TrustAudio 2026

TrustAudio: Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI

Proceedings of the Workshop

November 10, 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 979-8-89176-452-1

Preface

Welcome to TrustAudio, the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI, co-located with AACL-IJCNLP 2026 in Zhuhai, China, on November 10, 2026. The workshop brings together researchers in speech processing, natural language processing, machine learning, privacy, and security to study the reliability and trustworthiness of speech and audio systems.

Speech and audio foundation models support a growing range of applications, including voice assistants, transcription, accessibility tools, and multimodal agents. Their deployment raises questions about adversarial robustness, safety, deepfake misuse, and privacy. TrustAudio provides a forum for understanding these risks, developing realistic evaluations, and investigating safeguards across the model lifecycle.

This volume contains six contributed papers. They examine the geometry of speech representations for audio deepfake detection, robust Amharic speech recognition under telecommunication degradation, human perception of Arabic synthetic speech, transfer of text jailbreaks to audio agents, activation-level analysis of adversarial jailbreak transfer across speech language models, and transferable adversarial attacks in deep reinforcement learning. Together, these contributions offer complementary perspectives on the behavior and limitations of learned systems under challenging inputs.

We thank the authors for their contributions, the reviewers for their careful assessments, and the AACL-IJCNLP 2026 organizers and publication chairs for their support. We hope the workshop encourages useful exchanges across research communities and helps advance secure, safe, and privacy-preserving speech and audio AI.

Nancy F. Chen, Yiming Chen, Xiaoxue Gao, Haizhou Li,
Björn Schuller, Berrak Sisman, and Xianghu Yue
TrustAudio 2026 Workshop Organizers

Organizing Committee

Workshop Organizers

Nancy F. Chen, Agency for Science, Technology and Research, Singapore
Yiming Chen, ASUS, Singapore
Xiaoxue Gao, The Chinese University of Hong Kong, Shenzhen, China
Haizhou Li, The Chinese University of Hong Kong, Shenzhen, China
Björn Schuller, Technische Universität München (TUM), Germany
Berrak Sisman, Johns Hopkins University, USA
Xianghu Yue, Tianjin University, China

Program Committee

Program Chairs

Nancy F. Chen, Agency for Science, Technology and Research, Singapore
Yiming Chen, ASUS, Singapore
Xiaoxue Gao, The Chinese University of Hong Kong, Shenzhen, China
Haizhou Li, The Chinese University of Hong Kong, Shenzhen, China
Björn Schuller, Technische Universität München (TUM), Germany
Berrak Sisman, Johns Hopkins University, USA
Xianghu Yue, Tianjin University, China

Reviewers

Junyi Ao, The Chinese University of Hong Kong, Shenzhen
Yushi Cao, AICS
Ho-Lam Chung, National Taiwan University
Hongchao Jiang, Nanyang Technological University
Zexin Li, Nanyang Technological University
Zi Haur Pang, Kyoto University
Xinyuan Qian, University of Science and Technology Beijing
Yilin Ren, University of Science and Technology Beijing
Jiawen Wang, Technical University of Munich
Xueyi Zhang, National University of Singapore
Yuwen Wang, University of Science and Technology Beijing
Daipeng Zhang, Tianjin University
Jinhui Zhang, Tianjin University
Sen Zhang, Tianjin University
Feiyu Zhao, Tianjin University
Weihua Zheng, Singapore University of Technology and Design and A*STAR

Invited Talk

Multi-Layer Defenses Against Speech Deepfakes: Combining Passive Detection and Proactive Protection

Junichi Yamagishi
National Institute of Informatics, Tokyo, Japan
Tuesday, November 10, 2026 –

Abstract: Speech generative AI can now produce voices that are difficult for listeners to distinguish from real ones. These advances enable valuable applications, but also lower the barrier to impersonation fraud and disinformation. This talk examines how passive detection and proactive protection can work together to address these risks.

Passive defense uses deepfake detectors to identify traces of synthesis in speech signals. It requires no cooperation from the generator, but must continually adapt to new and unseen generation methods. The talk reviews recent progress in detection and its remaining limitations.

Proactive defense intervenes when speech is generated or published. Alongside C2PA provenance metadata, the talk introduces self-embedding watermarking, which conceals a compact representation of the original speech within the signal so that manipulated segments can be localized and partially restored. Such marks can provide strong evidence, but are available only for content marked in advance and may be stripped, forged, or degraded.

