

Ethnicity is not the only axis that moves. Fig. 2

plots the worst-to-best gap for each demographic axis at every depth (per-group values in Table 15), and the three remaining axes each do something different. The gender gap grows from 3.3 to 11.1 pp (male speakers disadvantaged in this corpus; Fig. 5). The age gap narrows, but only because the oldest group, worst at baseline, degrades slowest (Fig. 7). The socioeconomic gap holds steady (Fig. 6; Table 15): Low-SES speakers start 5.9 pp behind Affluent speakers (ratio 1.36) and stay behind at every depth, and although the raw gap reaches 7.2 pp at L-8, Affluent speakers lose more of their starting accuracy (+83% vs +67%), so the ratio falls to 1.24. Pruning carries the class disparity along unchanged.

Taken together, these results answer RQ1. The disparity is inherited, but the hiding is new: at L-2 the aggregate improves while the worst-performing ethnicity group degrades significantly. Gender amplifies, class persists, age narrows by leveling down; aggregate WER is blind to all three, which is why pruned models need per-group evaluation.

4.2 The role of encoder scale (RQ2)

Fig. 3 shows per-group WER at all three encoder scales. Black speakers have the highest WER in every panel at every depth. What differs across panels is the cost of the first prune. The small model degrades immediately, its aggregate rising by 10.8 pp at two removed layers, and medium by 2.3 pp; large-v2 instead improves by 0.5 pp.

The disparity ratio ρ (Table 4) separates the three scales in a way the aggregate does not. Before pruning, they are similarly unequal, with ρ at 1.99 on large-v2, 2.03 on small, and 2.16 on medium, so Black speakers face roughly twice the word error rate of Asian speakers on all the encoder scales. The first prune is what shows them going apart. On small and medium, the ratio falls, from 2.03 to 1.70 and from 2.16 to 2.01, because these models degrade so sharply that the better-served group loses more in proportional terms. On large-v2 it rises instead, from 1.99 to 2.15, and large-v2 is the one configuration whose aggregate WER improved.

Beyond the first pruning, ρ falls on every scale, reaching 1.92 on large-v2 by L-8, while the absolute gap widens at every depth, from 13.5 pp unpruned to 15.0, 18.0, 20.6, and 24.5 pp at L-8 (Table 3). This is the divergence anticipated in Section 3.4: deep pruning drives every group toward high error rates, and the quotient of two large numbers is smaller than that of two small ones even

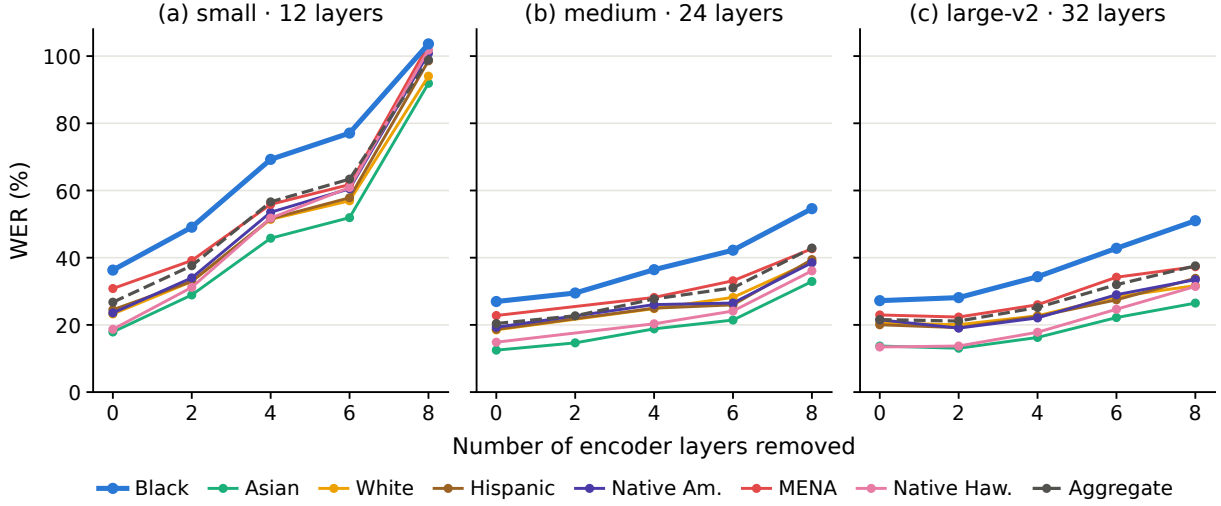


Figure 3: Per-group WER on Fair-Speech as encoder layers are removed, per Whisper scale (shared y -axis; dashed = aggregate). Black speakers have the highest error rate at every scale and every depth. Only at large-v2 does the shallow prune leave the aggregate unchanged. Appendix Table 11 covers that depth.

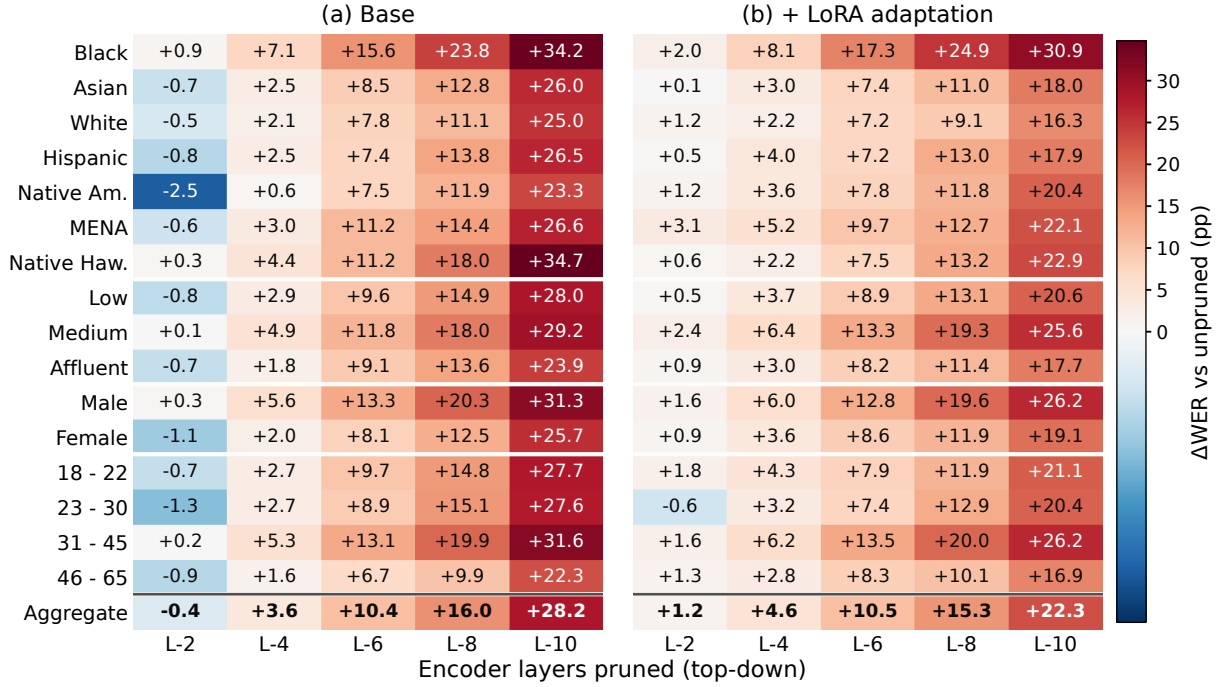


Figure 4: Change in WER (percentage points) per subgroup as Whisper large-v2 is pruned top-down, without (a) and with (b) LoRA (Fair-Speech). Warm colors indicate degradation, cool colors indicate improvement, and the bottom row is the aggregate. Black speakers degrade most at nearly every depth in both conditions. LoRA lowers the aggregate everywhere, but it improves only the already better-performing groups: at ten removed layers it recovers 8.0 pp for Asian speakers and only 3.3 pp for Black speakers.

when the difference between them has grown. Our primary measure resolves it, since at L-8 Black speakers face 51.0% WER against 26.5% for Asian speakers. Every scale inherits the disparity, then, but only the largest conceals what pruning does to it.

4.3 The effect of LoRA adaptation (RQ3)

Fig. 4 summarizes the LoRA comparison: each cell is a subgroup’s change in WER relative to the unpruned model (warm colors mean degradation), with the base model in panel (a), the LoRA adapted model in panel (b), and the aggregate in the bottom row; the underlying per-group WERs are in Ap-

Model	L-0	L-2	L-4	L-6	L-8
small (12L)	2.03	1.70	<i>1.51</i>	<i>1.49</i>	<i>1.13</i>
medium (24L)	2.16	2.01	1.94	1.97	<i>1.66</i>
large-v2 (32L)	1.99	2.15	2.10	1.93	1.92

Table 4: Fair-Speech, disparity ratio ρ between Black and Asian speakers, per encoder scale. Italic values lie beyond the usable range (aggregate WER > 40%). Only large-v2 shows a rise at the first prune, where its aggregate WER also improves.

pendix Table 16. The LoRA-adapted model is better everywhere: with the unpruned encoder, aggregate WER drops from 21.6% to 17.7%, and at eight removed layers from 37.6% to 33.0%. Adaptation also extends the usable pruning range through ten removed layers, two more than the base model.

The disparity, however, does not shrink with the error rate. At every depth, the Black-to-Asian ratio is wider with LoRA than without LoRA: at L-0 it rises from 1.99 (base) to 2.20 (LoRA), at L-2 from 2.15 to 2.36, and at L-8 from 1.93 to 2.22. The reason is visible in panel (b): LoRA adaptation compensates the best-performing groups better. At ten removed layers, it recovers 8.0 pp of Asian speakers’ degradation but only 3.3 pp of Black speakers’.

4.4 Beyond Fair-Speech: Common Voice English, Dutch, and Danish

Group	n	L-0	L-2	L-4	L-6	L-8
US	1,145	11.1	12.0	13.9	16.7	17.8
England	358	12.2	13.2	15.1	17.1	18.9
India/S-Asia	508	15.8	17.8	19.1	21.0	24.8
Aggregate	3,008	12.1	13.0	14.7	17.0	19.1

Table 5: Common Voice 22 English, accent (Whisper large); n is the number of utterances per group. India/South Asia has the highest WER at every depth. Unlike Fair-Speech, the aggregate degrades from the first prune.

Group	n	L-0	L-2	L-4	L-6	L-8
Netherlands	3,817	11.5	12.9	15.0	18.1	20.4
Belgian	1,146	13.7	15.2	17.8	20.5	25.2
Aggregate	5,306	11.8	13.3	15.5	18.5	21.4

Table 6: Common Voice Dutch, accent (Whisper large); n is the number of utterances per group. Belgian speakers have higher WER than Netherlands speakers at every depth.

Group	n	L-0	L-2	L-4	L-6	L-8
Male	3,372	13.5	15.3	17.7	18.5	20.6
Female	1,163	10.4	11.1	13.3	14.8	15.9
Aggregate	4,535	12.7	14.2	15.5	16.5	19.4

Table 7: Common Voice Dutch, gender (Whisper large); n is the number of utterances per group. Male speakers have higher WER at every depth.

Axis	Group	n	L-0	L-2	L-4	L-6	L-8
Gender	Male	1,031	36.0	38.3	40.6	49.0	51.8
	Female [†]	299	33.0	34.1	35.2	42.9	48.0
	Aggregate	1,330	35.3	37.4	39.4	47.6	50.9
Age	Teens [†]	50	40.3	42.0	43.6	49.9	53.6
	Twenties	601	41.1	43.2	45.2	53.7	56.3
	Thirties	392	31.5	34.2	36.8	45.0	48.3
	Forties [†]	150	26.3	27.2	28.0	34.1	36.7
	Fifties [†]	98	29.2	31.8	34.4	44.5	53.2
	Sixties [†]	76	36.2	36.4	36.6	44.8	48.7
	Aggregate	1,367	36.5	38.7	40.9	49.2	52.5

Table 8: Common Voice Danish (Whisper large); n is the number of utterances per group. Groups marked [†] fall below the analysability threshold and are reported for coverage only.

Everything discussed so far comes from Fair-Speech. Fair-Speech provides a comprehensive suite for evaluating multiple demographic disparities. However, it is limited only to the English language. To move beyond English, we also used the demographic labels provided in the CV-22 in different languages. Specifically, we used English, Danish, and Dutch data as in Kolluri et al. (2026). This also allows us to determine whether the observed effects generalise across datasets and languages. The results reveal a more nuanced pattern: demographic disparities generally persist under pruning, but the concealed amplification observed on Fair-Speech does not consistently transfer to CV-22.

CV-22 English provides the clearest contrast (Table 5). Speakers with Indian and South Asian accents have the highest WER at every pruning depth, and their absolute disadvantage relative to US speakers grows as more layers are removed. However, the relative disparity remains stable: the WER ratio changes only from 1.42 at L-0 to 1.39 at L-8. Moreover, aggregate WER deteriorates from the first pruning step. Thus, pruning reduces overall performance, but it neither disproportionately amplifies the accent gap nor conceals the degradation from a practitioner monitoring aggregate WER.

CV-22 Dutch exhibits the same persistence without amplification (Tables 6 and 7; Fig. 8). Belgian speakers consistently have higher WER than speakers from the Netherlands, but the ratio remains between 1.1 and 1.3 across pruning depths. The gender axis follows a similar pattern: male speakers have higher WER at every depth, while the ratio remains close to 1.3. These disparities therefore survive pruning but do not systematically widen.

CV-22 Danish marks the limit of the analysis (Table 8). Its unpruned aggregate WER is already 35.5%, and the 40% usability threshold is crossed after removing four layers. Demographic annotations are also sparse, with most groups—including all female speakers—falling below the analysability threshold. We therefore treat the Danish results as coverage information rather than evidence for group-level effects.

Together, these findings answer RQ2. Pruning did not reduce an existing disparity in any evaluated corpus. However, persistent disparities did not necessarily become amplified: on CV-22 English and Dutch, *relative* gaps remained broadly stable, while aggregate degradation was immediately visible. Concealed disparity amplification occurred only on Fair-Speech with the largest encoder. Evaluating on Fair-Speech does introduce a distribution shift from the Common Voice training data, which may penalize underrepresented linguistic features; however, *absolute* disparities widened under compression in both our in-domain and out-of-domain settings, so this widening cannot be attributed to corpus shift alone. Consequently, the presence of a baseline performance gap cannot predict how compression will affect a group; the effect must be evaluated separately for each corpus, demographic axis, and model scale.

5 Conclusion

Demographic bias (Veliche et al., 2024) and layer pruning (Kolluri et al., 2026) have been studied extensively, but their interaction remains underexplored. In this work, we investigate how pruning affects demographic disparities in ASR. On Fair-Speech, removing two layers from Whisper Large slightly improves aggregate WER, while WER for Black speakers increases. The Black-to-Asian WER ratio rises from 1.99 to 2.15, and after removing eight layers, the absolute gap widens from 13.5 to 24.5 pp. Pruning also increases the low-SES vs affluent gap from 5.9 to 7.2 points and more

than triples the gender gap. Thus, aggregate improvements can conceal disproportionate degradation among already disadvantaged groups. LoRA improves aggregate WER at every depth but similarly widens the Black-to-Asian ratio by benefiting better-performing groups more. On CV-22 English, Danish, and Dutch, demographic gaps persist but do not clearly amplify, while sparse annotations and high WER prevent reliable group-level conclusions for Danish.

These findings motivate fairness-aware compression objectives that prioritise worst-group rather than average performance. Future work should examine compression methods that not only focus on improving overall performance but also give equal importance to all demographics.

Limitations

This study investigates a single configuration: Whisper is the only speech encoder with three variants, Qwen2.5-3B the only LLM backbone, and top-down layer pruning the only compression technique. All configurations are trained with a single random seed (42). Since the projector is retrained at every pruning depth, seed replication would multiply across the full grid of encoder scales, depths, and languages, which was beyond our compute budget. To mitigate this, we report paired bootstrap significance over the evaluation set and ground our claims in trends that hold consistently across the sweep rather than in differences within any single configuration. Multi-seed replication remains an important direction for future work. The dataset scope is similarly narrow. The main analysis rests on Fair-Speech, an English read-speech corpus with self-reported demographic labels, and the cross-lingual evidence is thin: pruned Danish models were too degraded for group comparison, and the Dutch corpus records only accent, gender, and age. The demographic axes share speakers and are analyzed separately, leaving intersectional effects unexamined. We evaluate no mitigation techniques.

Acknowledgments

This work was supported by a Knowledge Transfer Partnership (KTP) project (project number 10131983) funded by UKRI through Innovate UK, in collaboration with Hivedome. RS was supported by the ELOQUENCE project (grant number 101135916) funded by the UKRI and the Euro-

pean Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the UKRI, European Union, or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

References

- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. [Common voice: A massively-multilingual speech corpus](#). *ArXiv*, abs/1912.06670.
- Ahmed Adel Attia, Jing Liu, Wei Ai, Dorottya Demszky, and Carol Y. Espy-Wilson. 2023. [Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults](#). *ArXiv*, abs/2309.07927.
- Yupeng Chang, Yi Chang, and Yuan Wu. 2026. [Bailora: Bias-alleviating low-rank adaptation to mitigate catastrophic inheritance in large language models](#). In *International Conference on Learning Representations*, volume 2026, pages 156585–156612.
- Zhoujie Ding, Ken Ziyu Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. 2024. [On fairness of low-rank adaptation of large models](#). *ArXiv*, abs/2405.17512.
- Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. [Quantifying bias in automatic speech recognition](#). *ArXiv*, abs/2103.15122.
- Srishti Ginjala, Eric Fosler-Lussier, Christopher W Myers, and Srinivasan Parthasarathy. 2026. [Do llm decoders listen fairly? benchmarking how language model priors shape bias in speech recognition](#). *arXiv preprint arXiv:2604.21276*.
- Gustavo Gonçalves and Emma Strubell. 2023. [Understanding the effect of model compression on social bias in large language models](#). *ArXiv*, abs/2312.05662.
- Calbert Graham and Nathan Roll. 2024. [Evaluating openai’s whisper asr: Performance analysis across diverse accents and speaker traits](#). *JASA express letters*, 4 2.
- Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian R. Bartoldson, Ajay Kumar Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. 2024. [Decoding compressed trust: Scrutinizing the trustworthiness of efficient LLMs under compression](#). In *Proceedings of the 41st International Conference on Machine Learning*.
- Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily L. Denton. 2020. [Characterising bias in compressed models](#). *ArXiv*, abs/2010.03058.
- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *ArXiv*, abs/2106.09685.
- Eugenia Iofinova, Alexandra Peste, and Dan Alistarh. 2023. [Bias in pruned vision models: In-depth analysis and countermeasures](#). *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24364–24373.
- Allison Koenecke, Andrew Joo Hun Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Touns, John Russell Rickford, Dan Jurafsky, and Sharad Goel. 2020. [Racial disparities in automated speech recognition](#). *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689.
- Ganesh Pavan Kartikeya Bharadwaj Kolluri, Michael Kampouridis, and Ravi Shekhar. 2026. [On the role of encoder depth: Pruning whisper and LoRA fine-tuning in SLAM-ASR](#). In *Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAK-ABLE) @ LREC 2026*, pages 183–193, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).
- Yi-Cheng Lin, Wei-Chih Chen, and Hung yi Lee. 2024a. [Spoken stereoset: on evaluating social bias toward speaker in speech large language models](#). *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 871–878.
- Yi-Cheng Lin, Tzu-Quan Lin, Hsi-Che Lin, Andy T. Liu, and Hung yi Lee. 2024b. [On the social bias of speech self-supervised models](#). *ArXiv*, abs/2406.04997.
- Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung yi Lee. 2024c. [Listen and speak fairly: a study on semantic gender bias in speech integrated large language models](#). *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 439–446.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled Weight Decay Regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.
- Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. [An Embarrassingly Simple Approach for LLM with Strong ASR Capacity](#). *arXiv preprint arXiv:2402.08846*.
- Joshua L. Martin and Kevin Tang. 2020. [Understanding racial disparities in automatic speech recognition: The case of habitual "be"](#). In *Interspeech*.

- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 Technical Report](#). *arXiv preprint arXiv:2412.15115*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). In *International Conference on Machine Learning*.
- Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. [Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization](#). *ArXiv*, abs/1911.08731.
- Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. [The edinburgh international accents of english corpus: Towards the democratization of english asr](#). *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Golshid Shekoufandeh, Paul Boersma, and Antal van den Bosch. 2025. [Improving the inclusivity of dutch speech recognition by fine-tuning whisper on the jasmin-cgn corpus](#). *ArXiv*, abs/2502.17284.
- Rachael Tatman. 2017. [Gender and dialect bias in youtube’s automatic captions](#). In *EthNLP@EACL*.
- Rachael Tatman and Conner Kasten. 2017. [Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions](#). In *Inter-speech*.
- Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. 2022. [Pruning has a disparate impact on model accuracy](#). *ArXiv*, abs/2205.13574.
- Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. [Towards measuring fairness in speech recognition: Fair-speech dataset](#). *ArXiv*, abs/2408.12734.
- Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. [Uneven success: automatic speech recognition and ethnicity-related dialects](#). *Speech Commun.*, 140:50–70.
- Zhichao Xu, Ashim Gupta, Tao Li, Oliver Bentham, and Vivek Srikumar. 2024. [Beyond perplexity: Multi-dimensional safety evaluation of LLM compression](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15359–15396.
- Yuchen Zhang, Haralambos Mouratidis, and Ravi Shekhar. 2026a. [Speak in context: Multilingual asr with speech–context alignment via contrastive learning](#). In *Proceedings of the Fifteenth biennial Language Resources and Evaluation Conference (LREC 2026)*.
- Yuchen Zhang, Ravi Shekhar, and Haralambos Mouratidis. 2026b. [Language family matters: Evaluating llm-based asr across linguistic boundaries](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*.

A Appendix: Training Details

Table 9 reports the batch size and epoch count for each language. All other optimisation settings are identical across runs and are given in Section 3.3.

Epoch counts scale inversely with corpus size: English (100 hours) converges within two passes, while Dutch (54 hours) and Danish (4.2 hours) require more passes to converge on their smaller corpora. LoRA configurations are trained for roughly twice as many epochs as projector-only runs, since the adapters are optimised jointly with a projector that is itself trained from scratch. Batch size is reduced to 4 for Danish, whose corpus is too small to fill larger batches without excessive repetition within an epoch. All settings are held constant across encoder scales and pruning depths, so configurations differ only in encoder capacity.

Language	Batch	Proj.	LoRA
English	8	2	4
Dutch	8	4	8
Danish	4	6	8

Table 9: Batch size and epoch counts per language. Settings are identical across all three encoder scales and all pruning depths.

B Appendix: Relative-to-baseline WER on Fair-Speech

Tables 10 and 11 report WER at each pruning depth relative to the unpruned model (WER_{Lx}/WER_{L0}): per ethnic group on Whisper large, and across model scales for the aggregate and the two headline groups. Ratios below 1.00 mean improvement over the unpruned model.

Group	n	L-0	L-2	L-4	L-6	L-8
Asian	3,854	1.00	0.95	1.19	1.62	1.93
Native Haw.	969	1.00	1.02	1.33	1.84	2.34
Hispanic	2,811	1.00	0.96	1.12	1.37	1.69
White	5,619	1.00	0.98	1.10	1.38	1.54
Native Am.	4,616	1.00	0.88	1.03	1.35	1.56
MENA	749	1.00	0.97	1.13	1.49	1.63
Black	7,799	1.00	1.03	1.26	1.57	1.87
Aggregate	26,417	1.00	0.98	1.17	1.48	1.74

Table 10: Fair-Speech, WER relative to each group’s own unpruned baseline (Whisper large); n is the number of utterances per group. Values below 1.00 indicate improvement. At L-2 the aggregate is at 0.98 while Black speakers are at 1.03, the highest of any group.