The talk concludes by exploring how the two approaches can cover each other's blind spots: detection can serve as a fallback when marks are removed or absent, while watermarks can ease the burden on detectors. It also considers open questions in designing and evaluating multi-layer defenses for real-world use.

Bio: Junichi Yamagishi received a Ph.D. from the Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis. He is a professor at the National Institute of Informatics, Tokyo, Japan. Previously, he held an EPSRC Career Acceleration Fellowship in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, U.K., from 2011 to 2016. He has authored and co-authored more than 400 refereed papers in international journals and conferences.

He served as an elected member of the IEEE Speech and Language Technical Committee from 2013 to 2019, as an Associate Editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing from 2014 to 2017, as a chairperson of ISCA SynSIG from 2017 to 2021, as a Senior Associate Editor for the IEEE/ACM Transaction on Audio, Speech and Language Processing from 2019 to 2023, and as a technical committee of the Asia-Pacific Signal and Information Processing Association (APSIPA) Multimedia Security and Forensics from 2018 to 2023. He has been a member of the IEEE Signal Processing Society Education Board since 2021.

He received the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, the Young Scientists' Prize from the Minister of Education, Science and Technology, the JSPS prize from the Japan Society for the Promotion of Science, the best paper award at the IEEE International Workshop on Information Forensics and Security (WIFS), the Docmo mobile science award, and IEEE Biometrics Council's BTAS/IJCB 5-Year Highest Impact Award in 2010, 2013, 2014, 2016, 2017, 2018, and 2023, respectively.

Invited Talk

Groundwork for Open Speech Foundation Models: Data, Recipes, and Task-Specific Models

Shinji Watanabe
Carnegie Mellon University
Tuesday, November 10, 2026 –

Abstract: Building speech foundation models is no longer the exclusive domain of large companies, but reproducing one still takes more than a published architecture. The hard part is the groundwork: collecting, cleaning, and automatically annotating data before training begins. Large-scale audio collections are abundant, yet raw audio and transcripts are often misaligned, mislabeled, or otherwise unusable without careful processing.

This talk uses fully transparent open Whisper-style speech models (OWSM) as a case study and focuses on the groundwork behind them. It shows how OWSM v4 narrowed a web-scale pool of multilingual speech, around 500K hours from YODAS v2, to a usable training set through alignment, language identification, and CTC-based scoring. It also examines how the same kind of pipeline carries over to YODAS v3, a larger, more multilingual release recorded at higher fidelity.

The talk argues that task-specific models are not a stepping stone to be replaced by foundation models; they are the tools that make the data usable. It presents the Phonetic Open Whisper-Style Speech Foundation Model (POWSM), which applies the same recipe to phone recognition, showing that task-specific linguistic and phonetic tools are key to this groundwork.

The talk concludes with a brief discussion of unified speech models spanning generation and understanding as a direction that follows the same open-science principles.

Bio: Shinji Watanabe is an Associate Professor at Carnegie Mellon University's Language Technologies Institute (LTI), where his research focuses on advancing speech processing, conversational AI, and end-to-end neural models for speech and language applications. He has made groundbreaking contributions to areas such as speech recognition, speech translation, speaker diarization, and speech enhancement. As the leader of the widely used ESPnet open-source toolkit, Dr. Watanabe has developed state-of-the-art tools for end-to-end speech processing, garnering significant recognition in the research and developer communities. His work has been presented at leading venues, including Interspeech and ICASSP.

Dr. Watanabe earned his Ph.D. in Computer Engineering from Waseda University, where his research centered on Bayesian approaches to speech recognition. He has held positions at NTT Communication Science Laboratories, Mitsubishi Electric Research Laboratories (MERL), and Johns Hopkins University before joining CMU. Dr. Watanabe has published over 500 peer-reviewed articles, many of which have received accolades, such as the Best Paper award from Interspeech in 2024 and EMNLP in 2024. He has also been recognized as an IEEE and ISCA Fellow for his fundamental contributions to speech recognition research and leadership.

At CMU, Dr. Watanabe teaches advanced courses such as Speech Technology for Conversational AIand Speech Recognition and Understanding,mentoring a diverse group of students and postdoctoral researchers. He actively serves the research community as an area chair or committee member for top conferences and chair of the IEEE Signal Processing Society Speech and Language Technical Committee (SLTC). Through his research, teaching, and service, Dr. Watanabe continues to push the boundaries of speech technology and conversational AI.

Invited Talk

Beyond the Transcript: Safety and Trustworthiness Challenges in Audio-Language Models

Hung-yi Lee
National Taiwan University
Tuesday, November 10, 2026 –

Abstract: Recent advances in speech and audio-language models have enabled increasingly natural voice interactions and sophisticated auditory reasoning. However, extending language models from text to speech introduces new safety and trustworthiness challenges that cannot be fully captured by text-based evaluations. Drawing on recent research, this talk highlights several overlooked vulnerabilities of audio-language models.

The research shows that different emotional deliveries of the same harmful instruction can lead to substantially different safety responses, while speaker characteristics such as gender and accent can introduce systematic biases. The talk also examines whether audio-language models truly ground their responses in acoustic evidence, revealing vulnerabilities such as sensitivity to irrelevant sounds and failures to recognize unanswerable questions.

The talk concludes with approaches to improving audio grounding and mitigating hallucinations. Together, these findings highlight the need for speech-aware approaches to evaluating and improving the safety and trustworthiness of audio-language models.

Bio: Hung-yi Lee is a professor in the Department of Electrical Engineering at National Taiwan University (NTU), with a joint appointment in the Department of Computer Science and Information Engineering. His recent research focuses on self-supervised learning and foundation models for speech. He received the Salesforce Research Deep Learning Grant in 2019, the AWS ML Research Award in 2020, the Outstanding Young Engineer Award from the Chinese Institute of Electrical Engineering in 2018, the Young Scholar Innovation Award from the Foundation for the Advancement of Outstanding Scholarship in 2019, the Ta-You Wu Memorial Award from Taiwan's Ministry of Science and Technology in 2019, and the 59th Ten Outstanding Young Persons Award of Taiwan in the category of Science and Technology Research and Development. He is a Fellow of the International Speech Communication Association (ISCA). He also runs a YouTube channel that teaches deep learning technologies in Mandarin and has more than 400,000 subscribers.

Table of Contents

What Frozen Speech Embeddings Reveal About Audio Deepfake Detection Difficulty
Yashas Harshad Patil, Krish Rawat and Anup Shanbhag 1

TeleASR-Amharic: A Benchmark for Robust Amharic ASR in Realistic Telecommunication Conditions
Dagnachew Mekonnen Marilign, Enrique Molero, Abrham Haile, Louis Powell and Israel Abebe Azime 11

Transferring Text Jailbreaks to Audio in Tool-Using LLM Agents
Kashyap R Manjusha and Murali Emani 24

Activation-Level Characterization of Cross-Model Adversarial Jailbreak Transferability in Speech Language Models
Rahaf Alshahrani and Ayana Niwa 29

Human Perception of Arabic Fake Audio
Elsayed Issa and Yi-Fang Cheng 35

Boosting Transferable Adversarial Attacks against Deep Reinforcement Learning
Zexin Li, Ruili Yao, Yiming Zeng and Xiaoxue Gao 50

What Frozen Speech Embeddings Reveal About Audio Deepfake Detection Difficulty

Yashas Harshad Patil **Krish Rawat** **Anup Shanbhag**
Washington High School James Logan High School Dublin High School
Fremont, California, USA Union City, California, USA Dublin, California, USA
yashaspatil@gmail.com krishrawat0222@gmail.com
anupshanbhag4@gmail.com

Abstract

Audio deepfake detectors perform unevenly across synthesis systems. Can frozen speech representations expose structure behind this variation? We construct a simple direction between genuine and synthetic WavLM embeddings and characterize each system's position and spread along it. Across 61 MLAAD text-to-speech system–language groups, projected spread is strongly associated with mean detector difficulty across three runs ($\rho = 0.60$). It is the only feature surviving correction over 83 recorded tests, and leave-one-system-out ridge prediction gives $R^2 = 0.277$. The genuine–synthetic directions also differ markedly across corpora, revealing domain-dependent representation geometry. As an auxiliary utterance score, the centroid projection reduces MLAAD detector equal error rate from 27.2% to 16.3%, improving on both component scores. Supervised probe comparisons further identify where adaptation gains originate. Together, these results establish a compact representation-space diagnostic of attack-specific detector behavior; the informative statistic and benefits of score adaptation depend on the corpus and detector.

1 Introduction

An audio deepfake detector can achieve a useful average error rate while missing speech from particular synthesis systems. It can also lose accuracy when recordings come from a different dataset (Müller et al., 2022). Understanding this uneven performance is an evaluation problem in its own right: what structure in speech representations helps describe why an existing detector succeeds on some synthesis systems and struggles with others?