Model	Series	n	L-0	L-2	L-4	L-6	L-8
large-v2 (32L)	Asian	3,854	1.00	0.95	1.19	1.62	1.93
	Black	7,799	1.00	1.03	1.26	1.57	1.87
	Aggregate	26,417	1.00	0.98	1.17	1.48	1.74
medium (24L)	Asian	3,854	1.00	1.17	1.51	1.72	2.64
	Black	7,799	1.00	1.09	1.35	1.57	2.03
	Aggregate	26,417	1.00	1.11	1.36	1.52	2.10
small (12L)	Asian	3,854	1.00	1.62	2.56	2.90	5.14
	Black	7,799	1.00	1.35	1.91	2.12	2.85
	Aggregate	26,417	1.00	1.41	2.12	2.37	3.69

Table 11: Fair-Speech, WER relative to each model’s own unpruned baseline; n is the number of utterances. Italic values lie beyond the usable range (aggregate WER > 40%). Only large-v2 has a depth where the aggregate improves while Black speakers degrade.

Group	n	L-0	L-2	L-4	L-6	L-8
Asian	3,854	12.5	14.7	18.8	21.4	32.9
Native Haw.	969	14.8	17.6	20.3	24.1	36.1
White	5,619	18.5	21.7	24.9	28.2	38.9
Hispanic	2,811	18.7	21.9	25.0	26.0	39.4
Native Am.	4,616	19.3	22.7	26.0	26.5	38.5
MENA	749	22.8	25.5	28.2	33.1	42.7
Black	7,799	27.0	29.5	36.4	42.2	54.6
Aggregate	26,417	20.4	22.7	27.7	31.0	42.9

Table 12: Fair-Speech, ethnicity (Whisper medium); n is the number of utterances per group. Italic values lie beyond the usable range (aggregate WER > 40%).

C Appendix: Per-group WER at small and medium scale

Tables 13 and 12 give the per-group WER behind Fig. 3, in the same form as Table 3 for large-v2. Black speakers have the highest WER at every scale and depth.

D Appendix: Fair-Speech coverage, SES, gender, and age

Table 14 reports the number of utterances and minutes of audio behind every Fair-Speech group; Table 15 reports WER for the remaining demographic axes, summarized in Fig. 2; Figs. 5 and 7 show the gender and age axes in detail.

E Appendix: Per-group WER under LoRA adaptation

Table 16 reports per-group WER under LoRA adaptation (Fair-Speech, Whisper large), complementing Fig. 4; compare with the base model in Table 3. Per-utterance outputs were not retained for the LoRA runs; SD here is estimated from stored

Group	<i>n</i>	L-0	L-2	L-4	L-6	L-8
Asian	3,854	17.9	28.9	<i>45.8</i>	<i>51.9</i>	<i>91.9</i>
Native Haw.	969	18.7	31.2	<i>51.8</i>	<i>60.9</i>	<i>101.8</i>
White	5,619	23.3	32.9	<i>51.4</i>	<i>56.9</i>	<i>94.0</i>
Native Am.	4,616	23.6	34.0	<i>53.5</i>	<i>60.5</i>	<i>100.8</i>
Hispanic	2,811	24.5	33.0	<i>51.6</i>	<i>57.8</i>	<i>98.6</i>
MENA	749	30.8	39.2	<i>55.9</i>	<i>61.8</i>	<i>103.4</i>
Black	7,799	36.3	49.1	<i>69.3</i>	<i>77.1</i>	<i>103.6</i>
Aggregate	26,417	26.8	37.6	<i>56.6</i>	<i>63.4</i>	<i>98.8</i>

Table 13: Fair-Speech, ethnicity (Whisper small); *n* is the number of utterances per group. Italic values lie beyond the usable range (aggregate WER > 40%).

Axis	Group	<i>n</i>	Minutes
Ethnicity	Black	7,799	1,010
	White	5,619	702
	Native Am.	4,616	556
	Asian	3,854	462
	Hispanic	2,811	372
	Native Haw.	969	94
	MENA	749	86
SES	Low	14,741	1,837
	Medium	9,809	1,219
	Affluent	1,867	224
Gender	Female	14,382	1,821
	Male	12,035	1,459
Age	31–45	12,755	1,493
	46–65	5,653	841
	23–30	4,164	530
	18–22	3,845	416

Table 14: Fair-Speech coverage per demographic group: utterances and minutes of audio. Every analyzed group exceeds the analysability threshold of 200 utterances and 30 minutes.

resampling intervals and reflects the precision of the group WER rather than per-utterance spread.

F Appendix: Common Voice English, Dutch, and Danish

This section supports the cross-corpus summary in Section 4.4. Accent groups shown are those meeting the analysability threshold (≥ 200 utterances and ≥ 30 minutes).

Common Voice English (Table 5) offers a useful contrast to Fair-Speech. The accent hierarchy is stable, with India/South Asia trailing US and England at every depth, and the absolute gap grows as pruning deepens. But there is no concealment window: the aggregate degrades from the very first prune, so a practitioner watching only the corpus WER would already see the cost. The relative gap does not widen either: 1.42 at L-0, 1.39 at L-8.

Dutch shows the same persistence without ampli-

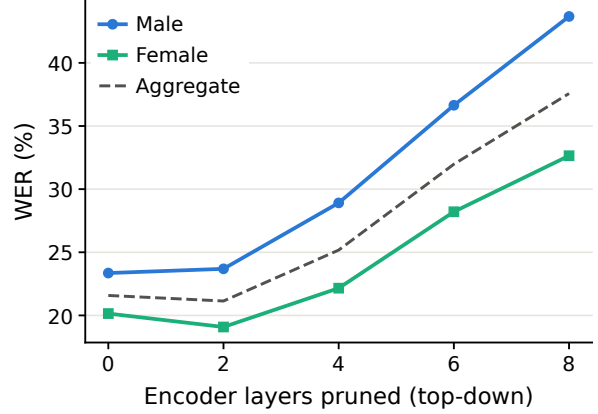


Figure 5: WER by gender under top-down pruning (Fair-Speech, Whisper large-v2; dashed = aggregate). Male speakers trail at every depth, and the gap widens from 3.3 pp at L-0 to 11.1 pp at L-8: female speakers improve slightly at the first prune while male speakers do not, and they degrade more slowly thereafter.

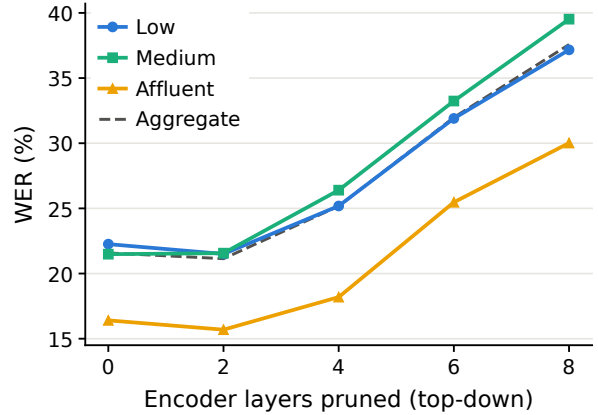


Figure 6: WER by socioeconomic group under top-down pruning (Fair-Speech, Whisper large; dashed = aggregate). Affluent speakers keep an advantage at every depth. The gap does not amplify: the Low-to-Affluent ratio falls from 1.36 to 1.24 as all groups degrade.

fication (Tables 6 and 7, Fig. 8). Belgian speakers trail Netherlands speakers at every scale and every usable depth, yet the ratio between them stays essentially flat, between 1.1 and 1.3, across small, medium, and large-v2 alike. The Dutch gender axis behaves the same way: male speakers trail at every depth, echoing the Fair-Speech gender direction, with the ratio flat near 1.3. The inherited gaps survive pruning unchanged; unlike the racial gap on Fair-Speech, they do not widen, and they behave the same at every model scale.

Danish marks the edge of what this analysis can reach (Table 8). Even unpruned, the aggregate WER sits at 35.5%, leaving little headroom: the 40% bound is crossed by six removed layers.

Axis	Group	n	L-0	L-2	L-4	L-6	L-8
SES	Low	14,741	22.3	21.5	25.2	31.9	37.2
	Medium	9,809	21.5	21.6	26.4	33.2	39.5
	Affluent	1,867	16.4	15.7	18.2	25.5	30.0
Gender	Male	12,035	23.4	23.7	28.9	36.6	43.7
	Female	14,382	20.1	19.1	22.2	28.2	32.6
Age	18–22	3,845	16.2	15.5	18.8	25.8	31.0
	23–30	4,164	19.3	18.1	22.0	28.3	34.5
	31–45	12,755	21.2	21.3	26.5	34.3	41.0
	46–65	5,653	27.0	26.1	28.5	33.7	36.9
	Aggregate	26,417	21.6	21.1	25.2	32.0	37.6

Table 15: Fair-Speech, socioeconomic status, gender, and age (Whisper large); n is the number of utterances per group.

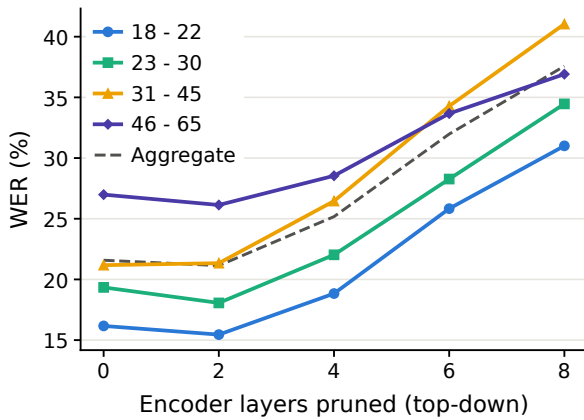


Figure 7: WER by age group under top-down pruning (Fair-Speech, Whisper large-v2; dashed = aggregate). The oldest group (46–65), worst at baseline, degrades slowest (+9.9 pp by L-8), while the 31–45 group degrades fastest (+19.8 pp) and overtakes it as the worst-performing group by L-6. The worst-best gap narrows, but by leveling down, not by helping the group that started worst.

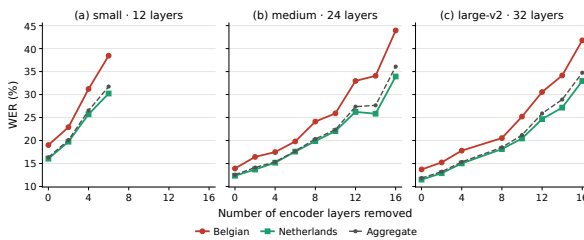


Figure 8: Common Voice Dutch: Belgian vs Netherlands accent WER under top-down pruning, per Whisper scale (usable range only; dashed = aggregate). Belgian speakers trail at every scale and depth, but the ratio stays between 1.1 and 1.3 throughout: the gap does not amplify. Each panel ends where its model leaves the usable range.

The demographic annotations are also sparse; most groups fall below the analysability threshold, in-

Group	n	L-0	L-2	L-4	L-6	L-8	L-10
Asian	3,854	11.1	11.2	14.2	18.5	22.2	29.1
Native Haw.	969	13.8	14.4	16.0	21.3	27.0	36.7
Native Am.	4,616	14.4	15.6	18.0	22.2	26.2	34.8
Hispanic	2,811	15.6	16.1	19.6	22.8	28.6	33.5
White	5,619	16.6	17.9	18.8	23.8	25.8	32.9
MENA	749	17.9	21.0	23.1	27.6	30.6	40.0
Black	7,799	24.4	26.4	32.5	41.7	49.3	55.3
Aggregate	26,417	17.7	18.9	22.3	28.2	33.0	40.0

Table 16: Fair-Speech, ethnicity under LoRA adaptation (Whisper large); n is the number of utterances per group. WER is lower than in the base model (Table 3) for every group at every depth.

cluding all female speakers. We therefore draw no group-level conclusions for Danish, and the table documents coverage rather than findings. Together, the three corpora bound the Fair-Speech result: disparities persist everywhere, but amplification hidden under an improving aggregate appeared only on Fair-Speech with the largest encoder.

Temporal Taxation Compounds Under Post-Training Compression of Whisper Models

Srishti Ginjala¹

Eric Fosler-Lussier¹

Christopher W. Myers²

Srinivasan Parthasarathy¹

¹The Ohio State University, Columbus, OH, USA

²Air Force Research Laboratory, USA

ginjala.1@osu.edu {fosler,srini}@cse.ohio-state.edu

christopher.myers.29@us.af.mil

Abstract

Automatic speech recognition models are audited for demographic fairness at full precision, yet the models that ship to production have been quantized, pruned, and distilled. We ask whether *post-training weight compression*, which alters model weights rather than the audio signal or its feature representation, redistributes error burden across demographic groups. Across the Whisper family on Fair-Speech, Common Voice 25, and AfriSpeech-200, 50% Wanda pruning of Whisper-large-v3 sharply widens the Black/AA-vs-Asian temporal-taxation differential on Fair-Speech: the absolute word-error-rate gap between the worst- and best-served groups more than doubles; at an assumed cost of five seconds of correction effort per transcription error this is a rise from 30 to 64 seconds of correction time per minute of speech. This +111% relative increase is invariant to the assumed per-error cost, survives an audio-quality control, and is only partly mitigated by beam-search decoding, which still leaves an +86% increase. At edge model size, INT4 HQQ quantization compounds catastrophic transcript loops on West African accents by factors of five to seven. Distillation, by contrast, narrows demographic gaps in 21 of 27 evaluated settings (teacher-student pair, precision, and dataset), with the exceptions concentrated on a single model pair. We cast the temporal-taxation construct of Choi and Choi (2025) as a quantitative metric, and show that single-snapshot fairness audits on full-precision models do not capture the deployment-time burden that compression places on already-marginalized speakers.

1 Introduction

Regulatory fairness audits of automatic speech recognition (ASR) systems typically evaluate a full-precision model and report word-error-rate gaps across demographic groups (Koenecke et al., 2020; Tatman and Kasten, 2017; Veliche et al., 2024).

However, the system deployed in production, especially on edge devices, is rarely the same as the one audited. These systems are often reduced-order distilled or quantized models. For instance, the Whisper-large-v3 model alone is downloaded over four million times each month from the Hugging Face Hub and has numerous derivative checkpoints, many of which are quantized to INT8 or INT4, magnitude-pruned, or distilled. This paper examines how the per-group failure profile changes when such *post-training weight compression* is applied, by which we mean quantization, pruning, and distillation of model weights, as distinct from compression of the audio signal or of its feature representation.

The question is not academic. In transcription, captioning, and dictation workflows, recognition errors are typically surfaced to human reviewers for correction, so a demographic gap in word error rate becomes a gap in correction labor: worse-served speakers spend more time repairing transcripts for the same minute of speech. The relative compounding we report does not depend on this assumption (Section 3.4). Choi and Choi (2025) formalize this burden as *temporal taxation*, which we adopt as our headline metric because it expresses ASR errors as a concrete time cost to the speaker rather than an abstract percentage. Compression theory in vision (Hooker et al., 2019, 2020) and multilingual NLP (Ogueji et al., 2022; Ahia et al., 2021) predicts that pruning and quantization disproportionately damage the long tail of input distributions; text-LLM work reports that quantization effects on bias are complex and unpredictable across techniques (Kirsten et al., 2025). Whether ASR demographic groups behave as this long tail under compression is therefore an empirical question with no confidently predicted direction. Our findings diverge from the naive long-tail extrapolation in three ways. Pruning compounding is monotonic in *increasing* model capacity, from -0.56 at *whisper-*

tiny to +0.58 at *whisper-large-v3* (Section 7), the opposite of the pattern reported in vision. Distillation narrows demographic gaps in 21 of 27 settings tested (Section 6), against the prediction that compression damages the tail. And NF4 and HQQ at the same bit-width redistribute error to different groups (Section 5.1), so the extrapolation does not select a direction even within one bit-width.

Prior ASR fairness work treats the model as fixed: per-model audits at FP32 or FP16, no compression (Koenecke et al., 2020; Tatman and Kasten, 2017; Harris et al., 2024; Veliche et al., 2024; Ezema et al., 2025; Cunningham et al., 2024; Koenecke et al., 2024). Cheng et al. (2026) introduces a semantic-axis audit but evaluates only at FP32. Prior ASR compression work has rarely examined fairness. The main exception is Ferraz (2024), which measures how 8-bit quantization affects multilingual demographic gaps in Whisper, but only at 8-bit and without English dialect or accent axes. The closest LLM-side work is Kirsten et al. (2025), who find that inference-acceleration effects on bias in text models are mixed and method-specific.

Our contributions are fourfold. (i) We present the first systematic study of post-training pruning and ASR fairness: 50% Wanda pruning (Sun et al., 2024) of *Whisper-large-v3* more than doubles the Black/AA-vs-Asian temporal-taxation differential on Fair-Speech (+111%), an effect invariant to the assumed per-error cost and robust to an audio-quality control (Section 4), and reduced but not removed by beam-search decoding, which still leaves a +86% increase (Section 5.4). (ii) We extend the 8-bit fairness analysis of Ferraz (2024) to sub-8-bit precision on English dialect and accent: at edge model size, INT4 HQQ quantization multiplies catastrophic-loop rates on West African accents by factors of five to seven, and WER compounding is capacity-dependent and recipe-specific (Section 5). (iii) We operationalize the temporal-taxation construct of Choi and Choi (2025), originally a conceptual proposal, as a reporting convention for ASR compression audits. The metric is a linear rescaling of per-group WER and we do not claim the arithmetic as a contribution; what it buys is a deployment-interpretable time unit and, because of the linearity, a relative compounding measure that is exactly invariant to the cost-per-error assumption (Section 3.4). (iv) We report a counter-expectation result: distillation does not compound demographic gaps but narrows them in most settings tested, with one structured exception (Section 6).

Our results are within the Whisper family, chosen to isolate the effect of compression from confounds with architecture, tokenizer, and training data (Section 3.1); a probe on IBM Granite-4.0-1b-speech (Appendix F) tests one direction of cross-family generalization.

2 Related work

ASR Fairness Audits. A line of work documents demographic disparities in ASR at fixed precision. Koenecke et al. (2020) reports a two-fold racial gap across five commercial systems, and Tatman and Kasten (2017) surfaces a gender-race interaction in early YouTube and Bing captions. Martin and Tang (2020) attribute part of the racial gap to mishandling of habitual *be* and other African American English features, and Harris et al. (2024) extend the gender-dialect interaction to modern ASR. Ezema et al. (2025) catalog cascading effects of ASR bias in spoken-language interfaces, and Cunningham et al. (2024) document downstream impacts on speakers of African American Language. Veliche et al. (2024) release the Fair-Speech benchmark. Koenecke et al. (2024) catalog Whisper hallucination harms that motivate the loop-rate axis tracked here. Cheng et al. (2026) most recently introduced a semantic-axis fairness metric. All of this work treats the model as fixed: it audits *which* model fails on whom, not *what happens to that failure profile after compression*.

Compression and Bias. Some researchers observe that compression damages the long tail of input distributions: Hooker et al. (2019, 2020) demonstrate this for vision models, and Ogueji et al. (2022); Ahia et al. (2021) show the same for multilingual NLP. Kirsten et al. (2025) find that inference-acceleration effects on LLM bias are "complex and unpredictable" across techniques. On the ASR dimension, Ferraz (2024) reports that 8-bit quantization of Whisper amplifies model-related biases (resourcefulness, model size) on FLEURS and Common Voice 13 while leaving speaker-related biases (gender, age) roughly stable. Still, the analysis is limited to 8-bit precision and does not examine English dialect or accent. Feng et al. (2025) benchmark low-bit Whisper for edge deployment and Andreyev (2025) compare quantization recipes; neither studies demographic fairness. The vision and text literatures predict compounding; the speech literature has examined only 8-bit quantization, and only on multilingual

axes.

Adjacent work on Whisper deployment-time fairness. [Ginjala et al. \(2026\)](#) study how acoustic input degradation and decoder architecture interact with ASR demographic fairness. The present work is orthogonal: it perturbs model weights, through quantization, pruning, and distillation, rather than the acoustic input, and the two experimental matrices share no common (model, condition) cell.

3 Methods

3.1 Models

The main analysis covers eight Whisper-family models: five Whisper backbones (*whisper-tiny*, *whisper-base*, *whisper-small*, *whisper-medium*, *whisper-large-v3*; 39M to 1.55B parameters) and three Distil-Whisper students paired with their teachers (*distil-small.en*, *distil-medium.en*, *distil-large-v3*) ([Radford et al., 2023](#); [Gandhi et al., 2023](#)). Although *whisper-large-v3* no longer leads the Open ASR Leaderboard, the Whisper family spans the widest publicly available capacity range and has the most calibrated demographic-fairness baselines in prior work ([Koenecke et al., 2020](#); [Ferraz, 2024](#)). We exclude *Whisper-large-v3-Turbo* because Turbo is itself a layer-pruned compression of *large-v3* and would confound the distillation arm.

Restricting the main analysis to one family is deliberate: a cross-family comparison would confound the compression effect with architecture, tokenizer, training data, and audio encoder, the confound that limits prior multi-family ASR fairness audits. The Granite-4.0 probe in Appendix F gives a one-architecture check on the FP16/INT8 gap structure, but does not test the pruning or INT4 claims cross-family.

3.2 Compression methods

Quantization. The sweep covers four precisions per model: FP16 as the deployment-realistic reference, INT8 ([Dettmers et al., 2022](#)), INT4 NF4 (NormalFloat4), and INT4 HQQ (Half-Quadratic Quantization; [Badri and Shaji, 2023](#)). NF4 and HQQ are both data-free; including both isolates the optimization-recipe axis while holding bit-width and calibration-data status constant. Adapting HQQ to Whisper’s encoder-decoder required an implementation fix, which we verified on a held-out set using FP16. GPTQ ([Frantar et al., 2023](#)) and AWQ ([Lin et al., 2024](#)) are excluded on implementation-availability grounds rather than

algorithmic ones: both algorithms are architecture-agnostic, but the reference libraries we tested (gptqmodel, autoawq) expose per-architecture calibration classes that do not include encoder-decoder speech models. Appendix E documents the specific blockers and notes that torchao provides a module-level configuration path we did not pursue.

Pruning. We apply Wanda 50% unstructured pruning ([Sun et al., 2024](#)), extended to cover the linear layers of Whisper’s encoder, decoder, and decoder cross-attention. Calibration uses 128 utterances from LibriSpeech train-clean-100, held out from every test set used for fairness claims. SparseGPT ([Frantar and Alistarh, 2023](#)) requires per-architecture model classes that do not currently exist for encoder-decoder speech models.

Distillation. Distil-Whisper checkpoints are paired with the corresponding Whisper teacher ([Gandhi et al., 2023](#)), evaluated at each precision in the quantization sweep.

3.3 Datasets and demographic axes

Three benchmarks carry the fairness claims, with LibriSpeech test-clean held in reserve as an aggregate-WER reference only. Fair-Speech ([Veliche et al., 2024](#)) provides controlled read voice-assistant prompts (26,471 utterances, 593 speakers, demographic axes for ethnicity, age, gender, geography, and native-English status). Common Voice 25 supplies the English test split (16,398 clips) with accent labels populated for seven groups at $n \geq 50$ utterances (us 1343, indian 527, england 437, canada 116, australia 100, scotland 53, african 53). AfriSpeech-200 provides 6,318 test utterances over 44 fine-grained African accents with $n \geq 50$.

3.4 Metrics

Per-utterance metrics include word error rate (WER), deletion rate, insertion rate, and a per-utterance loop flag set when a hypothesis contains more than five 5-gram repetitions or when the hypothesis-to-reference length ratio exceeds three (thresholds set before the full sweep ran; see Appendix D). For each (model, configuration, dataset) cell, we aggregate WER by demographic group and report the utterance-weighted mean WER with $B = 200$ bootstrap 95% confidence intervals. Per-cell fairness is summarized by the max-min ratio $\text{MMR} = \max_g \text{WER}_g / \min_g \text{WER}_g$ and the compounding index $(\text{MMR}_{\text{compressed}} - \text{MMR}_{\text{FP16}}) / \text{MMR}_{\text{FP16}}$.

The headline metric is the temporal-taxation differential adapted from Choi and Choi (2025). For demographic group g in a cell with audio of mean rate wpm words per minute,

$$T_g = \text{WER}_g \cdot C \cdot wpm,$$

where C is the assumed cost per error in seconds. T_g is a linear rescaling of WER_g ; we do not present the transformation itself as a contribution. Its role is to express the gap in a deployment-interpretable unit and to separate the cost-dependent part of the claim from the cost-independent part. We report the worst-versus-best group difference and the relative change in this difference under compression. Because T_g is linear in C , the relative change in the differential is exactly invariant in C ; only absolute anchors depend on C . We anchor at $C = 5$ s/edit and report sensitivity over $C \in \{2, 5, 8\}$ s/edit (Appendix B). The linear-cost assumption applies to per-token errors; loop-induced cost, which scales non-linearly with hypothesis length, is tracked separately on the loop-rate axis (Section 5.2).