We study this question using frozen WavLM representations (Chen et al., 2022). We average frame representations into one vector per utterance and use labeled speech to construct a direction from the genuine centroid to the synthetic centroid. Each utterance then has an inspectable scalar projection. The construction is deliberately simple: it lets us examine detector-relevant structure already present in the representation without modifying the detector architecture or fitting another classifier.

Our central finding concerns the spread of these projections within each synthesis system. Across 61 MLAAD text-to-speech system–language groups (Müller et al., 2024), wider spread is associated with greater WavLM-GAT difficulty, measured by one minus the area under the receiver operating characteristic curve (AUC). The rank correlation is 0.60, and leave-one-system-out ridge prediction gives $R^2 = 0.277$. Spread is the only feature surviving correction across the recorded 83-test screen. A one-dimensional geometric summary thus captures meaningful variation in the behavior of a much richer detector.

Position and spread offer complementary views of that structure. A wide synthetic distribution can place more utterances near genuine speech, while a narrow distribution can also be difficult when centered close to genuine speech. We make this distinction explicit in the method and schematic. Across corpora, the genuine–synthetic direction changes substantially, and different geometric summaries become informative. This provides a representation-space perspective on domain-specific detector behavior.

We also examine the direction as an auxiliary score for individual utterances. On MLAAD, centroid fusion improves on both the detector and the projection alone. Comparisons with supervised probes show which gains come from a simple centroid score and which come from fitting a stronger target-domain classifier. These are secondary applications of the geometric analysis; the primary contribution is understanding variation across synthesis systems.

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 1–10
November 10, 2026 ©2026 Association for Computational Linguistics

Our contributions are:

- **Geometric diagnostic.** We link within-system spread along a labeled genuine–synthetic direction to detector-specific synthesis-system difficulty.
- **Statistical evidence.** We characterize the MLAAD relationship across 61 systems and three detector runs using screen-wide correction, resampling, permutation analysis, and leave-one-system-out prediction of the selected feature.
- **Domain geometry.** We document substantial cross-corpus changes in direction, alongside stable within-corpus estimates.
- **Score augmentation and decomposition.** We evaluate when geometric scores complement existing detectors and separate centroid-based gains from supervised probe adaptation.

2 Background and evaluation setting

Self-supervised speech representations support spoofing detection through learned downstream classifiers (Tak et al., 2022; Wang and Yamagishi, 2022). Our primary detector, denoted WavLM-GAT, follows Zhang et al. (2025): WavLM features feed adaptive phoneme pooling, graph attention, a bidirectional LSTM, and a classification head. The WavLM-base encoder has approximately 95 million parameters and is frozen; the downstream network has approximately 13 million trainable parameters. AASIST (Jung et al., 2022) provides a second detector architecture. Its spectro-temporal network does not use WavLM, although the geometric scores added to it in our analysis still do.

Difference-of-means directions summarize a labeled distinction in an embedding space. Related constructions have been used to analyze word embeddings and language models (Bolukbasi et al., 2016; Arditi et al., 2024). Whereas SSL-based anti-spoofing methods use representations as inputs to learned detectors (Tak et al., 2022; Wang and Yamagishi, 2022), we analyze the frozen representation itself and ask whether a simple labeled direction captures variation in detector difficulty across synthesis systems. Unlike detector logits or trained classifier probes, the direction is estimated directly from class centroids. Unlike generic embedding distances, it separates position and spread along a specified contrast. Its simplicity makes the resulting diagnostic easy to inspect.

The primary analysis uses 61 MLAAD text-to-speech (TTS) system–language groups with at least eight utterances per group, abbreviated as systems below. The cached test subset contains 908 genuine and 938 synthetic clips across 63 such groups; two groups with fewer than eight clips are excluded from the system-level analysis. Additional analyses use ASVspoof 2019 logical access (LA) (Wang et al., 2020), ASVspoof 2021 (Yamagishi et al., 2021), and In-the-Wild (ITW) (Müller et al., 2022). ASVspoof groups utterances by attack identifier, while the ITW analysis groups them by speaker. The ITW results therefore describe speaker-level variation. Appendix A lists the analyzed subsets.

For detector score $d(x)$ oriented so that larger values indicate synthetic speech, system s has difficulty

$$h_s = 1 - \mathrm{AUC}\big(d(B), d(S_s)\big), \tag{1}$$

where B is the shared genuine-speech reference pool and S_s contains synthetic utterances from system s. Perfect ranking gives $h_s = 0$, chance ranking gives $h_s = 0.5$, and reversed rankings can give larger values. This is a property of a detector–dataset pair, not an intrinsic measure of synthesis quality. It does not require choosing a detection threshold.

Three saved WavLM-GAT checkpoints produce similar MLAAD difficulty rankings: pairwise Spearman correlations range from 0.89 to 0.97. The primary regression target is the mean of their three difficulty values for each system. This agreement supports studying a shared pattern of system-level difficulty across these detector runs.

3 Measuring geometry and difficulty

3.1 A labeled direction and scalar projections

For each 16-kHz, three-second clip, let $e_i \in \mathbb{R}^{768}$ be the mean of the frozen WavLM layer-12 frame embeddings. Layers 0, 9, and 12 were examined; layer 12 gave the strongest axis-based result. We use layer 12 following this exploratory comparison.

Let $\mu_{B,j}$ and $a_{B,j}$ be the mean and population standard deviation of coordinate j in the genuine-speech reference data. We standardize coordinates

as

$$\widetilde{e}_{ij} = \frac{e_{ij} - \mu_{B,j}}{a_{B,j} + \epsilon}. \tag{2}$$

with $\epsilon = 10^{-9}$. For a held-out synthesis system s, let S_{-s} contain the synthetic reference utterances from the other systems. The centroid difference and its unit direction are

$$v_{-s} = \text{mean}_{i \in S_{-s}} \widetilde{e}_i - \text{mean}_{i \in B} \widetilde{e}_i, \tag{3}$$

$$u_{-s} = v_{-s} / \|v_{-s}\|_2. \tag{4}$$

The scalar projection of utterance i from system s is $p_i = u_{-s}^{\mathsf{T}} \widetilde{e}_i$. Thus projection means a dot product with a unit vector. Its sign is fixed from genuine toward synthetic speech. The direction uses genuine/synthetic class labels and does not use detector difficulty values h_s.

For the n_s synthetic utterances in system s, define their mean projection and spread as

$$m_{\parallel}(s) = \frac{1}{n_s} \sum_{i \in S_s} p_i, \tag{5}$$

$$\sigma_{\parallel}(s) = \sqrt{\frac{1}{n_s} \sum_{i \in S_s} \big(p_i - m_{\parallel}(s)\big)^2}. \tag{6}$$

The spread is calculated within one synthetic system. It neither pools genuine and synthetic utterances nor combines their standard deviations. A large separation between the class centroids therefore does not itself make $\sigma_{\parallel}$ large. Equation 6 follows the population-SD convention of the audited headline result. The original exploratory table used sample SD; Appendix A records this distinction. For exploratory position comparisons, the saved statistic is the median projection, rather than $m_{\parallel}$.

3.2 Why both position and spread matter

Figure 1 distinguishes location from spread. Holding a synthetic mean on the synthetic side of the genuine distribution, increasing its spread can place more utterances near genuine speech. Conversely, a tightly clustered system can still be difficult if its mean lies inside the genuine distribution. A distant mean with a small spread can be easy to separate. For a distant mean and a large spread, the answer depends on their relative magnitudes and the distribution shapes.

Projected overlap supplies intuition for the diagnostic. WavLM-GAT and AASIST use richer decision functions, so the empirical question is whether spread along this simple direction covaries with their measured difficulty across systems.

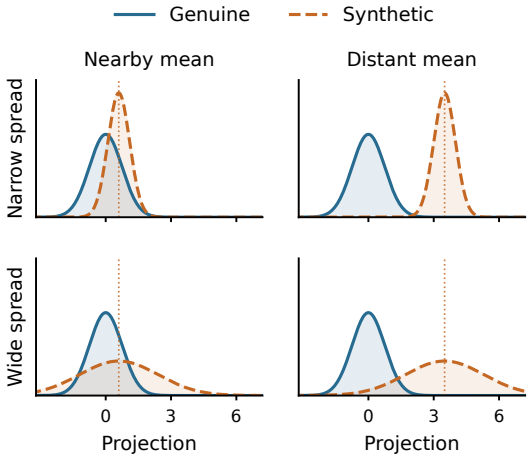


Figure 1: Position and spread both affect projected overlap. Each panel shows illustrative densities of genuine speech (blue) and one synthetic system (orange); these are schematics, not observations. Columns change the synthetic mean and rows change its spread. The same spread can accompany very different overlap, illustrating why position and spread describe distinct properties.