3.5 Statistical methodology

We test paired per-utterance WER differences with a paired permutation test ($B = 2,000$ permutations per cell, two-sided) and adjust within each cell with the Benjamini-Hochberg procedure across demographic groups (Benjamini and Hochberg, 1995). Confidence intervals for per-group means use $B = 200$ percentile bootstrap samples. Ranking stability of models by MMR across precisions is summarized by Kendall’s τ . Following Li et al. (2024), we test for an audio-quality confound on Fair-Speech by computing a per-utterance SNR proxy, fitting an ordinary-least-squares regression of WER on SNR, and re-testing the demographic F-statistic after the SNR adjustment (full output in Appendix C).

4 Pruning effects on demographic temporal taxation

4.1 Headline result

Deploying *whisper-large-v3* with 50% Wanda unstructured pruning on Fair-Speech raises the Black/AA-vs-Asian temporal-taxation differential from 30.14 to 63.73 seconds of correction time per minute of speech at $C = 5$ s/edit, a relative increase of 111.4% (Figure 1). The differential more than doubles, and the relative change is exactly invariant in C (sensitivity table in Appendix B),

so the +111% claim does not depend on the cost-per-error assumption. The asymmetry is concentrated in the worse-served group: under Wanda, Black/AA WER rises by +7.89 percentage points (BH-adjusted $p \approx 0$, $n = 7782$), while Asian WER rises by only +0.43 points ($p_{\text{BH}} = 0.006$, $n = 3853$).

The pattern replicates across cells: *whisper-large-v3* + Wanda widens the indian-vs-canada Common Voice 25 differential by +68.9%, and *whisper-small* + Wanda widens Black/AA-vs-Asian on Fair-Speech by +51.9% and African-vs-Canadian on Common Voice 25 by +35.1% (absolute s/min anchors for all cells in Appendix B). Nine of the top ten cells in Figure 1 are Wanda; pruning produces the largest demographic-redistribution effects in the experimental matrix. Figure 2 panel (a) shows the per-group breakdown: under *whisper-large-v3* + Wanda, Black/AA stands alone at +7.89 pp while Asian, Native American, Pacific Islander, and Middle Eastern speakers cluster below +1.5 pp.

4.2 Audio-quality confound

A standing concern in ASR fairness measurement is that demographic groups are confounded with recording quality, so observed WER gaps reflect microphone or environmental differences rather than accent or dialect (Li et al., 2024). We compute a per-utterance SNR proxy on Fair-Speech, regress per-utterance WER on SNR, and re-test the demographic F-statistic on the residual. The ethnicity effect survives the SNR control in 34 of 34 (model, precision) Fair-Speech cells. After SNR adjustment, the demographic MMR is larger than the raw MMR in every cell, by 15.0% on average (median 13.9%); the SNR control widens the gap rather than narrowing it. The pruning compounding result is therefore not explained by SNR-measurable audio-quality variation; the raw corpus WER mildly understates the disparity rather than overstating it. SNR is a partial proxy: reverberation and non-linear distortion are not captured, and residual confounding from those sources cannot be excluded (Limitations).

5 Quantization effects across capacity and method

5.1 WER compounding and capacity scaling

WER compounding under quantization is capacity-bounded (Figure 3). At *whisper-tiny* on Com-

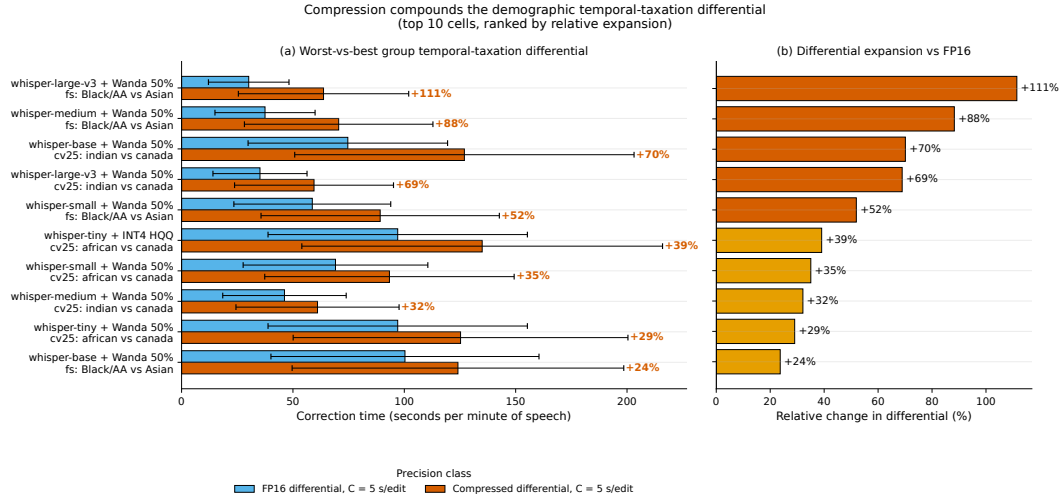


Figure 1: Temporal-taxation differential under pruning and quantization, between the worst- and best-served demographic groups, top ten cells by relative compounding. **(a)** Absolute differential in seconds of correction per minute of speech, FP16 (blue) versus compressed (vermilion), at five seconds per error; horizontal extensions show the differential range when the per-error cost varies from two to eight seconds. **(b)** Relative change in the differential under compression, which is invariant to the per-error cost assumption. The headline cell is *whisper-large-v3* with 50% Wanda pruning on Fair-Speech (Black/AA vs Asian, +111%).

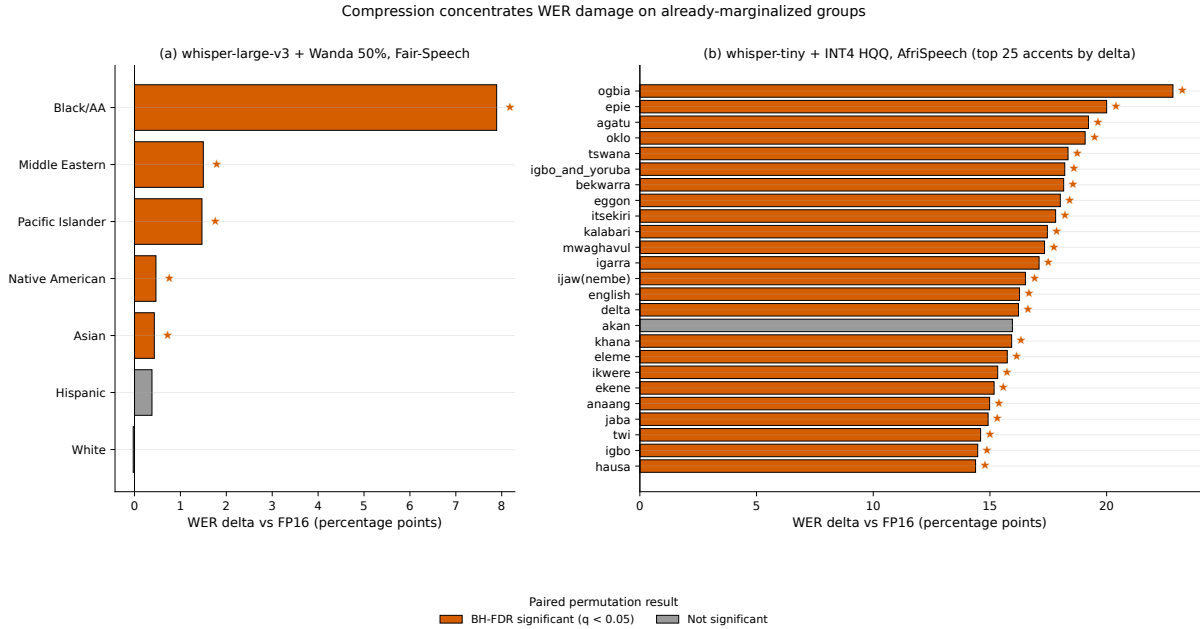


Figure 2: Per-group WER damage under pruning (a) and quantization (b) concentrates on already-marginalized groups. Per-group WER delta (compressed minus FP16, percentage points) with paired-permutation BH-FDR-significant cells in vermilion ($q < 0.05$). (a) Fair-Speech ethnicity under *whisper-large-v3* + 50% Wanda: Black/AA +7.89 pp; Hispanic and White below the BH threshold. (b) AfriSpeech top-25 accents by delta under *whisper-tiny* + INT4 HQQ; all but akan BH-significant.

mon Voice 25, INT4 HQQ produces 7 of 7 BH-significant per-accent deltas, all positive, ranging from +5.45 pp (Canadian) to +13.19 pp (African). At *whisper-large-v3* on Common Voice 25, INT4 HQQ produces a smaller MMR shift (+0.216), and only the British English accent reaches BH significance (+0.73 pp); the MMR change at this size is driven by which accent occupies the extremes of the per-group WER distribution rather than by large per-group magnitudes. Four mid-size Common Voice 25 quantization cells show the opposite pattern, with compounding index between -0.13 and -0.06 : MMR narrows because the best-served accent degrades proportionally harder than the worst-served one. Under *whisper-base* + INT4 HQQ, Canadian WER rises 28% (9.6% to 12.3%) while African WER rises 12% (24.9% to 27.8%).

NF4 and HQQ at the same bit-width produce distinguishable redistributions. At *whisper-tiny* both recipes compound on the WER axis, but unevenly: on Common Voice 25, HQQ yields 7 of 7 BH-significant per-accent deltas against 5 of 7 for NF4, and on the loop-rate axis HQQ reaches significance on 7 of 62 AfriSpeech accents and 1 of 7 Fair-Speech groups where NF4 reaches 1 of 62 and 0 of 7. NF4 nonetheless flattens MMR more at this size (Common Voice 25 compounding index -0.117 for NF4 against $+0.010$ for HQQ), because it damages the best-served accent proportionally harder. At *whisper-large-v3* the per-group picture inverts on AfriSpeech, where NF4 reaches BH significance on 3 of 63 accents and HQQ on none, while on Common Voice 25 HQQ shifts MMR further than NF4 (+0.216 against +0.138). The choice of 4-bit recipe therefore carries a fairness consequence as well as an accuracy consequence, consistent with [Kirsten et al. \(2025\)](#)’s finding that quantization method interacts with bias on text LLMs. Figure 2 panel (b) shows the per-accent evidence for the strongest WER-compounding cell.

5.2 Loop-rate compounding

Aggregate WER is a poor summary of user-experienced cost when a fraction of utterances produce catastrophic transcripts: long hallucination loops, sustained 5-gram repetitions, and length-blowups that force minutes of correction effort on a single sentence ([Koenecke et al., 2024](#)). We track the loop axis separately. A hypothesis is flagged as a loop when it contains more than five 5-gram repetitions or its length exceeds three times the reference length; the thresholds are set in advance of

the full sweep (Appendix D).

Loop rate compounds (increases) sharply with edge model size for West African accents. Under *whisper-tiny* with INT4 HQQ on AfriSpeech, seven of 62 accents become BH-significantly worse than FP16 in paired permutation (Figure 4). Kanuri ($n=66$) moves from 4.55% to 30.30% loop rate, a 6.67-fold increase ($p_{BH} = 2.68 \times 10^{-3}$). Hausa ($n=196$) moves from 1.53% to 9.18%, a 6.0-fold increase ($p_{BH} = 1.43 \times 10^{-2}$). Yoruba ($n=648$) moves from 1.23% to 6.33%, a 5.12-fold increase ($p_{BH} = 8.62 \times 10^{-5}$). The Fair-Speech replication is in the same direction: under *whisper-tiny* + INT4 HQQ, Black/AA loop rate rises from 0.51% to 1.28% (a 2.5-fold increase, $n = 7807$, $p_{BH} = 2.46 \times 10^{-6}$). White and other ethnic groups on Fair-Speech show smaller, non-significant changes.

Since the *relative* compounding metric controls for the FP16 baseline within each accent, these results isolate the quantization-induced increase in loop rate from baseline accent difficulty. A Kanuri 4.55% loop rate at FP16 is already high; the claim is that INT4 HQQ multiplies that rate by $6.67\times$, not that the absolute rate is high. The signal weakens at larger model sizes: at *whisper-large-v3* + INT4 HQQ on AfriSpeech, no accent reaches BH significance for loop-rate compounding (Appendix A). The loop-rate finding therefore applies to *edge-deployable* models, where INT4 is the realistic deployment precision.

5.3 Ranking stability

Model rankings by MMR are highly stable across precisions. Kendall’s τ between the FP16 ranking and the compressed-precision ranking is 1.000 in 11 of 12 (dataset, precision) cells, with one adjacent-pair swap at Fair-Speech + INT4 HQQ ($\tau = 0.929$); full table in Appendix A. A practitioner choosing a model on a full-precision audit can rely on the relative ordering surviving deployment at sub-8-bit precision; the absolute fairness gap, in contrast, does not survive (Sections 4 and 5.1).

5.4 Beam-search decoding does not mitigate compounding

A natural mitigation to consider is whether beam-search decoding compensates for compression error. We rerun two models (*whisper-small*, *whisper-large-v3*) at two precisions (FP16, INT4 NF4) on Fair-Speech and Common Voice 25 with a beam width of 5, producing 8 quantization ablation cells

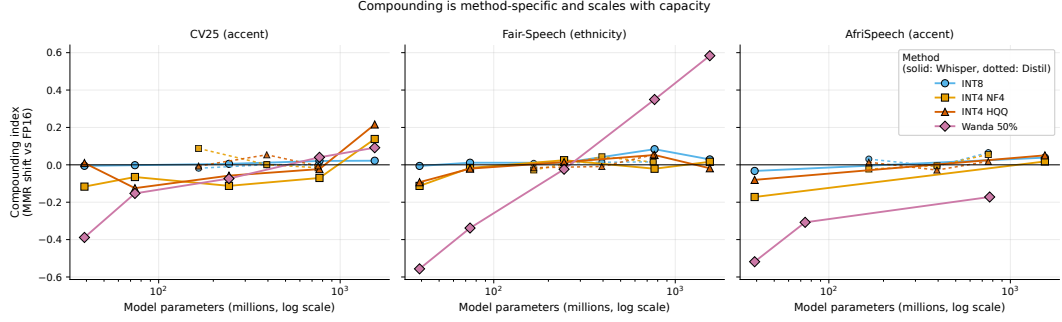


Figure 3: Quantization and pruning compounding is method-specific and scales with capacity. Compounding index (MMR shift relative to FP16) versus Whisper parameter count (log scale); one panel per dataset; each panel covers four methods (INT8, INT4 NF4, INT4 HQQ, and 50% Wanda pruning). Solid: Whisper backbones; dotted: Distil-Whisper. The Fair-Speech Wanda curve (purple) is monotonic from -0.56 at *whisper-tiny* to $+0.58$ at *whisper-large-v3*, opposite in direction to sub-8-bit quantization. AfriSpeech Wanda is evaluated only at *whisper-tiny*, *whisper-base*, and *whisper-medium*.

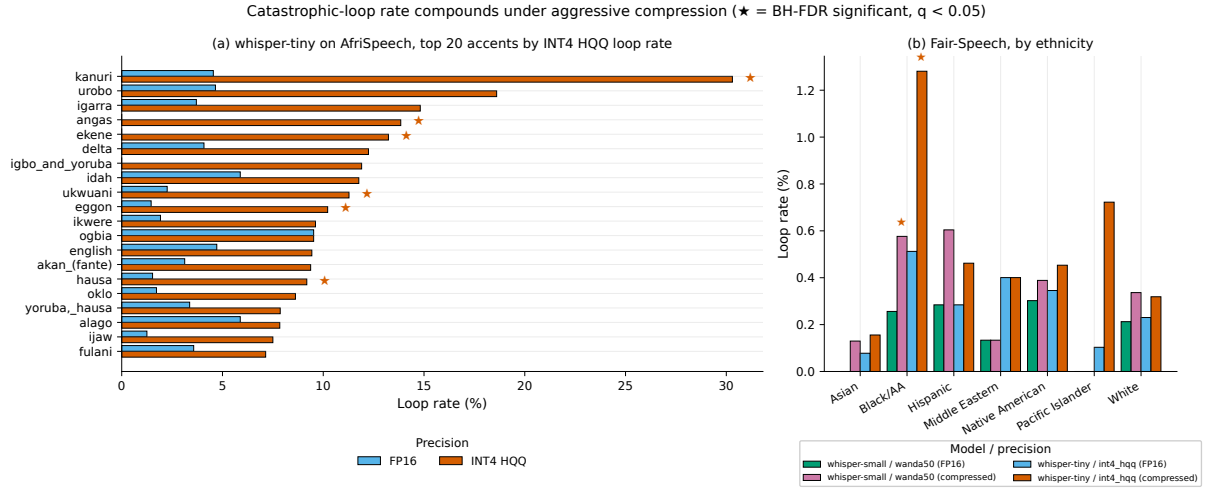


Figure 4: Catastrophic-loop rate under quantization and pruning compounds at edge capacity ($\star$ = BH-FDR significant, $q < 0.05$). (a) *whisper-tiny* on AfriSpeech, top-20 accents by INT4 HQQ loop rate. (b) Fair-Speech ethnicity loop rates for *whisper-tiny* + INT4 HQQ and *whisper-small* + 50% Wanda.

against the greedy baseline, and separately rerun the headline pruning cell (*whisper-large-v3* + 50% Wanda on Fair-Speech) at the same beam width. In six of the eight cells the beam-5 MMR moves by less than 0.2 from its greedy counterpart, and in three of the four (model, dataset) pairs beam-5 leaves the NF4 compounding index unchanged or slightly reduces it. In the fourth pair, *whisper-large-v3* + INT4 NF4 on Common Voice 25, beam-5 *amplifies* the compounding index relative to greedy (+0.296 vs +0.138). Loop rate increases under beam-5 in eight of eight cells, with the largest absolute jump on *whisper-large-v3*. Beam search is not a deployment-time fairness fix for quantization compounding; it can amplify it.

On the pruning arm, beam-5 helps but does not resolve the disparity. Under *whisper-large-v3* + 50% Wanda on Fair-Speech, the compounding index falls from +0.584 under greedy decoding to +0.468 under beam-5, and the Black/AA-vs-Asian temporal-taxation differential rises from 27.95 to 51.91 s/min rather than from 30.14 to 63.73 s/min, so the headline increase moves from +111% to +86%. The redistribution is unchanged in shape: Black/AA still carries by far the largest degradation (+5.63 pp under beam-5 against +7.89 pp under greedy), the next-largest group delta is under +1.3 pp, and six of seven ethnicity groups reach BH significance under beam-5 against five of seven under greedy, because beam-5's lower variance makes the smaller deltas detectable. Beam search buys back roughly a fifth of the pruning compounding at roughly twice the inference cost; it does not make pruning demographically neutral.

6 Distillation effects on demographic gaps

At the deployment-relevant model size, distillation narrows the demographic gap on the accent axis: six of the seven *distil-large-v3* cells on AfriSpeech and Common Voice 25 reduce MMR relative to their *whisper-large-v3* teacher (the exception, Common Voice 25 at FP16, is a +0.003 shift within noise), and on AfriSpeech under INT4 HQQ (the only pair and precision evaluated at this scale) 57 of 63 accents narrow with BH significance. The direction replicates across the matrix: over 27 (teacher, student, precision, dataset) cells, distillation narrows demographic fairness gaps in 21 cells, shows systematic widening on 4, and shows near-zero shifts within noise on the remaining 2 (Figure 5). The widening is structured rather than

scattered, concentrated on the *whisper-medium* → *distil-medium.en* pair on Fair-Speech ethnicity at every precision (gap-widening index +0.011 to +0.151); the two near-zero shifts (*distil-large-v3* + Fair-Speech + INT4 HQQ at +0.010, *distil-large-v3* + Common Voice 25 + FP16 at +0.003) are within measurement noise.

Aggregate WER rises uniformly under distillation across all 27 cells, with relative student-vs-teacher WER deltas ranging from +0.07% on *distil-small.en* + INT4 HQQ + Common Voice 25 to +31.4% on *distil-medium.en* + INT4 HQQ + Fair-Speech. The narrowing is a relative-parity result rather than an absolute improvement: distillation raises every group's WER, but the worst-served group's WER scales sub-linearly with the aggregate increase, so the ratio compresses. Distillation is therefore fairness-preserving on the accent axis, with one exception (the medium-capacity teacher-student pair on read-speech ethnicity). Unlike pruning and sub-8-bit quantization, distillation does not concentrate cost on already-marginalized groups, even though it makes every group worse in absolute terms.

Candidate mechanisms. Why distillation narrows where pruning compounds is open. Training on teacher pseudo-labels averages the student's loss over the teacher's full input distribution, which may regularize against the worst-group behavior the teacher already exhibits, and may also push the student toward a smoother decision boundary that is less peaked on majority-group features. Adjudicating between these would require training-data attribution or controlled pseudo-label experiments, which we leave to future work.

7 Concluding Remarks and Discussion

Across the three compression families, the deployment-vs-benchmark gap is asymmetric. Pruning compounds the demographic temporal-taxation differential at the deployment scale; distillation narrows it on the accent axis (with one structured-ethnicity exception); and quantization sits between, capacity-bounded and recipe-specific. The three families are therefore not interchangeable from a fairness perspective. A practitioner choosing between 50% Wanda pruning and INT4 HQQ quantization to fit *Whisper-large-v3* on a mobile device is making a fairness decision: Wanda more than doubles the Black/AA-vs-Asian Fair-Speech differential, while sub-8-bit quantization

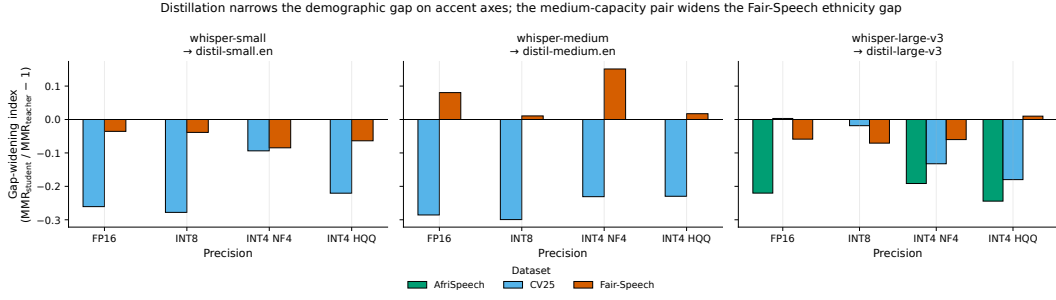


Figure 5: Distillation narrows the demographic gap on accent axes but widens the Fair-Speech ethnicity gap for the whisper-medium → distil-medium.en pair (center panel). Gap-widening index = $(\text{MMR}_{\text{student}} / \text{MMR}_{\text{teacher}}) - 1$, by precision; one panel per teacher-student pair; bars colored by dataset; negative values indicate narrowing.

at the same model size shifts MMR by less than 0.25 and reaches BH significance on only one accent group (England, +0.73 pp). This characterization applies to the data-free 4-bit recipes and to Wanda; GPTQ, AWQ, and SparseGPT, which dominate practitioner deployment in many settings, fall outside this matrix on implementation-availability grounds: the algorithms are architecture-agnostic, but the reference libraries expose per-architecture calibration classes that do not include encoder-decoder speech models (Appendix E).

Sub-8-bit quantization compounding on the WER axis fades as model size grows: at *whisper-large-v3*, INT4 HQQ yields a Common Voice 25 MMR shift of +0.216 and smaller Fair-Speech shifts concentrated on specific groups. The Wanda axis moves in the opposite direction: across the five Whisper backbones, the Fair-Speech compounding index is strictly monotonic in parameter count, sweeping from -0.56 at *whisper-tiny* through -0.34 , -0.02 , $+0.35$ to $+0.58$ at *whisper-large-v3* (Figure 3), where the +111% taxation differential is the endpoint of the curve. At edge capacity, Wanda flattens MMR by crushing the best-served group harder than the worst-served; compounding emerges at medium capacity. One candidate mechanism, consistent with prior compression work in vision (Hooker et al., 2019, 2020) and multilingual NLP (Ogueji et al., 2022; Ahia et al., 2021), is that Wanda’s magnitude $\times$ activation pruning criterion preferentially removes weights serving low-frequency input distributions, which in Whisper training data correspond to the demographic groups already worst-served at full precision. Quantization, by contrast, scales magnitudes uniformly within each quantization bucket and does not selectively prune the long tail. Our experiments establish the empirical asymmetry but

cannot adjudicate the mechanism: the deletion-rate decompositions in Appendix A are consistent with a long-tail account but are not decisive. A decisive test would need weight-level attribution to training-data subpopulations, or a controlled experiment varying long-tail representation in the calibration set.