3.3 Association and held-out prediction

We distinguish two uses of leaving a system out. For geometric measurement, its utterances are excluded when constructing its reference direction. For predicting difficulty, ridge regression with regularization $\alpha = 1$ is fitted to the other systems' measured pairs $(\sigma_{\parallel}, h)$ and predicts the held-out system. The scalar predictor is standardized across the 61 systems before these regression folds. The reported prediction statistic is

$$R^2 = 1 - \frac{\sum_s (h_s - \widehat{h}_s)^2}{\sum_s (h_s - \bar{h})^2}. \tag{7}$$

A negative value indicates worse prediction than the overall-mean reference in this definition. The regression evaluates the selected feature using a fixed table of system measurements; feature discovery preceded this analysis. We interpret the result as within-MLAAD evidence for that feature. Appendix A specifies the feature-table construction and scaling.

Spearman's ρ measures rank association across systems. The recorded screen contains 83 predictor–difficulty tests. We report its global

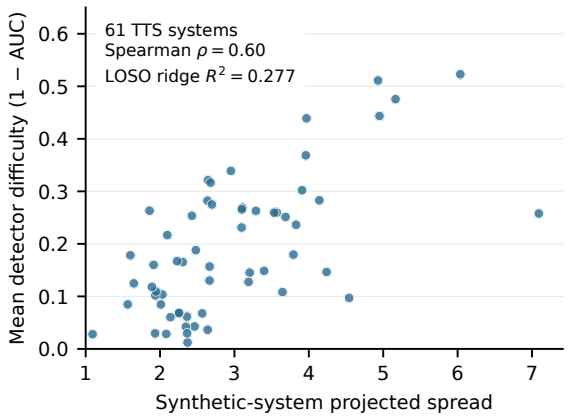


Figure 2: MLAAD system-level geometry and detector difficulty. Each point is one cached synthesis system–language group; difficulty averages three WavLM-GAT runs. Spread is expressed using the population-SD convention. The annotations summarize the audited association and ridge regression defined in Section 3.3.

Benjamini–Hochberg false-discovery-rate correction (Benjamini and Hochberg, 1995), using $q < 0.05$ as the significance threshold. Its rank-correlation p-values and the separate regression permutation test are distinct. The latter compares predictive performance with 4,000 shuffles of system difficulties (Ojala and Garriga, 2010). The permutation test keeps the geometric feature table fixed.

4 Geometric structure of detector difficulty

4.1 Spread predicts MLAAD difficulty

On 61 MLAAD systems, spread correlates positively with WavLM-GAT difficulty: $\rho = 0.60$, with a reported bootstrap 95% interval of $[0.41, 0.73]$ (Figure 2). The leave-one-system-out ridge regression attains $R^2 = 0.277$ with permutation $p = 2.5 \times 10^{-4}$. Table 1 compares representative features from the recorded screen. The spread statistic is its sole global FDR survivor, with $q \approx 3.1 \times 10^{-5}$. Projection position and axis-independent dispersion have smaller absolute correlations and do not pass this correction.

The association is stable across the available sensitivity checks. Removing one system at a time gives correlations from 0.58 to 0.63. Partial rank correlations controlling for root-mean-square signal energy or utterance count are both approximately 0.58.

The association is stronger when only systems with more utterances are retained: $\rho = 0.59$ for 61 systems with at least eight utterances, 0.64 for 51 systems with at least 12, and 0.69 for 28 systems with at least 16. Greater measurement precision is one possible explanation. Because filtering also changes the system population, we interpret this as a sensitivity check.

MLAAD feature	ρ	Global q
Projected spread, layer 12	$+0.597$	0.000031
Median projection	-0.262	0.148
Centroid norm, layer 12	$+0.289$	0.109
Radius of gyration, layer 0	-0.381	0.059
Radius of gyration, layer 12	-0.308	0.094
Trajectory entropy, layer 9	$+0.352$	0.072

Table 1: Representative entries from the 83-test screen. Radius of gyration measures dispersion without choosing an axis. Only projected spread passes global $q < 0.05$. The global correction spans all 83 recorded tests, including the additional corpora.

The raw $R^2 = 0.277$ is our primary prediction measure. A split-half analysis estimated difficulty reliability at 0.759, with a full-sample correction of 0.863. Their ratio, $0.277/0.863 \approx 32\%$, provides secondary context under that reliability model rather than a causal variance decomposition.