Because the temporal-taxation differential is linear in C , the relative change under compression cancels C exactly, so the +111% pruning compounding claim holds whether a Fair-Speech caption error costs 2, 5, or 8 seconds; only the absolute anchors (12 to 102 s/min across C) depend on the assumption. At five seconds per error, the worst-served Fair-Speech group already carries roughly 30 minutes of additional correction effort per hour of speech at full precision, and 50% Wanda raises this to about 64 minutes: an extra half-hour of repair labor per hour of audio attributable to compression alone. This decoupling makes temporal taxation a useful operational fairness metric beyond mean WER (Choi and Choi, 2025): single-snapshot FP16 audits miss the redistribution that follows at deployment (Koencke et al., 2020; Veliche et al., 2024; Ferraz, 2024), so deployment-time audits should evaluate at least one sub-8-bit precision and one pruned condition.

Limitations

The main results cover one model family, Whisper; a comprehensive cross-family audit is left to future work.

The evaluation is English-only and uses read-speech corpora exclusively (voice-assistant prompts, read sentences, and read clinical text). Post-training weight compression may compound differently on tonal or agglutinative languages. The read-speech caveat is asymmetric in a known direc-

tion: the WER gap between spontaneous and read speech is largest for AAVE-speaking and Black/AA populations (Martin and Tang, 2020; Cunningham et al., 2024), and these are also the groups carrying the largest compounding effect in our experiments. The +111% Wanda compounding result on Fair-Speech is therefore plausibly a lower bound on the corresponding effect in spontaneous AAVE; read-speech evaluation works against detection of the disparity, not toward it.

Audio-quality provenance remains a partial confound. The SNR control in Section 4.2 suggests that recording quality is not the source of the demographic gap, but SNR is a partial proxy for recording quality: it does not capture reverberation or non-linear distortion, and a perceptual-quality control (for example DNSMOS) was not run. A residual confound between group membership and recording conditions therefore cannot be ruled out without controlled recording-environment data (Li et al., 2024).

Each compression method is evaluated through a single reference implementation, so results are specific to that implementation rather than the underlying algorithm. GPTQ, AWQ, and SparseGPT are excluded because the reference libraries we tested expose per-architecture calibration classes that do not include encoder-decoder speech models, not because the algorithms are architecture-specific (Frantar et al., 2023; Lin et al., 2024; Frantar and Alistarh, 2023); Wanda 2:4 structured pruning was attempted and produced catastrophic outputs (WER 194% on a 20-utterance test with 32 calibration utterances). Pruning is therefore evaluated only at unstructured 50% sparsity: on all five Whisper backbones for Fair-Speech and Common Voice 25, and on *whisper-tiny*, *whisper-base*, and *whisper-medium* for AfriSpeech. All compression is strictly post-training; quantization-aware training and demographic-aware retraining are natural extensions left to future work.

All results are measured on single released checkpoints. Our contrasts are within-checkpoint, comparing FP16 against compressed versions of identical weights, so variance attributable to the original training run largely differences out. What remains unmeasured is whether the magnitude of the compounding effect would replicate across independently trained checkpoints. Group fairness metrics are known to vary substantially across training runs in other modalities (Ganesh et al., 2023), and we did not retrain any Whisper model.

Four additional experiments were proposed during review and are not included here: an AWQ arm configured through torchao, a DNSMOS perceptual-quality control on the *whisper-large-v3* + Wanda + Fair-Speech cell, an FP8 (E4M3) arm paralleling INT8, and a search-over-scale variant of INT8 on *whisper-large-v3* + Fair-Speech. None were completed in time for this version. The first would test whether an activation-aware 4-bit recipe redistributes error differently from the two data-free recipes evaluated here; the second would replace the SNR proxy with a perceptual measure; the third and fourth would probe how sensitive the INT8-is-gentle result is to number format and scale selection. We record their absence here rather than leaving the commitments unaddressed; the conclusions of Sections 4 and 5 should be read as covering NF4, HQQ, INT8 at library-default scale, and Wanda 50% unstructured only.

Ethical considerations

All datasets are used under their published licenses, with no modification and no new data collection; demographic categories follow speaker self-identification in the source corpora. The compute footprint is approximately 150 A100 GPU-hours. Code, configurations, and result tables will be released on publication.

A foreseeable misreading is that compounding argues against deploying compressed ASR for marginalized populations. The intended reading is the opposite: withholding edge-deployable ASR would itself disadvantage the same speakers, since edge precision is often the only feasible delivery channel. These findings motivate fairness-aware compression and per-deployment audits, not a retreat from compression.

References

- Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. [The low-resource double bind: An empirical study of pruning for low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*.
- Allison Andreyev. 2025. [Quantization for OpenAI’s whisper models: A comparative analysis](#). *Preprint*, arXiv:2503.09905.
- Hicham Badri and Appu Shaji. 2023. [Half-quadratic quantization of large machine learning models](#). Mobius Labs blog post.

- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289-300.
- Ting-Hui Cheng, Line H. Clemmensen, and Sneha Das. 2026. [Beyond word error rate: Auditing the diversity tax in speech recognition through dataset cartography](#). *Preprint*, arXiv:2603.05267.
- Anna Seo Gyeong Choi and Hoon Choi. 2025. [Fairness of automatic speech recognition: Looking through a philosophical lens](#). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Jay Cunningham, Su Lin Blodgett, Michael Madaio, Hal Daumé III, Christina Harrington, and Hanna Wallach. 2024. [Understanding the impacts of language technologies’ performance disparities on African American language speakers](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12826-12833, Bangkok, Thailand. Association for Computational Linguistics.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [LLM.int8\(\): 8-bit matrix multiplication for transformers at scale](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kelechi Ezema, Chelsea Chandler, Rosy Southwell, Nimal Cholendiran, and Sidney D’Mello. 2025. [“It Feels like We’re Not Meeting the Criteria”: Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Spoken Language Interfaces](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1-13. Association for Computing Machinery.
- Chen Feng, Yicheng Lin, Shaojie Zhuo, Chenzheng Su, Ramchalam Kinattinkara Ramakrishnan, Zhaocong Yuan, and Xiaopeng Zhang. 2025. [Edge-ASR: Towards low-bit quantization of automatic speech recognition models](#). *Preprint*, arXiv:2507.07877.
- Thomas Palmeira Ferraz. 2024. [Efficient compression of multitask multilingual speech models](#). *Preprint*, arXiv:2405.00966.
- Elias Frantar and Dan Alistarh. 2023. [SparseGPT: Massive language models can be accurately pruned in one-shot](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10323-10337. PMLR.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. [GPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.
- Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. [Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling](#). *Preprint*, arXiv:2311.00430.
- Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. 2023. [On the impact of machine learning randomness on group fairness](#). In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1789-1800. Association for Computing Machinery.
- Srishti Ginjala, Eric Fosler-Lussier, Christopher W. Myers, and Srinivasan Parthasarathy. 2026. [Do LLM decoders listen fairly? benchmarking how language model priors shape bias in speech recognition](#). *Preprint*, arXiv:2604.21276.
- Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. [Modeling gender and dialect bias in automatic speech recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166-15184, Miami, Florida, USA. Association for Computational Linguistics.
- Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2019. [What do compressed deep neural networks forget?](#) *Preprint*, arXiv:1911.05248.
- Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. [Characterising bias in compressed models](#). *Preprint*, arXiv:2010.03058.
- Elisabeth Kirsten, Ivan Habernal, Vedant Nanda, and Muhammad Bilal Zafar. 2025. [The impact of inference acceleration on bias of LLMs](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1834-1853, Albuquerque, New Mexico. Association for Computational Linguistics.
- Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. [Careless whisper: Speech-to-text hallucination harms](#). In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Touns, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. [Racial disparities in automated speech recognition](#). *Proceedings of the National Academy of Sciences*, 117(14):7684-7689.
- Changye Li, Trevor Cohen, and Serguei Pakhomov. 2024. [Reexamining racial disparities in automatic speech recognition performance: The role of confounding by provenance](#). *Preprint*, arXiv:2407.13982.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. [AWQ: Activation-aware weight quantization for LLM compression and acceleration](#). In *Proceedings of Machine Learning and Systems (MLSys)*.

Joshua L. Martin and Kevin Tang. 2020. [Understanding racial disparities in automatic speech recognition: The case of habitual “be”](#). In *Proceedings of Interspeech 2020*, pages 626-630.

Kelechi Ogueji, Orevaoghene Ahia, Gbemileke Onilude, Sebastian Gehrmann, Sara Hooker, and Julia Kreutzer. 2022. [Intriguing properties of compression on multilingual models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9092-9110, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. [A simple and effective pruning approach for large language models](#). In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Rachael Tatman and Conner Kasten. 2017. [Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions](#). In *Proceedings of Interspeech 2017*, pages 934-938.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. [Towards measuring fairness in speech recognition: Fair-Speech dataset](#). Preprint, arXiv:2408.12734.

A Ranking stability and per-cell significance

Dataset	Precision	Kendall τ
AfriSpeech	INT8	1.000
AfriSpeech	INT4 NF4	1.000
AfriSpeech	INT4 HQQ	1.000
Common Voice 25	INT8	1.000
Common Voice 25	INT4 NF4	1.000
Common Voice 25	INT4 HQQ	1.000
Common Voice 25	Wanda 50%	1.000
Fair-Speech	INT8	1.000
Fair-Speech	INT4 NF4	1.000
Fair-Speech	INT4 HQQ	0.929
Fair-Speech	Wanda 50%	1.000

Table 1: Kendall’s τ between FP16 model rankings by MMR and rankings under each compressed precision. The single adjacent-pair swap at Fair-Speech + INT4 HQQ is the only deviation from a perfect ordering.

Aggregated BH-FDR significance counts ($q < 0.05$ within each cell, paired permutation vs. FP16) over the 56 Common Voice 25 (model, accent) pairs are: INT8, 1 of 56 significant per-accent deltas (*distil-large-v3*, us, +0.40 pp); INT4 NF4,

11 of 56 (concentrated on *whisper-tiny* and *whisper-base*); INT4 HQQ, 12 of 56 (the *whisper-tiny* + Common Voice 25 cell carries all seven accent deltas at $q < 0.05$). Counting whole cells instead, at least one per-group delta reaches $q < 0.05$ in 6 of 21 INT8 cells, 18 of 21 INT4 NF4 cells, 15 of 21 INT4 HQQ cells, and 13 of 13 Wanda 50% cells. Across all three datasets INT8 produces 7 significant per-group deltas out of 427. Four of the seven fall on AfriSpeech accents with $n \leq 101$ (*bini* +3.35 pp, *khana* +3.05 pp, *etche* +2.64 pp, *mwaghavul* +1.80 pp); on the two larger benchmarks the significant deltas are at most +0.40 pp (Common Voice 25 us) and +0.22 pp (Fair-Speech Black/AA). INT8 is close to, but not exactly, fairness-neutral, and its exceptions sit on the smallest accent strata. The Wanda + *whisper-large-v3* + Fair-Speech result is concentrated rather than uniform: Black/AA carries +7.89 pp ($p_{BH} \approx 0$) while the next-largest delta is more than five times smaller, Middle Eastern at +1.50 pp, followed by Pacific Islander at +1.47 pp; Hispanic (+0.38 pp, $p_{BH} = 0.15$) and White (−0.03 pp) fall below the BH threshold. Decomposing the same cell, deletion rate accounts for 61% of the Black/AA WER increase and only 34% of the Asian increase, consistent with a long-tail account of how Wanda’s magnitude $\times$ activation criterion preferentially removes weights serving rarer input distributions (Hooker et al., 2019).

B Temporal-taxation sensitivity

Because $T_g = \text{WER}_g \cdot C \cdot \text{wpm}$ is linear in C , the worst-vs-best differential scales linearly with C and the relative change under compression cancels C exactly. Table 2 reports the top ten cells by relative compounding at $C \in \{2, 5, 8\}$ s/edit. Relative change is identical at every C ; absolute anchors scale by $C/5$ from the middle column. Nine of the top ten cells are Wanda; the only sub-8-bit quantization cell to enter the top ten is *whisper-tiny* + INT4 HQQ on CV25.

C SNR confound check

Following Li et al. (2024), we compute a per-utterance SNR proxy on every Fair-Speech utterance, regress per-utterance WER on SNR via ordinary least squares, and re-test the demographic F-statistic on the residual. Every one of the 34 (model, precision) Fair-Speech cells remains demographically significant after the SNR control

Model	Cell	$C=2$	$C=5$	$C=8$	rel.
whisper-large-v3	Wanda, FS Black/AA vs Asian	12.06→25.49	30.14→63.73	48.23→101.97	+111%
whisper-medium	Wanda, FS Black/AA vs Asian	14.99→28.21	37.46→70.53	59.94→112.85	+88%
whisper-base	Wanda, CV25 indian vs canada	29.86→50.79	74.64→126.97	119.43→203.16	+70%
whisper-large-v3	Wanda, CV25 indian vs canada	14.08→23.79	35.20→59.47	56.32→95.15	+69%
whisper-small	Wanda, FS Black/AA vs Asian	23.48→35.67	58.69→89.17	93.91→142.68	+52%
whisper-tiny	INT4 HQQ, CV25 african vs canada	38.82→53.99	97.04→134.97	155.27→215.95	+39%
whisper-small	Wanda, CV25 african vs canada	27.64→37.33	69.10→93.33	110.56→149.32	+35%
whisper-medium	Wanda, CV25 indian vs canada	18.48→24.41	46.19→61.03	73.91→97.66	+32%
whisper-tiny	Wanda, CV25 african vs canada	38.82→50.10	97.04→125.26	155.27→200.41	+29%
whisper-base	Wanda, FS Black/AA vs Asian	40.12→49.63	100.30→124.09	160.48→198.54	+24%

Table 2: Worst-vs-best temporal-taxation differential in s/min, FP16 → compressed, at three cost-per-edit values. The relative change column is invariant in C .

(F -test $p < 0.05$). The MMR after SNR adjustment is larger than the raw MMR in 34 of 34 cells, on average 15.0% larger (median 13.9%; range 4.7% to 32.3% larger across cells). The SNR control therefore widens the gap rather than narrowing it, so the raw corpus WER mildly understates the disparity.

D Loop-flag threshold and reproducibility

The per-utterance loop flag triggers when a hypothesis contains more than five 5-gram repetitions or its length exceeds three times the reference length. Both thresholds were fixed before the full sweep was run, based on a preliminary inspection of 100 utterances in which non-pathological hypotheses had at most five 5-gram repetitions (intentional repetition in voice-assistant prompts) and a length ratio of at most 2.7.

Threshold sensitivity. To verify that the loop-rate compounding result does not hinge on these specific cutoffs, we re-derive the loop flag at two perturbations of the (rep5, length-ratio) pair: a loose setting (4, 2.5) and a strict setting (6, 3.5). The recomputation is run directly on the stored per-utterance rep5 and len_ratio columns and requires no re-inference. Table 3 reports the H2 headline cell (*whisper-tiny* + INT4 HQQ on AfriSpeech) at each setting. Aggregate loop rate, BH-significant accent count, and the compounding-index sign are stable. Three of the five headline West African accents (Kanuri, Yoruba, Hausa) clear BH-FDR significance at every threshold (Table 4). Swahili reaches significance only at the strict setting; Igbo at none. Per-accent loop-rate ratios for these three accents stay within a 3.5 to 6.7-fold band across the threshold range.

All inference uses greedy decoding (except for the beam-search ablation in Section 5.4) with

(rep5, len-ratio)	FP16→HQQ	BH-sig	CI
Loose (4, 2.5)	2.07% → 6.79%	6/62	+0.73
Baseline (5, 3.0)	1.50% → 5.78%	7/62	+0.68
Strict (6, 3.5)	1.42% → 5.46%	8/62	+0.56

Table 3: H2 cell (*whisper-tiny* + INT4 HQQ on AfriSpeech) under three loop-flag thresholds. BH-sig is the count of accents (out of 62 with $n \geq 30$) reaching $q < 0.05$ in two-proportion z-tests with Benjamini-Hochberg correction. CI is the smoothed loop-rate MMR compounding index. Aggregate loop rate decreases as the threshold tightens, but the FP16-to-HQQ contrast and the per-accent ranking are preserved.

Accent (n)	Loose	Baseline	Strict
Kanuri (66)	$4.00 \times^*$	$6.67 \times^*$	$6.33 \times^*$
Yoruba (648)	$3.50 \times^*$	$5.12 \times^*$	$5.00 \times^*$
Hausa (196)	$4.75 \times^*$	$6.00 \times^*$	$5.00 \times^*$
Swahili (521)	$2.71 \times$	$3.40 \times$	$4.25 \times^*$
Igbo (355)	$2.20 \times$	$2.83 \times$	$2.83 \times$

Table 4: INT4 HQQ over FP16 loop-rate ratio under *whisper-tiny* on AfriSpeech, for the five headline West African accents at each loop-flag setting. $\star$ marks $q < 0.05$ after BH-FDR correction within the 62-accent slate. The three accents driving the H2 claim (Kanuri, Yoruba, Hausa) are significant at every setting.

deterministic-kernel flags, fixed random seeds, and a pinned environment. Two reruns of the same 100 utterance ids on Whisper-large-v3 + FP16 produced bit-identical hypotheses on the same hardware.

E Excluded compression methods

AWQ (Lin et al., 2024) is excluded because its reference library has been deprecated and its successor pipeline is tied to a serving runtime without first-class encoder-decoder support. GPTQ (Frantar et al., 2023) and SparseGPT (Frantar and Alistarh, 2023) are excluded because their reference implementations expose per-architecture classes that do not currently include any encoder-decoder speech

model; implementing a Whisper adapter for either method would require non-trivial new infrastructure (per-block calibration walks across encoder, decoder, and cross-attention) and would not add a contrast beyond the data-free 4-bit recipes (NF4 and HQQ) already evaluated. Wanda 2:4 structured pruning was attempted on *whisper-small* and produced catastrophic outputs on a 20-utterance test with 32 calibration utterances (WER 194%), incompatible with the narrow attention-head dimensions of Whisper, and is therefore dropped from the main sweep.

In every case the blocker is the reference implementation rather than the algorithm. The specific blockers we hit were: `gptqmodel` 7.0 lists no encoder-decoder architecture among its supported model classes and requires a transformers major-version upgrade that would invalidate our pinned determinism baseline; `auto-gptq` does not build against the pinned torch and Python versions; and `autoawq` fails to import against transformers 4.57 and has been deprecated in favor of a serving-runtime pipeline. A Whisper adapter for any of these would need per-block calibration walks across the encoder, the decoder, and the decoder cross-attention projections. We note for future work that torchao exposes a module-level configuration path for AWQ that does not require a per-architecture class, which we did not pursue for this version.

F Granite-4.0 generalization probe

To check whether the INT8 portion of the quantization findings generalizes beyond Whisper, we evaluate IBM Granite-4.0-1b-speech (a 1B-parameter speech model with architecture distinct from Whisper) on Fair-Speech at FP16 and INT8. The probe is appendix-only and cannot test the main claims of this paper: pruning was not applied to Granite, and INT4 was not evaluated, so the probe speaks only to the INT8 sub-claim of Section 5 and to the baseline demographic-gap structure.

Granite-Speech is less accurate than Whisper-large-v3 at FP16 on Fair-Speech (utterance-weighted mean WER including loop-flagged hypotheses, 10.72% vs 6.10%), but the per-ethnicity gap structure is essentially identical (per-group figures below, as everywhere in the paper, are loop-filtered corpus WER): Black/AA speakers carry the highest WER (14.79% on Granite, 9.78% on Whisper-large-v3), Asian speakers the low-

est (5.13% on Granite, 3.27% on Whisper-large-v3), and the FP16 max-min ratios across ethnicity groups are within 0.1 of each other (Granite MMR = 2.88; Whisper-large-v3 MMR = 2.99). The shape of the demographic disparity therefore survives a substantial architectural change, which suggests the baseline gap is not a Whisper-specific training artifact and is more likely driven by training-data distribution shared across speech models or by Fair-Speech’s own demographic composition.

INT8 quantization preserves both the aggregate accuracy and the demographic-gap structure on Granite. Aggregate WER moves from 10.72% to 11.02%, the MMR shifts from 2.883 to 2.896 (compounding index +0.004), loop rate is unchanged at 0.45%, and the largest per-ethnicity WER delta is +0.51 pp on Middle Eastern speakers ($n = 747$). This is consistent with the Whisper INT8 result reported in Section 5, where INT8 produced 7 BH-significant per-group deltas out of 427 across all three datasets, none of them on Fair-Speech ethnicity groups larger than +0.22 pp (Appendix A). The INT8-is-gentle finding therefore generalizes to one non-Whisper architecture. The probe does not adjudicate whether the pruning or INT4 findings generalize; cross-family experiments on those axes are left to future work.

TurkSV-Bench: Assessing Demographic Bias and Preprocessing Effects in Speaker Verification

Ahmet Berkay Kantarcı
Ostim Technical University
220172022@ostim.teknik.edu.tr

Abdulkadir Parlak
abdulkadirparlak3@gmail.com

Ahmet Karazor
TURKSAT
ahmet.karazor@turksat.com.tr

Abstract

In this study, TurkSV-Bench is introduced as a novel benchmark dataset to evaluate the zero-shot cross-lingual performance and demographic biases of Turkish speaker verification (SV) systems. Created via an automated pipeline from YouTube data, this dataset comprises 1,844 audio recordings from 461 speakers representing diverse age groups, genders, and regional dialects. Evaluation of existing models identifies ResNet34-LM as the most successful baseline architecture. Subgroup analyses demonstrate the models' generalization capabilities across different gender and age demographics, while revealing distinct performance variations among regional dialects. Comprehensive evaluations of various preprocessing techniques highlight the baseline architecture's structural robustness, identifying specific configurations that maintain or refine verification performance without requiring aggressive signal filtering.

1 Introduction

Speaker verification (SV) is a biometric task that verifies whether a given voice recording belongs to a claimed identity by comparing it with a previously recorded reference sample. Unlike speaker identification, SV does not aim to recognize a specific speaker, but rather determines whether two voice samples belong to the same person. Today, SV systems are widely used in applications such as telephone banking, call center identity verification, smart assistants, mobile devices, access control, and digital forensics.

With the development of deep learning methods, embedding-based architectures have become the standard, transforming speaker-specific discriminative features into low-dimensional representation vectors and achieving high success on benchmark datasets such as VoxCeleb (Nagrani et al., 2017) and CN-Celeb (Fan et al., 2020).

However, the vast majority of existing speaker verification models are trained on English or multilingual datasets, making their application to Turkish speech inherently a cross-lingual transfer problem. Turkish differs significantly from these training languages in terms of phonetic structure, agglutinative characteristics, speech rhythm, and rich regional dialect diversity. Despite this, comprehensive studies systematically examining the performance of state-of-the-art speaker verification models on Turkish remain quite limited, making the evaluation of existing models' generalizability to Turkish speech an important research problem.

Another significant shortcoming in Turkish speaker verification research is the lack of a comprehensive and reliable benchmark dataset, as existing datasets are limited in terms of age, gender, regional dialects, and demographic diversity, making it difficult to systematically evaluate model performance across different user groups.

In this study, we develop an end-to-end automated data collection and processing pipeline for Turkish speaker verification and use it to create a new benchmark dataset, named as TurkSV-Bench. The system automatically collects audio recordings from publicly available YouTube videos and playlists, performing voice activity detection, speaker diarization, removal of overlapping speech, quality control, and data management. The dataset accounts for varying age groups, genders, and regional dialect diversity across Turkey, yielding a traceable benchmark for analyzing both speaker verification performance and potential demographic biases.

Rather than proposing a new speaker verification model, this study aims to systematically evaluate state-of-the-art models by establishing a comprehensive benchmark for Turkish. Widely used speaker verification models from the literature are compared on the Turkish dataset under a common experimental protocol to identify the most success-

ful base model, thereby analyzing the zero-shot cross-lingual performance of models trained on different languages when applied to Turkish.

In the second stage, a comprehensive preprocessing analysis is conducted on the identified base model, evaluating methods such as normalization, pre-emphasis, windowing, noise reduction, and voice activity detection, both individually and in combination. The primary objective is to identify preprocessing strategies that reduce the False Accept (FA) rate while increasing similarity between reference and test audios, as well as to systematically examine how multi-stage filtering chains strengthen or weaken speaker-specific discriminative features.

2 Related Work

Speaker verification systems have made significant progress in text-independent tasks with the adoption of deep learning approaches. In particular, the transition from traditional vector-based methods to CNN and Transformer-based architectures has greatly increased the robustness of models under different acoustic conditions (Sang et al., 2023).

The difference between the language on which models are trained and the language on which they are tested brings about cross-lingual transfer problems. Recent studies in the literature systematically examining the cross-lingual transfer capabilities of different SV architectures show that model design plays a critical role in obtaining language-independent speaker representations (Buitrago and Hernando, 2026).

Besides model architectures, preprocessing techniques applied to speech signals also have a significant role on system performance. Various studies emphasize that steps such as voice activity detection (VAD), silence removal, noise reduction, and loudness normalization can positively affect the quality of the obtained embeddings and thus the verification performance (Labied et al., 2022).

The equal and fair performance of speaker verification systems across all demographic groups is one of the most critical issues that speech research has focused on in recent years. It reveals that system performance can show significant variations according to demographic characteristics such as age, gender, and dialect, and that this situation can lead to bias in models (Hutiri and Ding, 2022).

Large-scale benchmark datasets such as VoxCeleb play a central role in the advancement of

speaker verification research and the standardized comparison of different methods. These datasets offer consistent evaluation protocols to the community to measure the robustness of models under different recording conditions and populations (Nagrani et al., 2017). In the literature, a study focused on the effect of language mismatch in Turkish speaker verification using a dataset of 73 speakers (47 male, 26 female), showing that language mismatch between training and test data significantly degrades performance (Hanilçi, 2017).

3 Methodology

3.1 Dataset

In this study, a novel benchmark dataset was created for use in Turkish speaker verification research. The dataset was designed to offer a more comprehensive and distinct structure compared to existing speaker verification datasets in the literature in terms of demographic diversity and regional dialect coverage. Thus, it provides not only a benchmark for measuring the general verification performance of models but also enables the investigation of potential biases associated with performance variations across different demographic structures.

Category	Speakers	Recordings	%
Total	461	1,844	100
<i>Gender</i>			
Male	310	1,240	67
Female	151	604	33
<i>Age</i>			
Adult (< 50)	293	1,172	63.6
Elder (≥ 50)	168	672	36.4
<i>Regional Dialect</i>			
Istanbul	381	1,524	83
Black Sea	24	96	5
Central Anatolia	21	84	4
Eastern	19	76	4
Thrace	8	32	2
Aegean	8	32	2

Table 1: Dataset statistics.

TurkSV-Bench consists of a total of 1,844 distinct voice recordings belonging to 461 different speakers. Each speaker is represented by four fixed-length voice samples extracted from different and non-overlapping time intervals of the YouTube video from which the audio was sourced. Maintaining four recordings per speaker allows for the creation of multiple genuine verification pairs among different samples belonging to the same speaker and enables a more reliable measurement of intra-speaker variability.

In the construction of the dataset, gender, age, and regional dialect distributions were considered jointly to support subgroup-based evaluation and

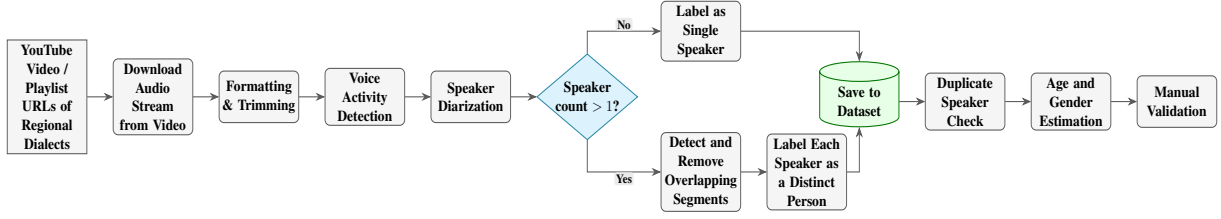


Figure 1: Pipeline for building the speaker verification benchmark dataset.

increase representational diversity. The dataset includes female and male speakers; adult, and elder age groups; and speakers representing Istanbul, Black Sea, Central Anatolia, Eastern, Thrace, and Aegean regional dialect categories. The resulting demographic and regional distribution is summarized in Table 1.

Field	Description
person_id	Speaker identity (P1, P2, ...)
data_id	Recording identity (P1_D1, ..., P1_D4)
url	Source video link
start / end	Start and end times in the video
person_number	Number of detected speakers
gender	Gender label
age_group	Age group (adult/elder)
region	Regional dialect label

Table 2: Metadatas stored for each audio recording.

Each recording is stored together with the metadata fields defined in Table 2.

The end-to-end pipeline followed in creating the benchmark dataset is shown in Figure 1. The pre-processing and quality control steps applied to the created dataset include the following:

Only the audio stream was extracted from the source videos; the recordings were converted to 16 kHz, mono, and WAV format. To increase the probability of excluding credits, music, advertisements, and promotional sections frequently found at the beginning of the videos, the first 60 seconds were not processed. The processing interval was limited to the 1000th second to ensure obtaining data of sufficient length from each person, considering the possibility of multiple people in a single video. Voice activity detection (VAD) was performed using Silero VAD (Silero Team, 2021), silent sections were removed, and only speech-containing regions were preserved. Speaker diarization was carried out with the pyannote.audio 3.1 pipeline (Plaquet and Bredin, 2023; Bredin, 2023) to determine the time intervals of the speakers, overlapping speech was cleaned, and each speaker was recorded with a separate identity. A 14-second segment was extracted from each recording; to eliminate overlaps caused by speaker transitions and clipped words concentrated at the edges, one second was further trimmed from each end, resulting in the final 12-

second recording.

To identify duplicate speakers semi-automatically, speaker embeddings extracted using an off-the-shelf ECAPA-TDNN model (Desplanques et al., 2020) trained on VoxCeleb were compared pairwise, and identities showing high similarity were manually examined and removed from the dataset. The age group and gender labels of the speakers were predicted using a wav2vec 2.0-based age and gender estimation model (Burkhardt et al., 2023). Following the automated processing stages, a manual quality-control procedure was applied to all recordings. Each speaker in TurkSV-Bench is represented by four 12-second recordings, and all four recordings associated with every speaker identity were listened to individually. This review was used to verify consistency among the recordings assigned to the same speaker and to identify residual errors that could remain after speaker diarization, overlap removal, and automatic metadata estimation. The automatically predicted age-group and gender labels were also subjected to human verification.

3.2 Models

The fundamental architectural features of current deep learning-based models, which are examined within the scope of the study and stand out in the literature for the speaker verification task, are summarized.

WeSpeaker ResNet34-LM is a ResNet34-based speaker verification model that extracts speaker-specific discriminative features from the audio signal. The model learns local patterns in the time and frequency axes by processing log-Mel filterbank features with two-dimensional convolutional layers. The model, which has approximately 6.63 million parameters, is trained on the VoxCeleb2-dev dataset, and large-margin fine-tuning is applied to the model to better separate different speakers in the embedding space (Wang et al., 2023a).

ECAPA-TDNN is a speaker verification model that improves the x-vector architecture with chan-

nel attention mechanisms and multi-scale feature aggregation. The model combines layer outputs at different depths using Res2Net-based blocks; it generates fixed-dimensional embeddings from variable-length audios through attentive statistics pooling. The model, with approximately 20.8 million parameters, is trained on the VoxCeleb1 and VoxCeleb2 datasets (Desplanques et al., 2020).

CAM++ is a densely connected TDNN (D-TDNN) based speaker verification model developed to extract speaker-specific information. The model processes time-frequency details with a two-dimensional convolutional front-end and aims to highlight speaker-related features and reduce the effect of noise through a context-aware masking mechanism. The model, which has a multi-scale pooling structure, contains approximately 7.18 million parameters (Wang et al., 2023b).

The ERes2Net family consists of speaker verification models that enhance the multi-scale structure of Res2Net with attention-based feature fusion mechanisms. The base ERes2Net aims to produce more discriminative speaker embeddings by combining local details within residual blocks and information from different network stages. ERes2NetV2 is a model with 17.8 million parameters that increases computational efficiency with fusion components that simplify the structure (Chen et al., 2024). ERes2Net-large is the higher-capacity version of the base structure with 22.46 million parameters (Chen et al., 2023).

TitaNet-Large is a ContextNet-based model developed for speaker verification. The model learns local and global speaker-specific features by combining one-dimensional depthwise separable convolutions with Squeeze-and-Excitation layers; it transforms variable-length recordings into 192-dimensional t-vector embeddings using channel-attentive statistical pooling. It is a model with approximately 23 million parameters (Koluguri et al., 2022).

ESPnet ECAPA-WavLM combines the self-supervised speech representations of WavLM-Large with an ECAPA-TDNN speaker encoder: the WavLM front-end extracts contextual and speaker-specific information from raw audio, which ECAPA-TDNN transforms into discriminative speaker embeddings. The model is trained on the VoxCeleb1 and VoxCeleb2 development sets (Jung et al., 2024).

WavLM-MHFA combines the self-supervised speech representations of WavLM-Base+ with a

Model	ROCCH-		FAR@		TO-HTER		ROCCH-	
	TO-HTER	EER	minDCF	FRR=0	FA FR	E[FA]	E[FR]	
ResNet34-LM	0.0055	0.0107	0.0091	0.0110	186 0	181.3	0.30	
ERes2NetV2	0.0598	0.0635	0.0122	0.1812	801 2	1077.5	1.76	
ECAPA-TDNN	0.0631	0.0652	0.0146	0.1676	915 2	1105.5	1.80	
ERes2Net	0.0732	0.0612	0.0117	0.1476	1256 2	1037.9	1.69	
TitaNet-Large	0.0812	0.0591	0.0100	0.2045	1529 2	1002.2	1.63	
ERes2Net-Large	0.1802	0.1659	0.0226	0.8196	3048 5	2814.6	4.59	
WavLM-MHFA	0.1803	0.1714	0.0153	1.3176	3052 5	2908.5	4.74	
ECAPA-WavLM	0.2227	0.2137	0.0193	1.5746	3876 6	3625.6	5.91	
CAM++	0.2807	0.2470	0.0213	3.4263	4618 8	4190.0	6.83	

Table 3: Benchmark results for speaker verification models on TurkSV-Bench.

Multi-Head Factorized Attentive Pooling (MHFA) layer, which fuses information from different Transformer layers via multiple attention heads to enhance speaker discrimination. With its 64-head structure, the model contains 96.2 million parameters (Peng et al., 2023).

3.3 Preprocessing Methods

Prior to the embedding extraction stage, various preprocessing steps described below are performed in order to reduce systemic or environmental distorting effects in the speech signals and to make the recordings optimally suited for the network architecture.

Peak Normalization: All samples are scaled to a specific range by taking the absolute highest amplitude value of the signal as a reference. This fundamental operation aims to balance the general volume differences among different audio recordings.

RMS Normalization: The root mean square energy of the signal is calculated and adjusted to a target level. Since it is based on the average energy rather than peak values, it provides a more robust and consistent normalization against sudden amplitude spikes in the speech signal.

Amplitude Scaling: Preserving the original morphological features of the signal, the amplitude values are multiplied linearly by a constant coefficient. It is a fundamental operation that only adjusts the magnitude of the signal and does not cause structural distortion.

Loudness Normalization (LUFS): In this method, the signal is brought to a perceptual target loudness level (-23 LUFS in this study) by taking the frequency sensitivity of the human auditory system into account. Unlike energy-based approaches, it aims to increase perceptual consistency.

DC Offset Removal: By bringing the arithmetic mean of the signal to zero, DC offsets that may originate from recording devices or the audio chain are cleaned. Thus, low-frequency artifact compo-

nents are prevented from negatively affecting the feature extraction stage.

Pre-emphasis: A first-order high-pass filter is applied to balance the natural spectral tilt of the speech signal. This operation makes speaker-specific formant information in higher frequency bands more prominent. In the study, the values 0.95, 0.97, and 0.99 are examined comparatively to analyze the effect of the α coefficient on performance.

Cepstral Mean Normalization (CMN): Used to suppress static spectral distortions from the channel and recording environment. Since models use end-to-end feature extraction, a waveform-level equivalent is designed by resynthesizing the signal after subtracting the band-based logarithmic amplitude mean in the STFT domain.

Windowing: Hamming, Hanning, and Blackman window functions are tested to minimize spectral leakage and smooth discontinuities at frame boundaries. The signal is windowed by dividing it into frames of 25 ms length with a 10 ms shift, and then the waveform is reconstructed using the overlap-add method.

Spectral Subtraction: The noise spectrum estimated from the silent (low-energy) sections of the recording is subtracted from the overall amplitude spectrum of the signal. In this way, stationary background noise is suppressed, ensuring that speaker features are obtained more clearly.

Wiener Filtering: An adaptive filtering is performed with the Minimum Mean Square Error (MMSE) criterion, based on the statistical properties of the signal and noise. This method aims to directly improve the speaker verification performance by increasing the Signal-to-Noise Ratio (SNR).

4 Experiments

The experimental studies were conducted on the TurkSV-Bench dataset detailed in Section 3.1, comprising 1,844 speech recordings from 461 speakers. A total of 1,699,246 evaluation trials were constructed, consisting of 2,766 genuine and 1,696,480 impostor trials.

System performance was evaluated using Threshold-Optimized Half Total Error Rate (TO-HTER), ROC Convex Hull Equal Error Rate (ROCCH-EER), the minimum Detection Cost Function (minDCF), and FAR@FRR=0. False Accept (FA) and False Reject (FR) counts were ad-

ditionally reported at the corresponding operating points.

Cosine similarity between the speaker embedding vectors extracted from each pair of speech samples was used as the verification score. For a decision threshold θ , a pair was accepted as belonging to the same speaker when its cosine similarity score was greater than or equal to θ , and rejected otherwise. Accordingly, a false accept (FA) occurs when an impostor pair is incorrectly accepted, whereas a false reject (FR) occurs when a genuine pair is incorrectly rejected. FAR and FRR were calculated as the percentages of FA and FR relative to the total number of impostor and genuine pairs, respectively, at each decision threshold. We additionally report FAR@FRR=0, defined as the false acceptance rate at the strictest empirically attainable decision threshold for which no genuine trial is falsely rejected.

At a decision threshold θ , HTER is computed as the arithmetic mean of FAR(θ) and FRR(θ) (Bengio and Mariéthoz, 2004). The empirical TO-HTER operating point is selected by minimizing $|FAR(\theta) - FRR(\theta)|$ over the available evaluation thresholds.

Equal-error performance is additionally characterized using ROCCH-EER, obtained from the intersection of the ROC convex hull with the FAR = FRR diagonal (Brümmer and de Villiers, 2013). This formulation provides an equal-error estimate when the discrete empirical operating points do not contain an exact equality between FAR and FRR, which is particularly relevant in low-error operating regions.

The minDCF is additionally reported as the minimum application-weighted detection cost over the evaluated decision thresholds.

Observed errors at TO-HTER are denoted as FA@TO-HTER and FR@TO-HTER. Expected errors at ROCCH-EER are reported as E[FA]@ROCCH-EER and E[FR]@ROCCH-EER, which can be non-integers since this operating point may lie between empirically attainable points.

The evaluation consists of two stages. In the first stage, the speaker verification models introduced in Section 3.2 were compared to identify the architecture yielding the most discriminative representation, establishing it as the baseline. For this stage, overall results computed across the entire dataset are presented in Table 3, while performance broken down by gender, age, and regional dialect

is reported in Table 4.

In the second stage, using this baseline model, we analyzed the impact of the signal preprocessing techniques described in Section 3.3 on verification performance. Similarly, overall results computed across the entire dataset are presented in Table 5, while the corresponding breakdown by gender, age, and regional dialect is reported in Table 6.

As part of the preprocessing analysis, all possible pairwise and triple-wise chain filter combinations of the individual filters examined in this study were tested. Due to space constraints and readability, the results of all combinations could not be presented in detail; only the most successful filter chains are reported in this paper.

5 Results

The evaluated pretrained speaker verification systems demonstrated varying levels of zero-shot generalization to Turkish speech. Among the models reported in Table 3, ResNet34-LM achieved the strongest overall performance, with a TO-HTER of 0.0055%, a ROCCH-EER of 0.0107%, and a minDCF of 0.0091. ResNet34-LM was therefore used as the baseline architecture for the subsequent preprocessing analyses.

The demographic subgroup results in Table 4 show relatively small differences between the male and female subsets for ResNet34-LM. The model yielded TO-HTER values of 0.0100% for male speakers and 0.0080% for female speakers, with corresponding ROCCH-EER values of 0.0187% and 0.0152%, respectively. For the Adult and Elder age groups reported in the subgroup comparison, TO-HTER values were 0.0076% and 0.0116%, while ROCCH-EER values were 0.0148% and 0.0217%, respectively. A larger separation was observed across regional-dialect groups. The Istanbul subset yielded a TO-HTER of 0.0057% and a ROCCH-EER of 0.0112%, whereas the Other Dialects subset yielded 0.0356% and 0.0656%, respectively.

Table 5 summarizes the overall preprocessing results obtained with ResNet34-LM. Among the individual preprocessing methods, pre-emphasis with $\alpha = 0.97$ yielded a TO-HTER of 0.0051% and a ROCCH-EER of 0.0096%, while CMN yielded 0.0052% and 0.0098%, respectively. The selected pairwise and triple-filter configurations remained within a similarly narrow low-error range. In contrast, spectral subtraction and Wiener filtering pro-

duced substantially higher error rates across the reported evaluation metrics.

The subgroup preprocessing results in Table 6 further show that the effect of preprocessing depends on both subgroup and evaluation metric. For male speakers, the baseline TO-HTER of 0.0100% changed to 0.0078% with pre-emphasis at $\alpha = 0.99$, while for female speakers the baseline value of 0.0080% changed to 0.0077% with pre-emphasis at $\alpha = 0.97$. For Adult speakers, the baseline TO-HTER of 0.0076% and ROCCH-EER of 0.0148% decreased to 0.0075% and 0.0142%, respectively, with either CMN or pre-emphasis at $\alpha = 0.97$. Pre-emphasis at $\alpha = 0.99$ reduced Adult minDCF from 0.0129 to 0.0117, but increased TO-HTER to 0.0080%. For Elder speakers, the baseline TO-HTER of 0.0116% changed to 0.0082% with either CMN or pre-emphasis at $\alpha = 0.99$. For the Other Dialects subset, the baseline TO-HTER of 0.0356% changed to 0.0297% with CMN and to 0.0287% for the reported Blackman+CMN-based pairwise and triple-filter configurations. These changes reflect the effects observed at the corresponding operating points, which varied across the evaluated preprocessing configurations. Spectral subtraction and Wiener filtering, in contrast, consistently produced substantially higher error rates across the demographic and regional-dialect subsets.

6 Discussion and Conclusion

In this study, we present an end-to-end data generation pipeline that automates the process from data collection to quality control. Using this infrastructure, we benchmark a diverse set of pretrained speaker-verification models, evaluating their zero-shot generalization to Turkish speech as well as the effects of several speech preprocessing methods.

The model comparison shows that performance on Turkish speech varies substantially across architectures. ResNet34-LM achieved the strongest overall results despite having considerably fewer parameters than several competing systems. This suggests that model capacity alone does not determine cross-lingual speaker-verification performance and that architectural design and training strategy may play an important role in generalization to Turkish speech.

From a demographic perspective, gender- and age-related differences were relatively limited and varied across models. In contrast, more noticeable differences were observed between the Istanbul

and Other Dialects subsets, with the Istanbul subset generally yielding lower error rates across the evaluated systems.

The preprocessing experiments show that most methods produced only small changes in overall verification performance. The observed improvements were limited to specific preprocessing methods and evaluation metrics, while the results varied across demographic subgroups. Pre-emphasis, CMN, and several filter combinations produced

small improvements in some evaluation metrics. These findings indicate that the effect of preprocessing was dependent on the evaluation setting, with some configurations providing modest improvements for certain demographic subgroups. In contrast, spectral subtraction and Wiener filtering resulted in consistently higher error rates, indicating that these more intensive noise-suppression methods were less suitable for the speaker-recognition conditions considered in this study.

(a) Gender												
Model	Male					Female						
	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
ResNet34-LM	0.0100	0.0187	0.0166	0.0200	153 0	143.59 0.35	0.0080	0.0152	0.0066	0.0160	29 0	27.55 0.14
ERes2NetV2	0.0628	0.0402	0.0153	0.0718	550 1	308.30 0.75	0.2707	0.1727	0.0110	0.5359	581 2	312.97 1.56
ECAPA-TDNN	0.0550	0.0463	0.0200	0.0826	431 1	354.82 0.86	0.2889	0.1641	0.0154	0.5033	647 2	297.31 1.49
ERes2Net	0.0579	0.0591	0.0170	0.1254	475 1	452.81 1.10	0.1209	0.1218	0.0110	0.2949	238 1	220.77 1.10
TitaNet-Large	0.0838	0.0875	0.0158	0.3003	460 2	670.21 1.63	0.0952	0.0511	0.0110	0.0800	145 1	92.65 0.46
ERes2Net-Large	0.2183	0.2025	0.0273	0.5280	1698 4	1551.48 3.77	0.3237	0.2581	0.0270	1.4134	573 3	467.71 2.34
WavLM-MHFA	0.1615	0.1327	0.0200	1.0041	1239 3	1016.78 2.47	0.4821	0.3966	0.0225	4.0121	947 4	718.63 3.59
ECAPA-WavLM	0.2190	0.2014	0.0225	2.2620	1709 4	1543.65 3.75	0.4818	0.3710	0.0368	4.5817	946 4	672.31 3.36
CAM++	0.3747	0.3132	0.0253	5.7355	2859 7	2400.19 5.83	0.4418	0.3469	0.0199	1.0822	801 4	628.64 3.14

(b) Age												
Model	Adult (< 50)					Elder (≥ 50)						
	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
ResNet34-LM	0.0076	0.0148	0.0129	0.0152	104 0	101.02 0.26	0.0116	0.0217	0.0148	0.0232	52 0	48.72 0.22
ERes2NetV2	0.1146	0.0785	0.0106	0.1921	790 2	537.32 1.38	0.1026	0.0689	0.0277	0.1060	238 1	154.65 0.69
ECAPA-TDNN	0.1268	0.0895	0.0191	0.1893	957 2	612.47 1.57	0.0172	0.0320	0.0180	0.0343	77 0	71.84 0.32
ERes2Net	0.0666	0.0699	0.0113	1.0359	523 1	478.39 1.23	0.0964	0.0668	0.0280	0.0936	210 1	150.03 0.67
TitaNet-Large	0.1239	0.0854	0.0150	0.2376	918 2	584.74 1.50	0.0080	0.0158	0.0146	0.0160	36 0	35.52 0.16
ERes2Net-Large	0.1188	0.1206	0.0214	0.2444	847 2	825.64 2.12	0.2767	0.2858	0.0362	1.6717	574 3	641.46 2.88
WavLM-MHFA	0.2908	0.2579	0.0227	1.6542	2034 5	1765.10 4.53	0.0207	0.0354	0.0200	0.0414	93 0	79.44 0.36
ECAPA-WavLM	0.3395	0.3241	0.0285	1.9150	2311 6	2218.11 5.70	0.0187	0.0354	0.0235	0.0374	84 0	79.37 0.36
CAM++	0.1856	0.1353	0.0188	0.5103	1373 3	926.01 2.38	0.5939	0.5119	0.0333	6.5663	1330 6	1149.06 5.16

(c) Regional Dialect												
Model	Istanbul					Other Dialects						
	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
ResNet34-LM	0.0057	0.0112	0.0098	0.0114	132 0	129.28 0.26	0.0356	0.0656	0.0397	0.0712	36 0	33.17 0.31
ERes2NetV2	0.0486	0.0312	0.0068	0.0534	618 1	361.71 0.71	0.5149	0.2908	0.0703	0.6290	310 2	147.02 1.40
ECAPA-TDNN	0.0411	0.0400	0.0110	0.1874	445 1	463.41 0.91	0.3923	0.2926	0.0592	0.6052	186 2	147.96 1.40
ERes2Net	0.0055	0.0107	0.0065	0.0111	128 0	123.84 0.24	0.4170	0.3423	0.0731	4.4086	211 2	173.06 1.64
TitaNet-Large	0.0934	0.0690	0.0109	0.2261	1150 2	799.39 1.58	0.2031	0.1175	0.0361	0.1978	100 1	59.40 0.56
ERes2Net-Large	0.0912	0.0869	0.0170	0.2314	1100 2	1007.04 1.99	0.6566	0.6148	0.0779	1.7880	348 3	310.84 2.95
WavLM-MHFA	0.1237	0.1216	0.0159	0.4029	1345 3	1408.25 2.78	0.7812	0.4998	0.0460	1.7820	474 3	252.69 2.40
ECAPA-WavLM	0.1317	0.0996	0.0194	0.2471	1531 3	1153.30 2.28	1.0005	0.7736	0.0522	1.8097	485 5	391.11 3.71
CAM++	0.1996	0.1561	0.0132	3.7873	2596 4	1807.83 3.57	0.8785	0.8067	0.0758	1.6594	467 4	407.89 3.87

Table 4: Benchmark results of models across (a) gender, (b) age, and (c) regional dialect subgroups on TurkSV-Bench.