4.2 Other corpora favor different statistics

The exploratory comparisons show that different corpora emphasize different geometric statistics. On ASVspoof 2019 LA, median position along the direction correlates with attack difficulty at $\rho = -0.71$ over 13 attacks ($p = 0.006$). The ITW speaker analysis instead reports a correlation with orthogonal position ($\rho = 0.53$, 29 speakers, $p = 0.003$). Here orthogonal position summarizes the norm of the embedding component perpendicular to the direction. An axis-independent dispersion statistic has a slightly stronger ITW correlation ($\rho = -0.56$); neither has a clear independent association after controlling for the other. These are exploratory associations with different statistics and grouping units. ASVspoof 2021 within-corpus position correlations of 0.60 and 0.64 for the two detectors do not survive the reported family correction.

MLAAD emphasizes spread, whereas the smaller ASVspoof and ITW samples emphasize position-related or axis-independent statistics. This motivates studying which properties of the genuine and synthetic distributions determine the most informative summary. Their means, disper-

Corpus pair	MLAAD frame	Raw
MLAAD / ITW	$+0.103$	$+0.027$
MLAAD / ASVspoof 2021	-0.207	-0.140
ITW / ASVspoof 2021	-0.238	-0.118

Table 2: Audited cross-corpus direction cosines. The middle column uses the same MLAAD genuine-speech scaling for all corpora; the last uses raw embeddings. These values depend on the coordinate metric.

sion, and other structure can all affect projected overlap.

Using WavLM spread to rank fine-tuned AASIST difficulty on MLAAD gives a smaller association, $\rho = 0.35$ ($p = 0.006$, global $q = 0.072$). Its leave-one-system-out regression gives $R^2 = -0.032$ ($p = 0.13$). The weaker result highlights the detector-specific nature of the diagnostic: geometry is shared, but difficulty depends on the decision function.

4.3 Corpus directions reveal domain dependence

For two directions u_a, u_b, cosine similarity is $u_a^\mathsf{T} u_b$ when both are unit normalized. Values near 1 indicate similar orientations, values near 0 indicate weak alignment, and negative values indicate opposing components. Table 2 compares corpus directions in two coordinate metrics. We also assess stability within each corpus by splitting its data in half. Split-half cosines measure how consistently the direction is estimated within the sampled corpus.

The audited pairs have weak or negative alignment, while within-corpus split-half cosines are 0.926 for MLAAD, 0.982 for ITW, and 0.981 for ASVspoof 2021. This contrast is consistent with domain-specific representation geometry. The marked change in direction offers a geometric perspective on why a detector-relevant signal may not transfer unchanged across corpora. The magnitudes depend on the coordinate metric, as the two columns show.

We treat these cosines as descriptive measurements. Transferring the hardness predictor to 13 ASVspoof 2021 attacks gives $\rho = 0.25$ and $p = 0.41$, leaving cross-corpus prediction an open question in this small sample.

5 Using labeled geometry to adapt scores

5.1 Centroid projection and supervised probes

The preceding sections use geometry to describe differences among synthesis systems. Here we ask a separate question: can the same labeled direction supply a score for individual utterances? For centroid fusion, that value is its projection $p(x)$ onto a direction estimated from reference data. If z_d and z_p standardize detector and projection scores using training-partition means and standard deviations, the fused score is

$$f(x) = z_d(d(x)) + \lambda z_p(p(x)). \tag{8}$$

Both components are oriented toward synthetic speech. In the MLAAD study, the coefficient λ maximizes training-partition AUC over $\{0, 0.25, 0.5, 0.75, 1, 1.5, 2, 3\}$. The axis and coefficient use the same training partition, rather than a separate inner validation split. Test labels are excluded from these choices. No per-system difficulty labels enter this fusion rule.

A supervised linear discriminant analysis (LDA) probe is a different score. It estimates a class-separating linear function using the class means and within-class covariance of the embeddings. The implementation fits raw embeddings using the least-squares solver with automatic covariance shrinkage. The logistic comparator fits genuine-standardized embeddings with inverse regularization $C = 0.1$ and a 2,000-iteration limit. These probes use target-domain class labels and estimate more structure than a centroid difference. Their component-only results separate the value of this additional fitting from the centroid construction.

Score adaptation uses labeled genuine and synthetic calibration examples to estimate the direction, standardize scores, and select the coefficient; the detector weights remain fixed. A supervised probe additionally fits a classifier. WavLM-GAT can reuse its existing representations, whereas adding these scores to AASIST requires a WavLM encoder. This distinction determines the added inference cost.