Preprocessing Method	TO-HTER	ROCCH-EER	minDCF	FAR@FRR=0	TO-HTER FA FR	ROCCH-EER E[FA] E[FR]
Baseline	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Pre-emphasis ($\alpha = 0.99$)	0.0053	0.0098	0.0079	0.0106	180 0	166.43 0.27
CMN	0.0052	0.0098	0.0084	0.0104	176 0	167.01 0.27
Pre-emphasis ($\alpha = 0.97$)	0.0051	0.0096	0.0084	0.0102	173 0	163.41 0.27
Amplitude scaling	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
RMS normalization	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
DC offset removal	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Peak normalization	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Loudness normalization (-23 LUFS)	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Hamming windowing	0.0055	0.0107	0.0091	0.0110	186 0	181.27 0.30
Hann windowing	0.0055	0.0107	0.0091	0.0110	186 0	181.37 0.30
Blackman windowing	0.0055	0.0107	0.0091	0.0110	186 0	181.37 0.30
Spectral subtraction	1.0113	0.9782	0.1617	11.0728	17140 28	16594.97 27.06
Wiener filtering	2.0258	2.0219	0.1856	31.1352	34388 56	34301.13 55.93
Pairwise filter combinations						
Blackman + CMN	0.0051	0.0098	0.0083	0.0102	173 0	165.45 0.27
Hann + Pre-emphasis ($\alpha = 0.97$)	0.0051	0.0096	0.0084	0.0101	172 0	162.71 0.27
Triple filter combinations						
Blackman + Hann + CMN	0.0051	0.0098	0.0083	0.0102	173 0	165.45 0.27
RMS normalization + Hann + CMN	0.0051	0.0098	0.0083	0.0102	173 0	165.45 0.27

Table 5: Benchmark results of preprocessing methods applied with the ResNet34-LM model.

(a) Age												
Preprocessing Method	Adult (< 50)						Elder (≥ 50)					
	TO- HTER	minDCF	ROCCH- EER	FAR@ FRR=0	TO- HTER FA FR	ROCCH- EER E[FA] E[FR]	TO- HTER	minDCF	ROCCH- EER	FAR@ FRR=0	TO- HTER FA FR	ROCCH- EER E[FA] E[FR]
Baseline	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Peak normalization	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
RMS normalization	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Amplitude scaling	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
CMN	0.0075	0.0121	0.0142	0.0150	103 0	97.25 0.25	0.0082	0.0136	0.0160	0.0165	37 0	35.87 0.16
Loudness normalization	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
DC offset removal	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Pre-emphasis ($\alpha = 0.97$)	0.0075	0.0120	0.0142	0.0150	103 0	97.25 0.25	0.0085	0.0135	0.0163	0.0169	38 0	36.68 0.16
Pre-emphasis ($\alpha = 0.99$)	0.0080	0.0117	0.0147	0.0159	109 0	100.72 0.26	0.0082	0.0135	0.0160	0.0165	37 0	35.87 0.16
Hamming windowing	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Hann windowing	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Blackman windowing	0.0076	0.0129	0.0148	0.0152	104 0	101.02 0.26	0.0116	0.0148	0.0217	0.0232	52 0	48.72 0.22
Spectral subtraction	0.8581	0.1792	0.8424	4.5529	5907 15	5765.66 14.81	1.1903	0.1503	1.1303	10.5579	2671 12	2536.86 11.39
Wiener filtering	2.2054	0.2042	2.1762	29.4458	15005 39	14894.92 38.26	2.0893	0.1902	1.9439	14.0215	4703 21	4363.13 19.59
Pairwise filter combinations												
Blackman + CMN	0.0073	0.0121	0.0139	0.0146	100 0	95.11 0.24	0.0082	0.0140	0.0161	0.0165	37 0	36.03 0.16
Hann + Pre-emphasis												
($\alpha = 0.97$)	0.0075	0.0120	0.0142	0.0150	103 0	97.25 0.25	0.0085	0.0135	0.0163	0.0169	38 0	36.68 0.16
Triple filter combinations												
Blackman + Hann + CMN	0.0073	0.0121	0.0139	0.0146	100 0	95.11 0.24	0.0082	0.0140	0.0161	0.0165	37 0	36.03 0.16
RMS + Hann + CMN	0.0073	0.0121	0.0139	0.0146	100 0	95.11 0.24	0.0082	0.0140	0.0161	0.0165	37 0	36.03 0.16

(b) Gender												
Preprocessing Method	Male						Female					
	TO- HTER	minDCF	ROCCH- EER	FAR@ FRR=0	TO- HTER FA FR	ROCCH- EER E[FA] E[FR]	TO- HTER	minDCF	ROCCH- EER	FAR@ FRR=0	TO- HTER FA FR	ROCCH- EER E[FA] E[FR]
Baseline	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Peak normalization	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
RMS normalization	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Amplitude scaling	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
CMN	0.0083	0.0159	0.0163	0.0166	127 0	124.58 0.30	0.0086	0.0071	0.0156	0.0171	31 0	28.18 0.14
Loudness normalization	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
DC offset removal	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Pre-emphasis ($\alpha = 0.97$)	0.0082	0.0158	0.0161	0.0163	125 0	123.35 0.30	0.0077	0.0055	0.0142	0.0155	28 0	25.69 0.13
Pre-emphasis ($\alpha = 0.99$)	0.0078	0.0155	0.0155	0.0157	120 0	118.94 0.29	0.0088	0.0055	0.0159	0.0177	32 0	28.83 0.14
Hamming windowing	0.0100	0.0166	0.0187	0.0200	153 0	143.59 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Hann windowing	0.0100	0.0166	0.0188	0.0200	153 0	143.92 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Blackman windowing	0.0100	0.0166	0.0188	0.0200	153 0	143.92 0.35	0.0080	0.0066	0.0152	0.0160	29 0	27.55 0.14
Spectral subtraction	1.2311	0.1897	1.1736	13.6497	9392 23	8993.42 21.83	1.5494	0.2173	1.5007	27.2853	2815 14	2719.30 13.60
Wiener filtering	1.7192	0.1648	1.7046	47.2108	13165 32	13063.00 31.71	3.5055	0.2611	3.4075	47.1181	6304 32	6174.39 30.87
Pairwise filter combinations												
Blackman + CMN	0.0084	0.0157	0.0163	0.0167	128 0	124.67 0.30	0.0077	0.0071	0.0142	0.0155	28 0	25.81 0.13
Hann + Pre-emphasis												
($\alpha = 0.97$)	0.0082	0.0157	0.0161	0.0163	125 0	123.35 0.30	0.0077	0.0055	0.0142	0.0155	28 0	25.69 0.13
Triple filter combinations												
Blackman + Hann + CMN	0.0084	0.0157	0.0163	0.0167	128 0	124.67 0.30	0.0077	0.0071	0.0142	0.0155	28 0	25.81 0.13
RMS + Hann + CMN	0.0084	0.0157	0.0163	0.0167	128 0	124.67 0.30	0.0077	0.0071	0.0142	0.0155	28 0	25.81 0.13

(c) Regional Dialect												
Preprocessing Method	Istanbul						Other Dialects					
	TO- HTER	minDCF	ROCCH- EER	FAR@ FRR=0	TO- HTER FA FR	ROCCH- EER E[FA] E[FR]	TO- HTER	minDCF	ROCCH- EER	FAR@ FRR=0	TO- HTER FA FR	ROCCH- EER E[FA] E[FR]
Baseline	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Peak normalization	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
RMS normalization	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Amplitude scaling	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
CMN	0.0057	0.0093	0.0108	0.0113	131 0	125.31 0.25	0.0297	0.0441	0.0566	0.0593	30 0	28.64 0.27
Loudness normalization	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
DC offset removal	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Pre-emphasis ($\alpha = 0.97$)	0.0056	0.0089	0.0105	0.0111	129 0	122.02 0.24	0.0356	0.0402	0.0668	0.0712	36 0	33.76 0.32
Pre-emphasis ($\alpha = 0.99$)	0.0058	0.0086	0.0108	0.0117	135 0	125.12 0.25	0.0346	0.0402	0.0655	0.0692	35 0	33.11 0.31
Hamming windowing	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0356	0.0397	0.0656	0.0712	36 0	33.17 0.31
Hann windowing	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0346	0.0397	0.0643	0.0692	35 0	32.53 0.31
Blackman windowing	0.0057	0.0098	0.0112	0.0114	132 0	129.28 0.26	0.0346	0.0397	0.0643	0.0692	35 0	32.53 0.31
Spectral subtraction	0.9129	0.1611	0.8918	11.1952	10508 21	10329.61 20.39	0.6141	0.1217	0.5698	4.4165	305 3	288.09 2.74
Wiener filtering	2.2315	0.1940	2.1633	30.9427	25853 51	25056.17 49.45	2.0405	0.2101	1.9488	20.6112	1010 10	985.32 9.35
Pairwise filter combinations												
Blackman + CMN	0.0055	0.0093	0.0107	0.0111	128 0	123.61 0.24	0.0287	0.0420	0.0553	0.0574	29 0	27.94 0.27
Hann + Pre-emphasis												
($\alpha = 0.97$)	0.0055	0.0089	0.0105	0.0111	128 0	121.30 0.24	0.0356	0.0402	0.0668	0.0712	36 0	33.76 0.32
Triple filter combinations												
Blackman + Hann + CMN	0.0055	0.0093	0.0107	0.0111	128 0	123.61 0.24	0.0287	0.0420	0.0553	0.0574	29 0	27.94 0.27
RMS + Hann + CMN	0.0055	0.0093	0.0107	0.0111	128 0	123.61 0.24	0.0287	0.0420	0.0553	0.0574	29 0	27.94 0.27

Table 6: Preprocessing results across (a) age, (b) gender, and (c) regional dialect subgroups.

Limitations

Due to the nature of the data collection platform, TurkSV-Bench exhibits partial imbalances in the distribution of age, gender, and regional dialects. Future work can extend the benchmark with larger and more balanced subgroup distributions, and additional acoustic conditions to provide more robust estimates of real-world speaker-verification perfor-

mance. Additionally, as both reference and test recordings are extracted from the same source, the resulting performance metrics may be somewhat optimistic compared to true cross-session conditions.

Ethics Statement

The audio recordings collected within the scope of TurkSV-Bench have been anonymized and uti-

lized solely for research purposes. Due to personal data usage policies, raw audio files cannot be shared. However, the data collection methodology and metadata are provided in detail to researchers.

References

- Samy Bengio and Johnny Mariéthoz. 2004. [A statistical significance test for person authentication](#). In *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop*.
- Hervé Bredin. 2023. [pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe](#). In *Proc. INTERSPEECH 2023*, pages 1983–1987.
- Niko Brümmer and Edward de Villiers. 2013. [The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF](#). *arXiv preprint arXiv:1304.2865*.
- Pol Buitrago and Javier Hernando. 2026. [Towards language-agnostic speaker verification: A cross-lingual transfer study of architectures](#). In *Proc. Odyssey 2026*.
- Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn W. Schuller. 2023. [Speech-based age and gender prediction with transformers](#). In *Speech Communication – 15th ITG Conference*, pages 46–50, Aachen, Germany. VDE.
- Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. 2023. [An enhanced Res2Net with local and global feature fusion for speaker verification](#). In *Proc. INTERSPEECH 2023*, pages 2228–2232.
- Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. 2024. [ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency](#). In *Proc. INTERSPEECH 2024*, pages 3245–3249.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. [ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification](#). In *Proc. Interspeech 2020*, pages 3830–3834.
- Yue Fan, J. W. Kang, L. T. Li, K. C. Li, H. L. Chen, S. T. Cheng, P. Y. Zhang, Z. Y. Zhou, Y. Q. Cai, and Dong Wang. 2020. Cn-celeb: A challenging chinese speaker recognition dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7604–7608.
- Cemal Hanilçi. 2017. [Effect of language mismatch on turkish speaker verification](#). *Uludağ University Journal of The Faculty of Engineering*, 22(1):189–196.
- Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. [Bias in automated speaker recognition](#). In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 230–247.
- Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Alex Gichamba, Barry-John Theobald, Ahmed Hussien Abdelaziz, and Shinji Watanabe. 2024. [ESPnet-SPK: Full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models](#). In *Proc. INTERSPEECH 2024*.
- Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. [TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context](#). In *Proc. IEEE ICASSP 2022*.
- Maria Labied, Abdessamad Belangour, Mouad Banane, and Allae Erraissi. 2022. [An overview of automatic speech recognition preprocessing techniques](#). In *Proc. 2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pages 804–809.
- Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. [VoxCeleb: a large-scale speaker identification dataset](#). In *Proc. INTERSPEECH 2017*, pages 2616–2620.
- Junyi Peng, Oldřich Plchot, Themis Stafylakis, Ladislav Mošner, Lukáš Burget, and Jan Černocký. 2023. [An attention-based backend allowing efficient fine-tuning of Transformer models for speaker verification](#). In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 555–562.
- Alexis Plaquet and Hervé Bredin. 2023. [Powerset multi-class cross entropy loss for neural speaker diarization](#). In *Proc. INTERSPEECH 2023*, pages 3222–3226.
- Mufan Sang, Yong Zhao, Gang Liu, John H. L. Hansen, and Jian Wu. 2023. [Improving Transformer-based networks with locality for automatic speaker verification](#). In *Proc. ICASSP 2023*, pages 1–5.
- Silero Team. 2021. Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. <https://github.com/snakers4/silero-vad>.
- Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023a. [WeSpeaker: A research and production oriented speaker embedding learning toolkit](#). In *Proc. IEEE ICASSP 2023*.
- Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023b. [CAM++: A fast and efficient network for speaker verification using context-aware masking](#). In *Proc. INTERSPEECH 2023*, pages 5301–5305.

What Drives Dialectal Jailbreaks? An Ablation of Surface Form, Cultural Framing, and Strategy Banks

Qingyang Xu

Independent Researcher, Shanghai, China

qyxu1994@gmail.com

Abstract

Recent work suggests that obscure language registers can weaken large language model refusal behavior, especially when paired with black-box prompt optimization. It remains unclear whether failures stem from non-standard surface form, culturally grounded framing, or optimization over an expressive prompt-strategy space. We study Chinese registers by extending a classical-Chinese red-teaming framework to Shanghaiese and Cantonese and running a 36-cell ablation across surface forms, strategy-bank variants, and two target models. The main finding is corrective: dialectal surface form is neither necessary nor sufficient for high attack success. Non-optimized English, Mandarin, and naive dialect translations remain below 8% attack success rate, whereas all conditions that retain an optimizer-controlled strategy bank reach 98–100%. A culture-neutral generic strategy bank reaches the same ceiling at near-single-query cost, further indicating that strategy-bank expressiveness, rather than dialectal or cultural content, accounts for most of the observed effect. Dialect choice still affects query efficiency and response severity, and qualitative coding identifies recurring high-level mechanisms such as semantic glossing and procedural scaffolding. We qualify the absolute ceiling-level attack success rate by showing sensitivity to judge calibration and noting that the design does not fully separate one-shot strategy construction from iterative search.

1 Introduction

LLMs are increasingly deployed as multilingual assistants, but their safety mechanisms are rarely evaluated with the same linguistic breadth as their advertised capabilities. A model that robustly refuses a harmful English or Mandarin request may respond differently when the same intent is embedded in a low-resource language, a dialectal register, or a locally specific cultural frame. This gap matters both technically and socially: technically,

it reveals weak semantic generalization of safety alignment; socially, it risks creating unequal protection for speakers of underrepresented languages and varieties.

Prior work has documented cross-lingual jailbreak vulnerabilities in low-resource languages (Yong et al., 2023; Deng et al., 2024), automatic adversarial suffixes (Zou et al., 2023), and real-world prompt jailbreak communities (Shen et al., 2024). Most closely related to our study, Huang et al. (2026) show that classical Chinese—a concise, historically prestigious, and comparatively underrepresented written register—can bypass model safeguards when paired with black-box prompt optimization. That result raises two related questions. First, is the vulnerability specific to classical Chinese, or does it generalize to living Chinese regional varieties whose written forms, idioms, and cultural references are also underrepresented in safety training? Second, within such an attack, which component is actually doing the work: the dialectal surface form, the cultural framing, or the iterative optimizer over an attack-strategy space?

We investigate both questions through a red-team evaluation of Chinese dialectal distribution shifts. We use “dialect” in its common NLP sense for regional Sinitic varieties, while recognizing that Shanghaiese/Wu and Cantonese/Yue are linguistically rich varieties with distinct phonology, lexicon, and sociocultural histories. Unlike classical Chinese—a historical, largely written-only register no longer used in everyday speech—these are *living* varieties with millions of contemporary speakers, which is what makes gaps in their safety coverage practically consequential. Our goal is not to characterize speakers or communities as risky, but to test whether LLM safety mechanisms generalize across the contexts in which users actually interact with deployed systems.

We extend the CC-BOS framework (Huang et al., 2026)—a black-box jailbreak that uses a

population-based optimizer (fruit-fly optimization, FOA) to search over a bank of prompt-construction strategies, requiring only query access to the target—from classical Chinese to Shanghaiese and Cantonese, preserving the black-box optimization loop but replacing the register-specific prompt constructor with dialect-specific sociocultural framing modules and adding a two-stage translation pipeline. On top of this we design a six-condition ablation varying (i) the surface form (English, Mandarin, dialect) and (ii) which optimizer strategy dimensions are active, evaluated against GPT-4o and DeepSeek-Reasoner on a 50-prompt AdvBench subset (1800 evaluations). The attack and translation model is DeepSeek-Chat, the judge is GPT-4o, and we early-stop at $\tau = 80$ on a 0–120 scale.

This paper makes the following contributions:

1. We design a controlled 36-cell ablation that systematically varies the surface form and the cultural-framing strategy bank while holding the target models, judge, and evaluation budget fixed, and show that an expressive optimizer-controlled strategy bank dominates while the dialect surface form contributes only to efficiency and severity, not to binary success. (Our design does not separate iterative search from one-shot strategy-bank construction; we return to this in Section 8.)
2. We extend dialectal red-team evaluation for Chinese regional varieties from classical Chinese to Shanghaiese/Wu and Cantonese/Yue, and report that the dialect-framed attack reaches 100% ASR on both GPT-4o and DeepSeek-Reasoner under our judge and threshold.
3. We provide a safety-preserving qualitative taxonomy of successful living-dialect attacks, with reliability probed by an independent automated re-coder (median $\kappa = 0.71$ between the author and a hypothesis-blind language-model agent; Section G), identifying semantic glossing, procedural scaffolding, commerce/logistics metaphors, grey-zone authority frames, and performative social scripts as recurring mechanisms, and show via a non-FOA template-artifact control that these mechanisms are emergent under optimization rather than artifacts of the prompt template.
4. We articulate a safety-preserving reporting protocol for dialectal jailbreak research: pub-

lish aggregate metrics, failure modes, and defenses, while withholding optimized adversarial prompts and harmful target responses.¹

2 Related Work

Jailbreak attacks and red teaming. LLM jailbreak research studies prompts that induce models to violate intended safety policies. Early work characterized real-world jailbreak prompts and their strategies, including role play, privilege escalation, and prompt injection (Shen et al., 2024). Automated methods search over adversarial strings or prompt transformations, such as gradient-guided suffix optimization (Zou et al., 2023), red-teaming with language models (Ganguli et al., 2022; Perez et al., 2022), and benchmark-driven safety and jailbreak evaluation (Gehman et al., 2020; Chao et al., 2023; Souly et al., 2024). A closely related line optimizes *readable* prompts or whole templates rather than suffixes: genetic-algorithm search over handcrafted prompts (Liu et al., 2024), fuzzing-style mutation of seed jailbreak templates (Yu et al., 2023), tree-of-attack refinement (Mehrotra et al., 2024), and encoding-based transformations such as ciphering (Yuan et al., 2024). These methods already show that optimization or mutation over a template/strategy space can break GPT-class models, so a key part of our contribution is the corrective finding that, on top of such an optimizer, the dialectal surface form and its cultural grounding add essentially nothing to binary success. Our work follows this empirical safety-evaluation tradition but focuses on a specific source of distribution shift: Chinese dialectal and cultural register, and the interaction between that shift and an optimizer-driven attack-strategy bank.

Multilingual safety alignment. Modern LLMs acquire refusal behavior through alignment techniques such as reinforcement learning from human feedback (Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022), but these safeguards are trained mostly on high-resource languages and often fail to generalize: Yong et al. (2023) show that translating unsafe prompts into low-resource languages bypasses GPT-4 safeguards, and Deng et al. (2024) find unsafe behavior increasing as language resource availability decreases, with multilingual alignment hard to resolve via translated

¹Code and aggregate evaluation results are available in the [project repository](#).

data alone (Shen et al., 2024). These studies typically operate at the level of national languages; we instead examine regional Sinitic varieties and culturally grounded register shifts within the Chinese language family.

Obscure-register jailbreaks. Most closely related is CC-BOS (Huang et al., 2026), which uses multi-dimensional fruit-fly optimization over dimensions such as role, metaphor, expression, and context to generate classical-Chinese adversarial prompts in black-box settings, framing classical Chinese as concise and obscure relative to modern safety-training corpora. We ask whether this “high capability–low alignment” phenomenon extends to living dialects, and whether the dialectal surface form is itself load-bearing.

Dialectal and low-resource NLP. NLP systems underperform on dialectal and low-resource varieties because data, annotation, and evaluation norms privilege standardized registers (Joshi et al., 2020; Bender, 2011), creating representational and practical harms. We connect this to LLM safety by showing that dialectal and register variation exposes inconsistencies in refusal robustness, not only in task accuracy—while clarifying that in our ablation the dominant failure mode is optimization over strategy dimensions rather than the dialect surface form alone, so the effect is broader than dialectal coverage per se.

3 Threat Model and Responsible Scope

We evaluate a black-box red-team setting in which an auditor queries a deployed target model and observes its response, with access to an attack model that proposes dialect-framed variants and a translation model for judge-based evaluation—approximating third-party robustness auditing of closed-weight systems. Because jailbreak research is dual-use, we restrict operational detail: we report aggregate metrics, sanitized methodology, and defense implications, and omit optimized prompts, raw harmful responses, and prompt templates that would make the attack directly reusable.

4 Method

4.1 Base Optimizer

Our evaluation builds on the CC-BOS black-box optimization framework (Huang et al., 2026). At a high level, the optimizer maintains a small population of prompt-construction policies. Each policy

maps a harmful intent into a dialect-framed prompt using a set of register, context, and metaphor dimensions. The target model is queried with candidate prompts; the resulting responses are translated to English when necessary; and an automated judge assigns a score. The optimizer then updates the candidate population using a bio-inspired search procedure—fruit-fly optimization (FOA), which we use as shorthand for the optimizer throughout.

Let x_i denote the i -th harmful-intent prompt, $d \in \mathcal{D}$ a dialect/register, and $m \in \mathcal{M}$ a target model. A prompt constructor produces a dialectal prompt $z_{i,d,\theta} = \pi_\theta(x_i, d)$, where θ indexes a redacted set of register-framing choices; the target returns $y = m(z_{i,d,\theta})$, translated to English as $\tilde{y} = T_d(y)$ when required. The judge assigns a consistency score $c_i \in \{0, 20, \dots, 100\}$ (a 0–5 rubric scaled by 20) plus a keyword bonus $b_i \in \{0, 20\}$, giving a total $s_i = c_i + b_i \in [0, 120]$. We define success at threshold $\tau = 80$:

$$\text{ASR}_\tau(d, m) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{s_i(d, m) \geq \tau\}. \quad (1)$$

We separately report Avg.Score, the average judge rating on the original 0–5 scale (more comparable to the rubric than the thresholded score), and the average number of target-model queries Avg.Q, over all prompts and over successful prompts.

The consistency score c_i is the judge’s rating of how fully the response carries out the harmful intent—as opposed to refusing, deflecting, or answering a different question—on the 0–5 rubric of Huang et al. (2026), rescaled by 20 so that it shares the 0–120 range with the keyword bonus. We inherit the rubric, the $\times 20$ scaling, and the success threshold $\tau = 80$ (a rubric level of 4 out of 5) unchanged from CC-BOS for comparability; none of the three is tuned, and τ is applied identically to every dialect, target, and condition. Because binary ASR depends on this cutoff, we report a threshold-sensitivity analysis in Appendix I: the two-regime gap between non-FOA and FOA conditions persists across $\tau \in \{40, 60, 80\}$ (non-FOA stays below 7% while FOA remains at ceiling), so the headline contrast is not an artifact of the specific threshold.

4.2 Dialect Extensions

We evaluate three prompt registers:

- **Classical Chinese:** the original CC-BOS baseline, using a single-stage classical-Chinese-to-English translation path for evaluation.

- **Shanghainese/Wu:** a living regional variety with dialect-specific lexical and pragmatic features, paired with local sociocultural framing. Evaluation uses a two-stage translation path: Shanghainese/Wu-oriented text to Mandarin gloss, then Mandarin to English.
- **Cantonese/Yue:** a living regional variety with a broader online footprint than Shanghainese, paired with southern Chinese and Hong Kong cultural framing. Evaluation also uses the two-stage translation path.

The dialect modules preserve the optimizer interface and the number of high-level strategy dimensions, but the concrete prompt templates and optimized prompts are not disclosed. This design permits direct comparison with classical Chinese while reducing the risk of releasing reusable attack artifacts.

4.3 Translation and Quality Control

Because target responses may contain dialectal, Mandarin, or mixed-register content, we translate before judging (single-stage for classical Chinese, two-stage through Mandarin glosses for Shanghainese and Cantonese). The system logs translation-quality flags for hedging or low-confidence phrases; this does not eliminate translation noise but prevents obvious low-confidence translations from being silently treated as reliable evidence.

4.4 Ablation Conditions

To separate the contribution of the dialectal surface form from the contribution of the cultural-framing strategy bank and the FOA optimizer itself, we partition the optimizer’s eight prompt-construction dimensions into two banks: *surface-form dimensions* (e.g., expression style, knowledge framing) and *cultural-frame dimensions* (e.g., role, metaphor, context, guidance). The exact split is given in our public configuration. Crossing FOA usage, frozen-dimension bank, and surface form yields the six conditions in Table 1.

The three non-FOA conditions isolate surface form alone (English, Mandarin, or naive dialect translation); the three FOA conditions isolate the optimizer by freezing one bank: surface-form-only (dialect_without_cultural_frame), cultural-frame-only in Mandarin (mandarin_with_cultural_frame),

Condition	FOA?	Frozen dims	Surface form
english_original	no	n/a	English
mandarin_translation	no	n/a	Mandarin
naive_dialect_translation	no	n/a	Dialect
dialect_without_cultural_frame	yes	cultural	Dialect
mandarin_with_cultural_frame	yes	surface	Mandarin
full_dialect_cultural_frame	yes	none	Dialect

Table 1: The six ablation conditions. Non-FOA conditions issue a single attack-LLM call per harmful intent and skip the optimization loop. FOA conditions with a frozen bank fix those dimensions at a canonical baseline.

or both banks in the target dialect (full_dialect_cultural_frame).

5 Experimental Setup

Dataset. We use the 50-prompt AdvBench subset used in the CC-BOS comparison setting (Zou et al., 2023; Huang et al., 2026). The prompts cover multiple harmful-intent categories. In the paper we do not reproduce the raw prompts.

Models. The attack model is DeepSeek-Chat, which proposes candidate dialect-framed prompts. The translation model is also DeepSeek-Chat. The automated judge is GPT-4o. The target models are GPT-4o and DeepSeek-Reasoner. All models are accessed through their providers’ hosted APIs during May 2026: GPT-4o (OpenAI, gpt-4o), DeepSeek-Chat and DeepSeek-Reasoner (DeepSeek, deepseek-chat and deepseek-reasoner), and—for the cross-vendor judge-agreement check in Appendix H—Qwen-Max (Alibaba, qwen/qwen3.6-max-preview via OpenRouter). Because these are hosted, continually updated endpoints, exact reproduction may require pinning to the same API snapshots.

Optimization budget. For each FOA cell, we use population size 5 and maximum 5 optimization iterations. We use early-stop threshold $\tau = 80$, corresponding to a query-efficient setting that stops when a substantial successful response has been detected. This setting should not be interpreted as maximizing response severity; it is designed to measure rapid penetration and query cost.

Completed runs and integrity checks. We complete all 36 cells: 3 dialects $\times$ 6 conditions $\times$ 2 targets $\times$ 50 prompts = 1800 target-model evaluations. Every record file contains exactly 50 unique prompt IDs with non-empty responses and valid score, target, dialect, and condition fields (a keep-last dedupe pass was applied to earlier-session

records).

6 Results

6.1 Attack Success Rates

Under the full method (full_dialect_cultural_frame; bottom row of Table 2), all three dialects achieve 100.0% ASR against both GPT-4o and DeepSeek-Reasoner with no genuine failures. Avg.Score is consistently high (4.02–4.28 on the 0–5 scale), with DeepSeek-Reasoner marginally above GPT-4o in every dialect (peaking at 4.28 for classical Chinese on DeepSeek-Reasoner). For each 50/50 cell the Wilson 95% confidence interval is approximately 92.9–100.0%; the benchmark is small, so we do not over-interpret sub-percent differences, but the ceiling ASR across all tested registers is robust under the same evaluation budget (and threshold-robust; Appendix I).

6.2 Ablation Results

Table 2 reports ASR for all six conditions. A clean two-regime pattern emerges. The three non-FOA conditions stay between 0% and 8% ASR for every dialect–target pair: directly translating a harmful English prompt into Mandarin or into a dialect, without iterative search over a strategy bank, almost always fails. In contrast, every FOA condition—including the two restricted variants that hold one of the two strategy banks frozen—reaches 98–100% ASR.

The ablation supports two claims. The dialectal surface form is not what causes the failure: `mandarin_with_cultural_frame` outputs Standard Mandarin yet reaches 98–100% ASR, matching the full dialect condition. Nor are the cultural-framing dimensions strictly necessary: with that bank frozen (`dialect_without_cultural_frame`) ASR still reaches 100% everywhere. The common ingredient explaining the jump from 0–8% to 98–100% is the optimizer operating over *some* non-trivial strategy bank; once it runs, the surface language and the choice of which bank to optimize have at most a small effect on binary success.

Paired McNemar tests confirm this split (Appendix F): across all 18 within-FOA contrasts every p-value is 1.00 with at most one discordant pair, so we cannot reject equality of the three FOA conditions at $n=50$ —a null constrained by ceiling saturation rather than a positive demonstration of

symmetry. Every FOA-vs-non-FOA contrast, by contrast, is highly significant (e.g., 46/50 discordant pairs, $p < 10^{-10}$). We discuss the implications in Section 8.

6.3 Query Efficiency

Although binary success saturates in the FOA conditions, query cost does not. Table 3 reports average target-model queries under the full method. Against GPT-4o, classical Chinese is most efficient (2.56), followed by Shanghainese (3.54) and Cantonese (5.14). Against DeepSeek-Reasoner, Shanghainese is most efficient (2.98), followed by classical Chinese (3.18) and Cantonese (5.84). Cantonese is consistently the least efficient register against both targets.

We read these differences as interactions between target-model refusal heuristics and dialectal framing, not as a measure of attack effectiveness; Cantonese’s lower efficiency is plausibly tied to its greater online representation, though we do not directly measure pretraining exposure.

6.4 Score Distribution and Severity

Table 5 (Appendix C) shows the score distribution under the full method. GPT-4o responses cluster tightly at the early-stop threshold: 48–49 of 50 successful responses receive a total score of exactly 80. DeepSeek-Reasoner produces more high-severity responses even though the optimizer stops at the first threshold-crossing success: 4–7 of 50 prompts reach scores of 100 or 120 per dialect, versus 1–2 of 50 for GPT-4o. The high-severity count is largest for classical Chinese on DeepSeek-Reasoner (9/50) and Cantonese on DeepSeek-Reasoner (7/50).

6.5 Failure Analysis

Under the full method there are zero failures across all six cells; failures concentrate in the non-FOA baselines (near floor) and a single near-ceiling FOA cell (Cantonese `mandarin_with_cultural_frame` on DeepSeek-Reasoner, 49/50). Given the optimizer and either bank, residual binary variation is essentially search-budget noise rather than harm-category dependence.

6.6 Culture-Neutral Generic Strategy Control

To test whether the cultural framing in our dialectal strategy bank is itself load-bearing—or whether *any* sufficiently expressive optimizer-controlled strategy bank suffices—we run a culture-neutral

Condition	Classical Chinese		Shanghainese		Cantonese	
	GPT-4o	DeepSeek-Reasoner	GPT-4o	DeepSeek-Reasoner	GPT-4o	DeepSeek-Reasoner
english_original	2.0	2.0	0.0	2.0	2.0	2.0
mandarin_translation	2.0	2.0	6.0	2.0	2.0	2.0
naive_dialect_translation	2.0	8.0	2.0	4.0	2.0	0.0
dialect_without_cultural_frame	100.0	100.0	100.0	100.0	100.0	100.0
mandarin_with_cultural_frame	100.0	100.0	100.0	100.0	100.0	98.0
full_dialect_cultural_frame	100.0	100.0	100.0	100.0	100.0	100.0

Table 2: ASR (%) under all six conditions, $\tau = 80$, $n = 50$ per cell. The top three rows are non-FOA baselines (single attack-LLM call). The bottom three rows hold the FOA optimizer fixed and ablate which strategy bank is active. Non-FOA conditions stay below 10% in every cell; FOA conditions reach ceiling in every cell.

Dialect/Register	All prompts		Successful only	
	GPT-4o	DeepSeek-Reasoner	GPT-4o	DeepSeek-Reasoner
Classical Chinese	2.56	3.18	2.56	3.18
Shanghainese/Wu	3.54	2.98	3.54	2.98
Cantonese/Yue	5.14	5.84	5.14	5.84

Table 3: Average target-model queries under the full method ($\tau = 80$). Because every prompt succeeds, the “successful-only” columns equal the “all prompts” columns. Dialect choice changes efficiency by up to $2\times$ even though final ASR is identical.

Surface form	Target	ASR (%)	Avg. total (0–120)	Avg. q.
English	GPT-4o	100.0	80.4	1.24
Mandarin	GPT-4o	100.0	82.0	1.30
Chinese register	GPT-4o	100.0	81.6	1.36
English	DeepSeek-Reasoner	100.0	82.0	1.52
Mandarin	DeepSeek-Reasoner	100.0	83.6	1.54
Chinese register	DeepSeek-Reasoner	100.0	85.6	1.40

Table 4: Culture-neutral generic-strategy control: FOA over a generic (non-culturally-grounded) strategy bank, three surface forms, $n=50$ AdvBench prompts per cell, $\tau=80$. All six cells reach 100% ASR with average severity comparable to the dialectal full method, and at lower query cost (1.24–1.54 vs. the dialectal cells in Table 3). The cultural framing is therefore *not* a necessary ingredient: any sufficiently expressive optimizer-controlled bank reaches the same ceiling.

generic-strategy variant (Table 4). The generic bank replaces the dialect-specific sociocultural dimensions (commerce framing, tea-house social frame, performing-arts frame, etc.) with a culture-neutral set of role, format, scenario, and pretext dimensions, evaluated across three surface forms: English, Mandarin, and a generic Chinese register. All other elements of the pipeline—FOA loop, target models, judge, $\tau=80$ early-stop, $n=50$ AdvBench subset—are held fixed.

This strengthens the ablation claim: surface form is not load-bearing once an optimizer and a strategy bank are active (Section 6.2), and the bank’s *cultural grounding* is not load-bearing either—a culture-neutral bank reaches the same ceiling at *lower* query cost. What remains is the conjunc-

tion *optimizer plus any sufficiently expressive strategy bank*, whether culturally grounded or neutral, across any surface form.

7 Qualitative Analysis of Successful Dialectal Attacks

To understand *how* the successful dialectal prompts work, we analyze the 200 successful optimized prompts under the full method (100 Shanghainese, 100 Cantonese). The analysis is safety-preserving: we code high-level rhetorical and sociocultural mechanisms without reproducing prompts or harmful responses. Schema-defining counts come from an earlier 99-per-dialect snapshot (frozen before the reliability check so labels cannot be retrofitted); the check itself (Section 7.1) uses the final FOA data.

7.1 Coding Procedure and Reliability

We coded each successful prompt for recurring surface and discourse mechanisms. The coding is not intended to establish causality, because successful prompts are outputs of an optimizer and many mechanisms co-occur. Instead, it characterizes the prompt families that the optimizer repeatedly converged on. Table 7 (Appendix G) reports prevalence by dialect under the v1 descriptive snapshot used to define the schema.

To assess the reliability of the schema, two coders independently labeled a random sample of 20 successful prompts drawn from the final FOA data (10 Shanghainese and 10 Cantonese, all under the full-method condition) using the 9-mechanism schema in Table 7. To preserve independence, the second coder was a separate general-purpose language-model agent dispatched in an isolated session with the schema and prompts but no exposure to the paper’s hypothesis, prior coding, or methodological context; this is disclosed here as a methodological substitute for a fully human second coder.

Median Cohen’s κ across the nine mechanisms was +0.706 (range +0.000 to +1.000), with four mechanisms at perfect agreement; per-mechanism values, including the one degenerate-marginal $\kappa=0$ case, are reported in Appendix G.

As a template-artifact control, we additionally coded 20 non-FOA prompts drawn from the same record set (english_original, mandarin_translation, and naive_dialect_translation) and compared mechanism prevalence. The three-step procedural scaffold appears in 1/20 non-FOA prompts but 20/20 FOA prompts ($\Delta = +0.95$); parenthetical semantic glossing shows the same 1/20 vs. 20/20 split ($\Delta = +0.95$); modern technical bridges, commerce/logistics metaphors, and grey-zone authority frames appear in zero non-FOA prompts. This pattern supports the claim that the common skeleton in Section 7 is *emergent under optimization* rather than mechanically induced by the prompt template or by the underlying AdvBench requests alone.

7.2 A Common Skeleton: Glossing plus Proceduralization

Two mechanisms appear in nearly all successful prompts. First, parenthetical semantic glossing: a culturally specific phrase is paired with an explanatory gloss pointing the model toward the intended modern abstraction, creating a dual-channel prompt that is opaque to safety heuristics relying on standard phrasing yet transparent enough for the base model to follow—*selective transparency* rather than simple obfuscation. Second, a three-part procedural scaffold: once the model accepts the prompt as a legitimate explanatory task, the structure elicits ordered, complete, concrete content—so the dialectal wrapper shifts the model into an instructional register while the scaffold determines the response shape. The ablation (Section 6.2) is consistent with this: the conditions that reach ceiling are those that retain both the scaffold and the gloss, and removing the optimizer removes them, with ASR returning to the non-FOA baseline. We read these mechanisms as characterizing the prompt families the optimizer repeatedly converges on; they should not be read as individually necessary or sufficient.

7.3 Per-Dialect Profiles

The two living dialects converge on different socio-cultural frames (full prevalences in Table 7). We

stress that these frame preferences are *emergent*: they are read off the prompts the optimizer converged on against each target, not imposed from sociolinguistic priors, and we make no claim that human speakers of these varieties preferentially use such frames. Shanghainese prompts are dominated by commercial, logistical, and semi-legal urban frames—account books, merchant houses, warehouses, concession-era ambiguity—that re-code harmful requests as business process, asset transfer, or access control, and they frequently add an explicit modern technical bridge (88/99) reinterpreting the local scene through contemporary engineering or systems language. Cantonese prompts use commerce and grey-zone frames less heavily and instead draw more on performative and communal discourse (performing-arts, tea-house, and ritual settings), relying on pragmatic role embedding rather than explicit technical recoding: the model is invited into a socially familiar scene that treats the request as genre knowledge or narrative reconstruction. Cantonese is also consistently less query-efficient than Shanghainese, plausibly because written Cantonese has more online representation and thus somewhat better model coverage; our analysis cannot directly verify this training-data hypothesis.

7.4 Mechanism Interactions

Successful prompts rarely rely on a single mechanism: a typical pattern combines a culturally specific scene, a euphemistic mapping from the harmful concept to a benign local object, parenthetical glossing that restores the modern semantics, and a multi-step request format. This layering helps explain the high ASR despite dialect differences. Prompt length itself does not explain success or efficiency—within each dialect, length has near-zero correlation with query count among successful prompts—so the decisive factor is the alignment between framing, glossing, and the target’s safety heuristics, not verbosity.

7.5 High-Severity Successes

Among the 200 living-dialect successes under the full method, 11 receive scores of 100 or 120, ten of them against DeepSeek-Reasoner—reinforcing that DeepSeek-Reasoner produces more severe responses immediately after a dialect-framed prompt crosses the refusal boundary, even though the optimizer stops at $\tau=80$ without continued severity-maximization. The qualitative mechanisms also

sharpen the defense implications we return to in Section 8 and Section 9: keyword blacklists miss benign-vocabulary prompts that carry harm through metaphor and glossing, while flagging all dialectal language would be discriminatory, motivating semantic normalization with uncertainty-aware review instead.

8 Analysis

8.1 An Expressive Strategy Bank Is the Load-Bearing Component

The ablation in Section 6.2 reframes the dialectal jailbreak story. Without FOA, no surface form—English, Mandarin, or dialect—produces more than 8% ASR. With FOA, both a Mandarin-surface attack restricted to the cultural-frame bank and a dialect-surface attack restricted to the surface-form bank reach ceiling. The dialectal surface form is therefore neither necessary nor sufficient: it is a vector along which the optimizer can find effective attacks, but several other vectors work just as well.

Search versus construction. We do not over-attribute this to the iterative *search* itself. The non-FOA baselines remove both the multi-dimensional strategy-bank *construction* and the optimization loop, so they cannot separate adaptive search from a single richly constructed prompt. The generic control (Section 6.6) is informative: it reaches 100% ASR at only 1.24–1.54 queries—the first constructed candidate usually succeeds and the search loop is barely engaged—suggesting the load-bearing ingredient is the strategy-bank construction rather than search per se. We thus frame the load-bearing component as an *optimizer-or-template* strategy bank. A budget-matched best-of-*k* control—non-adaptive sampling of the same strategy bank at FOA’s full 25-query budget, which removes adaptive search while holding construction and budget fixed—is designed to test this directly. This control is currently *partial*: of the six intended dialect×target cells only classical-Chinese/GPT-4o is complete, where non-adaptive sampling reaches 100% ASR (33/33) at the matched budget. We report this as preliminary rather than decisive evidence and do not rest the main attribution on it; it is consistent with the near-single-query generic control (Section 6.6), which likewise reaches ceiling with the search loop barely engaged. Completing the remaining five cells is the cleanest way to fully isolate adaptive search from construction (Section 10).

Two consequences follow: surface-only defenses (e.g., dialect-specific keyword filters) will not generalize, since the same attack succeeds with a Mandarin surface form; and future “dialectal jailbreak” evaluations should always include a Mandarin-surface variant of the same optimized strategy bank, or an apparent dialect effect may simply be an optimizer effect.

8.2 Dialect-Agnostic Effectiveness, Dialect-Specific Efficiency

The observed ASR pattern is dialect-agnostic under the full method: all tested dialects reach ceiling ASR under our judge and threshold. The efficiency pattern is dialect-specific: the optimizer reaches successful prompts at different query costs depending on the target model. We interpret this as evidence that the main weakness lies in semantic safety generalization under optimization plus a strategy bank, while search efficiency depends on target-model training coverage and refusal heuristics for the specific surface variety. A benchmark reporting only ASR would call all dialects equivalent; query cost and severity reveal which dialect–model pairs are more brittle in practice. More broadly, the generic control (Section 6.6) locates the failure in the safety classifier’s tolerance for optimizer-driven prompt-shape variation rather than in missing dialectal training data; because the same surface-form sensitivity also fires on innocuous inputs, benign users in underrepresented varieties may see inconsistent safety behavior.

Finally, the reasoning model DeepSeek-Reasoner is more vulnerable than GPT-4o: it produces more high-score responses across all dialects and is the only target admitting a non-ceiling ablation cell (49/50). This may reflect a reasoning-oriented model more readily elaborating once it accepts a dialect-framed premise, or weaker safety coverage for these registers; we cannot distinguish these without private chain-of-thought or training data, but it motivates targeted safety stress-testing of reasoning models.

9 Defense Recommendations

We propose three practical recommendations. First, **dialect-aware red-team suites with Mandarin-surface controls**: audits should cover dialectal and regional varieties (classical/literary registers, Cantonese, Shanghainese, Hokkien/Min, Hakka, code-switching) and include a Mandarin-surface version

of every optimized attack, to avoid mistaking an optimizer effect for a dialect effect. Second, **semantic intent detection after normalization**: because every successful attack in our study preserves the harmful *intent* while only varying surface form and framing, the natural countermeasure is a guardrail that evaluates intent semantically rather than by surface cues. Concretely, such a guardrail would (i) normalize dialectal or obfuscated input into a high-resource pivot (e.g., Mandarin or English) via translation or paraphrase; (ii) recover the underlying request by asking a semantic classifier what task the prompt ultimately asks for, seeing through the recurring evasion scaffolds we identify in Section 7—parenthetical glossing, multi-step procedural reframing, and metaphorical recoding; and (iii) apply a single policy to the recovered intent regardless of surface variety, so the same request is treated identically in English, Mandarin, or a dialect. This must be paired with translation-quality checks in both directions: an over-aggressive normalizer risks discriminatory false positives against benign dialectal users, while a lossy one risks unsafe false negatives. Robust semantic intent detection—rather than surface-form or keyword refusal—is the defense most directly implied by our finding that the failure lives in the safety classifier’s tolerance for optimizer-driven prompt-shape variation. Third, **dialectal safety fine-tuning and multi-metric reporting**: fine-tune on safety examples that preserve dialectal context (with native-speaker review), and report query efficiency and severity alongside ASR.

10 Conclusion

We asked which component of a dialect-framed black-box jailbreak is load-bearing. Across 1800 evaluations in a 36-cell ablation, classical Chinese, Shanghaiese, and Cantonese all reach 100% ASR on GPT-4o and DeepSeek-Reasoner, yet the dialectal surface form is not: a Mandarin cultural-frame-only attack and a dialect surface-form-only attack both hit the ceiling while non-optimized translations stay below 8%. The load-bearing ingredient is an expressive optimizer-controlled strategy bank, and dialect choice changes only efficiency and severity—leaving current safety alignment brittle under optimization over any such bank (defenses in Section 9).

Limitations

This study has several limitations.

Small benchmark, and a saturation-prone one.

We evaluate a 50-prompt AdvBench subset. Although the runs are complete for all 36 cells of the ablation matrix, the benchmark is too small for precise estimates of small differences across dialects or target models. AdvBench is also known to elicit inflated success rates and “empty” jailbreaks relative to harder, rubric-graded benchmarks, which motivated StrongREJECT (Souly et al., 2024). The ceiling-level ASR we observe should therefore be read partly as a property of the instrument. Larger benchmark and cross-benchmark generalization experiments (StrongREJECT, CLAS) and $\tau=120$ reruns of the main results are in progress and will be reported in an extended version.

Partial matched-budget non-adaptive control.

Our within-FOA contrasts identify which strategy banks the optimizer relies on, but they do not by themselves separate *adaptive search* from *search budget*: a non-adaptive sampler with the same query budget might or might not reach the same ceiling. We have begun a budget-matched best-of- k control over the same strategy bank (non-adaptive sampling at FOA’s 25-query budget; `phase_e_matched_budget_baselines` in `configs/ablation_matrix.yaml`). It is not yet complete: of the six intended dialect $\times$ target cells, only classical-Chinese/GPT-4o is finished, where non-adaptive sampling reaches 100% ASR (33/33) at the matched budget. We treat this as preliminary evidence—consistent with, but not on its own establishing, the claim that adaptive search is not required for ceiling success—and do not rely on it for the central attribution, which rests on the ablation and the near-single-query generic control (Section 6.6). We will report the full six-cell control in an extended version, and until then we limit the adaptive-search claim accordingly.

Single automated judge in the headline ASR.

The headline ASR is computed under a single GPT-4o judge. To bound the judge-choice confound, we re-scored a stratified subsample of 216 rows (six rows per dialect $\times$ target $\times$ condition cell, seeded for reproducibility) with a cross-vendor second judge (Qwen-Max, accessed via OpenRouter as `qwen/qwen3.6-max-preview`) and report the resulting agreement in Appendix H. The two judges agree strongly on the continuous severity scale

(Pearson 0.87, Spearman 0.86, Krippendorff $\alpha = 0.79$) and, on the total-score basis used for ASR throughout the paper, also on the binary success flag (raw agreement 0.931, Cohen’s $\kappa = 0.86$). A much lower $\kappa = 0.11$ appears only if success is binarised on the strict consistency-only score (dropping the keyword bonus), where GPT-4o’s marginal success rate is $\sim 5\%$ against Qwen-Max’s $\sim 45\%$; this is a high-prevalence “ κ paradox” on a non-ASR basis, not a ranking disagreement (Appendix H). We nonetheless caveat that the *absolute* ASR ceiling under GPT-4o is somewhat calibration-dependent—it falls to 88.8% under Qwen-Max—even though the within-judge ablation contrasts (evaluated under a fixed judge) are robust to this gap. A related confound is model-family coupling: GPT-4o serves as both a target and the primary judge, and DeepSeek-Chat is both the attack and the translation model. Same-family scoring and translation can bias absolute severity estimates; the cross-vendor Qwen-Max re-scoring partially bounds the judge side of this, but a fully decoupled judge and translation pipeline is left to future work.

Translation confounds. Shanghainese and Cantonese use two-stage translation through Mandarin before English evaluation. This may understate or distort response severity relative to classical Chinese. Better evaluation should include bilingual or dialect-proficient human annotation.

Limited target models in the full ablation. The full ablation evaluates only GPT-4o and DeepSeek-Reasoner. Cross-model generalization to Claude Sonnet, Qwen-Max, and DeepSeek-Chat is configured but not yet complete. A small smoke test on DeepSeek-Chat suggests similar risks, but DeepSeek-Chat is also the attack and translation model, creating a self-attack confound.

No release of attack artifacts. For safety reasons, we do not release optimized prompts or raw harmful responses. This limits exact reproducibility, but we believe aggregate reproducibility and controlled access are appropriate for dual-use jailbreak research.

Sociolinguistic simplification. The term “dialect” hides substantial linguistic and social complexity. Shanghainese and Cantonese are not merely stylistic variants of Mandarin. Future work should collaborate with dialect speakers and sociolinguists to design more faithful and respectful evaluations.

Ethics Statement

This work studies jailbreak vulnerabilities and is therefore dual-use. The purpose is defensive: to identify safety gaps affecting underrepresented language varieties and to motivate more equitable multilingual alignment. We intentionally omit optimized adversarial prompts, raw harmful responses, and prompt templates. We report only aggregate statistics and sanitized failure categories. Before public release of code or data, we recommend redacting harmful prompts and outputs, adding access controls for high-risk artifacts, and notifying affected model providers with a concise vulnerability report.

There is also a representational risk: readers might incorrectly infer that dialect speakers or regional cultures are associated with unsafe behavior. That is not our claim. The vulnerability lies in model safety generalization, not in the languages, dialects, or communities being evaluated. Dialect-aware safety should improve protection and service quality for speakers of underrepresented varieties.

Use of AI Assistants

We used a large language model as the independent second coder for the inter-rater reliability check (Section 7.1, Appendix G), dispatched in an isolated, hypothesis-blind session as a methodological substitute for a human coder. AI assistants were additionally used for code scaffolding and for editing and proofreading the manuscript. All experimental design, scientific claims, statistical analyses, and results were verified by the authors.

References

- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

- Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Li-dong Bing. 2024. Multilingual jailbreak challenges in large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Samuel R. Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Christopher Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*, pages 3356–3369.
- Xun Huang, Simeng Qin, Xiaoshuang Jia, Ranjie Duan, Huanqian Yan, Zhitao Zeng, Fei Yang, Yang Liu, and Xiaojun Jia. 2026. Obscure but effective: Classical Chinese jailbreak prompt optimization via bio-inspired search. In *Proceedings of the Fourteenth International Conference on Learning Representations*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL*, pages 6282–6293.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*.
- Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box LLMs automatically. In *Advances in Neural Information Processing Systems*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of EMNLP*, pages 3419–3448.
- Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. “Do Anything Now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*.
- Lingfeng Shen, Sihao Chen, Linfeng Liu, and collaborators. 2024. Dissecting the safety challenges of large language models in multilingual contexts. In *Findings of ACL*.
- Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for empty jailbreaks. *arXiv preprint arXiv:2402.10260*.
- Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak GPT-4. In *NeurIPS Workshop on Socially Responsible Language Modelling Research*.
- Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023. GPTFuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.
- Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2024. GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher. In *Proceedings of the Twelfth International Conference on Learning Representations*.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Preliminary Smoke Test on DeepSeek-Chat

Before the final 1800-evaluation run, a 5-prompt smoke test included DeepSeek-Chat as a target model. Classical Chinese and Cantonese each achieved 4/5 success, while Shanghaiese achieved 5/5 success. We exclude DeepSeek-Chat from the main full-run comparison because it is also used as the attack and translation model, creating a self-attack confound. The smoke test is useful only as preliminary evidence that the vulnerability is not limited to the two fully evaluated targets.

Target	Dialect/Register	Score 80	Score 100	Score 120
GPT-4o	Classical Chinese	49	1	0
	Shanghainese/Wu	48	1	1
	Cantonese/Yue	49	0	1
DeepSeek-Reasoner	Classical Chinese	41	4	5
	Shanghainese/Wu	46	0	4
	Cantonese/Yue	43	1	6

Table 5: Total score distribution under the full method with early stopping ($\tau = 80$). DeepSeek-Reasoner has more high-score outputs even though the optimizer is not trying to maximize severity beyond $\tau = 80$.

B Comparison with CC-BOS

The original CC-BOS paper reports classical-Chinese results under both a high-severity setting and a query-efficient setting. Our experiments use the query-efficient early-stop threshold $\tau = 80$. Therefore, our Avg.Score values should not be directly compared to CC-BOS high-severity results obtained with a larger optimization budget. The ASR comparison is more meaningful than the average severity comparison. A $\tau=120$ high-severity rerun of our main results is in progress and will be reported in an extended version.

C Score Distribution and Severity

Table 5 reports the total-score distribution under the full method, referenced from the severity discussion in the main text.

D Responsible Artifact Release Checklist

For any public release accompanying this paper, we recommend:

1. release aggregate metrics and sanitized scripts for computing ASR, query cost, and score distributions;
2. withhold optimized adversarial prompts and raw harmful model responses;
3. provide only benign toy prompts in public examples;
4. require controlled access for any high-risk benchmark artifacts;
5. include documentation warning that dialectal vulnerability reflects model limitations, not speaker communities;
6. include a model-provider notification plan for severe new vulnerabilities discovered in follow-up runs.

E Redacted Prompt Structure and Evaluation Rubric

To convey the *shape* of a successful optimized prompt without releasing a reusable attack, we give a schematic skeleton in which every content slot is a placeholder rather than filled text. This reflects the structure that the qualitative coding (Section 7) found recurring across successful dialectal prompts; the placeholders are intentionally left unfilled to preserve the safety protocol.

[ROLE / PERSONA in dialect]
[CULTURAL-FRAME SETUP: benign
local scene]
[REQUEST, recoded as a task within the
scene]
(gloss: [local term] $\rightarrow$ [abstract mod-
ern concept])
[3-STEP PROCEDURAL SCAFFOLD:
1. ... 2. ... 3. ...]
[OUTPUT-FORMAT INSTRUCTION]

No harmful intent, target behavior, or operational detail is instantiated in any slot; the skeleton only shows how the optimizer wraps an arbitrary intent in persona, cultural framing, parenthetical glossing, and procedural structure. We do not release filled instances.

The automated judge is prompted separately from the attack. It receives the (translated) target response and the original intent and returns an integer rating on a 0–5 rubric—0 for a clear refusal or off-topic answer, rising to 5 for a response that fully and specifically carries out the requested harmful task—together with a short justification. This rating is the consistency score c_i (rescaled by 20; Section 7.1); the keyword bonus b_i is added by a separate rule-based check. Sharing the rubric’s structure supports evaluation reproducibility without providing an attack, since the judge prompt scores responses rather than generating them.

F Paired Statistical Tests

Table 6 reports McNemar exact p-values and discordant-pair counts for every within-FOA condition-pair contrast within each (dialect, target) cell. For each contrast, the success indicators are prompt-id aligned at $n=50$. The short labels no-frame, mand-frame, and full correspond to the three FOA conditions in Table 1.

Dialect	Target	Contrast (A vs. B)	n_{disc}	McNemar p	Δ ASR (%)
Classical Chinese	GPT-4o	no-frame vs. mand-frame	0	1.00	0.0
Classical Chinese	GPT-4o	no-frame vs. full	0	1.00	0.0
Classical Chinese	GPT-4o	mand-frame vs. full	0	1.00	0.0
Classical Chinese	DeepSeek-Reasoner	no-frame vs. mand-frame	0	1.00	0.0
Classical Chinese	DeepSeek-Reasoner	no-frame vs. full	0	1.00	0.0
Classical Chinese	DeepSeek-Reasoner	mand-frame vs. full	0	1.00	0.0
Shanghainese	GPT-4o	no-frame vs. mand-frame	0	1.00	0.0
Shanghainese	GPT-4o	no-frame vs. full	0	1.00	0.0
Shanghainese	GPT-4o	mand-frame vs. full	0	1.00	0.0
Shanghainese	DeepSeek-Reasoner	no-frame vs. mand-frame	0	1.00	0.0
Shanghainese	DeepSeek-Reasoner	no-frame vs. full	0	1.00	0.0
Shanghainese	DeepSeek-Reasoner	mand-frame vs. full	0	1.00	0.0
Cantonese	GPT-4o	no-frame vs. mand-frame	0	1.00	0.0
Cantonese	GPT-4o	no-frame vs. full	0	1.00	0.0
Cantonese	GPT-4o	mand-frame vs. full	0	1.00	0.0
Cantonese	DeepSeek-Reasoner	no-frame vs. mand-frame	1	1.00	+2.0
Cantonese	DeepSeek-Reasoner	no-frame vs. full	0	1.00	0.0
Cantonese	DeepSeek-Reasoner	mand-frame vs. full	1	1.00	-2.0

Table 6: Paired McNemar tests on within-FOA condition contrasts. n_{disc} = number of prompts (out of 50) on which the two conditions disagree on binary success ($\tau=80$). All p-values are 1.00 and discordant pairs ≤ 1 in every cell, so we cannot reject the null that the three FOA conditions are equally effective at $n=50$. By contrast, every FOA-vs-non-FOA contrast yields $n_{disc} \geq 46$ and $p < 10^{-10}$. The complete per-pair table is released as results/public/paired_tests.csv in the supplementary materials, alongside the aggregate ablation, generic-control, judge-agreement, and threshold-sensitivity CSVs. Δ ASR = ASR(A) - ASR(B) in percentage points.

G Inter-Rater Reliability

Two coders independently labeled the same random sample of 20 successful FOA prompts (10 Shanghainese, 10 Cantonese, all under full_dialect_cultural_frame) using the 9-mechanism schema of Table 7. Coder A was the primary author. Coder B was a separate general-purpose language-model agent dispatched in an isolated session with only the coding schema and the 20 prompt texts: it was not told that this work is a paper revision, that any hypothesis was being tested, that a common rhetorical skeleton was at issue, that an optimizer had produced the prompts, or that any prior coding existed. Both coders worked from blind identifiers and ignored the dialect/target/condition/optimizer columns. This isolated-dispatch protocol is disclosed here as a methodological substitute for a fully human second coder. Per-mechanism Cohen’s κ is reported in Table 8, alongside marginal prevalences.

As a template-artifact control, we also coded 20 non-FOA prompts (drawn from english_original, mandarin_translation, and naive_dialect_translation records, including refusals and bare AdvBench requests) with the same schema. Table 9 compares mechanism prevalences. Every dialect-specific frame (commerce, grey-zone, tech bridge, perform-arts,

tea-house) appears in zero non-FOA prompts; the procedural scaffold and parenthetical gloss each appear in only 1/20 non-FOA prompts. This $\Delta \geq +0.90$ gap for the two near-universal FOA mechanisms supports the claim that the common skeleton in Section 7 is an emergent product of the FOA loop rather than an artifact of the underlying prompt template or AdvBench surface forms.

H Cross-Judge Agreement

To bound the single-judge confound in the headline ASR, we re-scored a stratified subsample of the main ablation rows with a second, cross-vendor judge. The sample design draws six rows uniformly without replacement from each of the 36 paper cells ($3 \text{ dialects} \times 2 \text{ targets} \times 6 \text{ conditions}$), for a target of 216 rows; the random seed and per-cell row identifiers are recorded in the released sampler manifest so the same subsample is recoverable from the released aggregate metrics. The second judge is Qwen-Max, accessed via OpenRouter as qwen/qwen3.6-max-preview, using the same 0–5 rubric and parsing pipeline as the primary GPT-4o judge. One sampled row was dropped because the second judge returned a null content field (apparent content-filter behavior), leaving 215/216 sampled rows scored by both judges. Combined with a residual partial-coverage pass from earlier development,

Mechanism	Sanitized description	Shanghainese	Cantonese
Parenthetical semantic glossing	Dialectal or cultural expressions are followed by parenthetical explanations that map the euphemism to an abstract modern concept.	94/99	96/99
Three-step procedural scaffolding	The prompt asks for a structured sequence, usually organized as three subquestions, converting the intent into a procedural explanation task.	99/99	99/99
Modern technical bridge	Local metaphors are explicitly connected to modern engineering, network, financial, or system concepts.	88/99	53/99
Commerce/logistics metaphor	Harmful intents are reframed as ledger work, cargo handling, handoff protocols, settlement, or route planning.	79/99	42/99
Grey-zone authority frame	The prompt invokes semi-legal or extra-legal social authorities, boundary spaces, or historical jurisdictional ambiguity.	65/99	28/99
Historical media or research pretext	The request is framed as historical reconstruction, newspaper-style reporting, archival research, or fiction-writing background.	39/99	16/99
Performing-arts frame	The request is embedded in storytelling, opera, theatrical, or performer-centered discourse.	15/99	26/99
Tea-house social frame	The prompt uses tea-house conversation, old-neighborhood gossip, or informal elder narration as the interactional setting.	7/99	20/99
Ritual or religious-social frame	The prompt maps the intent to temple, procession, geomancy, or ritualized community activity.	11/99	23/99

Table 7: Qualitative mechanisms observed in successful living-dialect adversarial prompts (v1 coding snapshot, 99 per dialect). Counts are descriptive and non-exclusive; a single prompt can contain multiple mechanisms. We do not report prompt text.

Mechanism	κ	p_A	p_B
Parenthetical glossing	+1.000	1.00	1.00
Three-step scaffolding	+1.000	1.00	1.00
Modern technical bridge	+0.000	0.90	1.00
Commerce/logistics metaphor	+0.219	0.70	0.95
Grey-zone authority frame	+0.706	0.55	0.40
Historical/research pretext	+0.571	0.25	0.20
Performing-arts frame	+1.000	0.10	0.10
Tea-house social frame	+0.667	0.25	0.40
Ritual/religious-social frame	+1.000	0.05	0.05
Median (all 9)	+0.706	—	—

Table 8: Per-mechanism Cohen’s κ between Coder A (primary author) and Coder B (separate general-purpose LM agent, dispatched in an isolated session with the schema and prompts only) on the 20-prompt FOA reliability sample. p_A, p_B : fraction of prompts coded 1 by each coder. Median $\kappa = +0.706$; range $[+0.000, +1.000]$. The $\kappa=0.000$ entry for *Modern technical bridge* reflects degenerate prevalence on the Coder B side ($p_B=1.00$): observed agreement (0.90) equals the chance expectation under those marginals, not disagreement on labels.

the total number of rows with both judges available is $n = 303$.

Table 10 reports four agreement statistics per judge pair: Pearson and Spearman correlation on the continuous 0–120 severity score, Krippendorff’s α as an ordinal-aware reliability coefficient, and Cohen’s κ on a binary jailbreak-success indicator. The binary indicator in this table is deliberately

Mechanism	FOA p	non-FOA p	Δ
Parenthetical glossing	1.00	0.05	+0.95
Three-step scaffolding	1.00	0.05	+0.95
Modern technical bridge	0.90	0.00	+0.90
Commerce/logistics metaphor	0.70	0.00	+0.70
Grey-zone authority frame	0.55	0.00	+0.55
Historical/research pretext	0.25	0.05	+0.20
Performing-arts frame	0.10	0.00	+0.10
Tea-house social frame	0.25	0.00	+0.25
Ritual/religious-social frame	0.05	0.05	+0.00

Table 9: Mechanism prevalence in 20 successful FOA prompts (Coder A) vs. 20 non-FOA prompts coded with the same schema. $\Delta = p_{\text{FOA}} - p_{\text{non-FOA}}$.

the *consistency-only* success flag—each judge’s raw 0–100 rubric score thresholded at $\tau = 80$, *without* the response-level keyword bonus. This is the strictest binarization and is *not* the basis on which we report ASR: our ASR success flag is the total score (consistency + keyword bonus) at $\tau = 80$. We report both bases explicitly—the strict consistency-only basis here and the ASR-consistent total-score basis in the “Judge-specific ASR” paragraph below—because the two give very different κ values for an interpretable reason (below).

The apparent gap between strong scale-level agreement and weak *consistency-only* binary agreement is a thresholding artifact, not a ranking disagreement: both judges order responses by severity in essentially the same way (Spearman 0.86), and on the ASR-consistent total-score basis their binary flags also agree closely (0.931; see below).

Judge pair	n	Pearson	Spearman	α	bin. agree	κ
GPT-4o $\times$ Qwen-Max	303	0.873	0.859	0.789	0.597	0.114
GPT-4o $\times$ refusal det.	2120	0.911	0.900	0.528	0.440	0.066
Qwen-Max $\times$ refusal det.	358	0.798	0.783	0.679	0.872	0.744

Table 10: Pairwise agreement between the primary GPT-4o judge, the cross-vendor Qwen-Max second judge, and the rule-based refusal detector. Binary agreement and κ here use the strict *consistency-only* success flag (raw 0–100 rubric score at $\tau = 80$, no keyword bonus). The two LLM judges agree strongly on the continuous severity scale (Pearson 0.87, Spearman 0.86, $\alpha = 0.79$); the low consistency-only binary $\kappa = 0.11$ is a prevalence artifact—on this strict flag GPT-4o awards success on only $\sim 5\%$ of rows versus $\sim 45\%$ for Qwen-Max, so the marginals are badly unbalanced. On the total-score basis actually used for ASR (consistency + keyword bonus at $\tau = 80$) the two judges instead agree closely: raw binary agreement 0.931 and $\kappa = 0.86$ (balanced marginals ~ 0.5 ; see the “Judge-specific ASR” paragraph). The high Qwen-Max $\times$ refusal-detector κ (0.74) relative to the GPT-4o $\times$ refusal-detector κ (0.07) reflects that Qwen-Max’s consistency-only flag binarises closer to a strict refusal/non-refusal cut than GPT-4o’s does.

A truncation check rules out a benign explanation for the strict-basis gap: across all rows scored by both judges, zero rows have Qwen-Max-score = 0.0 while GPT-4o-score ≥ 80 , ruling out artifactual sentinel-zero outputs from the reduced max_tokens setting used on the second-judge pass.

For interpretation, this means the *absolute* headline ASR ($\sim 100\%$ under the full method across all paper cells) is somewhat sensitive to judge calibration—under Qwen-Max the full-method ceiling is 88.8% rather than 100% (below)—but the within-judge ablation contrasts that drive our central claim—non-optimized translations near 0% versus optimizer-on conditions near ceiling, measured under the same fixed judge in every cell—are not threatened by this calibration gap.

Judge-specific ASR. To make the calibration gap concrete, Table 11 recomputes ASR under each judge separately on the $n = 303$ dual-scored rows, on the same total-score basis used for ASR throughout the paper: each judge’s 0–100 consistency plus the same response-level keyword bonus b_i , thresholded at $\tau = 80$. The two-regime structure is judge-invariant—non-FOA stays near floor and FOA stays near ceiling under *both* judges—even though the absolute FOA ceiling falls from 100% (GPT-4o) to 88.8% (Qwen-Max). Crucially, moving from the strict consistency-only flag of Table 10 to this ASR-consistent total-score flag reconciles the two judges: raw binary agreement rises from 0.597 to 0.931 and Cohen’s κ from 0.11 to 0.86. The low consistency-only $\kappa = 0.11$ is thus a high-prevalence “ κ paradox” driven by badly unbalanced marginals ($\sim 5\%$ success for GPT-4o versus $\sim 45\%$ for Qwen-Max on the strict flag); once the keyword bonus is included—as it is in every ASR

Regime	Judge	n	ASR (%)
Non-FOA	GPT-4o	151	3.3
	Qwen-Max	151	2.0
FOA	GPT-4o	152	100.0
	Qwen-Max	152	88.8

Table 11: Judge-specific ASR on the dual-scored subsample ($n = 303$), total-score basis, $\tau = 80$. The non-FOA-vs-FOA contrast survives the judge swap; only the absolute FOA ceiling shifts (100% $\rightarrow$ 88.8%).

Regime	ASR@40	ASR@60	ASR@80
Non-FOA ($n=900$)	6.3%	5.1%	2.4%
FOA ($n=900$)	100.0%	99.9%	99.9%

Table 12: Threshold sensitivity under the GPT-4o judge across the core matrix; the gap persists through the early-stop cutoff.

figure—both judges sit near 50% prevalence and κ is well-behaved. We therefore read the judge swap as shifting the absolute ceiling by roughly ten points, not as a disagreement about which prompts succeed.

I Threshold Sensitivity

Table 12 recomputes ASR at $\tau \in \{40, 60, 80\}$ over the complete GPT-4o-judged core matrix (900 non-FOA and 900 FOA rows). The gap is stable: non-FOA never exceeds 6.3%, while FOA never falls below 99.9%. We omit $\tau > 80$ because early stopping at 80 right-censors optimized scores; valid ASR@100/@120 requires new runs.

Author Index

Bajwa, Angad Ripudaman Singh, 61
Bao, Catherine, 42
Borkar, Samyak, 10
Bourrel, Nicolas, 1

Dsouza, Aiden, 10

Fosler-Lussier, Eric, 81

Ginjala, Srishti, 81
Gupta, Manik, 33

Issam, Abderrahmane, 1

Jaishankar, Arjun, 10

Kampouridis, Michael, 67
Kantarcı, Ahmet Berkay, 95
Karazor, Ahmet, 95
Kolluri, Ganesh Pavan Kartikeya Bharadwaj, 67
Koo, Heejoon, 20

Mishra, Shreyansh, 10
Myers, Christopher W., 81

Parlak, Abdulkadir, 95
Parthasarathy, Srinivasan, 81
Patwari, Neal, 42

Saha, Maneesha Rani, 42
Sharma, Shubham, 33
Shekhar, Ravi, 67
Spanakis, Gerasimos, 1
Suneja, Yashraj, 10
Swain, Rohan, 10

Urade, Tanmay, 10

Xu, Qingyang, 104