5.2 Centroid fusion benefits detectors with headroom

Table 3 reports equal error rate (EER), the error level where the false-positive and false-negative rates are equal as the threshold varies. All EER

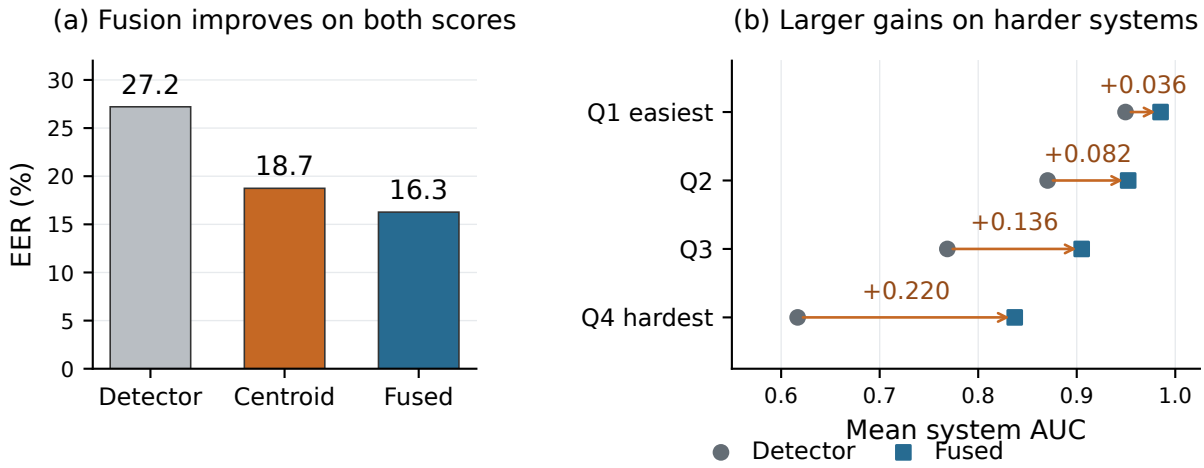


Figure 3: Two views of MLAAD centroid fusion, redrawn from the original saved results. (a) Mean EER across three WavLM-GAT runs: fusion improves on the detector and centroid score alone. (b) Mean system AUC within quartiles defined by baseline detector difficulty; arrows label absolute AUC gains. The largest gains occur in the hardest quartile. This breakdown conditions on baseline difficulty and describes the evaluated systems.

WavLM-GAT setting	Detector	Axis	Fused
ASVspoof 2019, weak	14.4	—	11.4
MLAAD, within corpus	27.2	18.7	16.3

Table 3: Centroid-projection fusion EER (%). The weak detector was trained with 500 utterances. An axis-alone value for that row is not reported here. The MLAAD fusion improves on both component scores.

values in this paper are percentages; differences are percentage points. EER measures the quality of a score ranking and does not supply a deployment threshold.

For the deliberately undertrained ASVspoof 2019 detector, EER decreases by 3.0 points, a relative reduction of approximately 21%. The reported paired-bootstrap interval for the EER difference is $[-4.6, -1.3]$ points, with $p < 10^{-3}$. All five evaluated folds improve. A stronger WavLM-GAT starting at 7.8% EER shows small changes, from approximately -0.04 to $+0.4$ points across the reported variants, with p between 0.56 and 0.80. The stronger detector changes little, with both small improvements and small increases in error.

On MLAAD, the fused EER of 16.3% improves on both the detector's 27.2% and the centroid score's 18.7% (Figure 3a). Five folds hold out synthetic system–language groups and independently partition genuine utterances; this is repeated for three detector runs. Each fold estimates the scaler, axis, score standardizers, and coefficient on its training portion. Improvements occur across all 15 fold-by-run combinations. The selected fusion coefficients are 1.5 or 2.0. Across these settings, larger gains occur for weaker starting detectors. The MLAAD component comparison shows that the centroid score supplies complementary information in that setting.

The original per-system breakdown locates these gains within MLAAD (Figure 3b). Mean AUC gains increase from 0.036 in the easiest baseline-difficulty quartile to 0.220 in the hardest. Quartiles are defined by the three-run mean baseline difficulty, and each point averages system AUCs across systems and runs. This descriptive comparison shows where the observed gain is concentrated, with greater room for improvement in the harder groups.

5.3 AASIST separates probe information from fusion gains

The AASIST comparison reveals where the adaptation gain comes from. For the unadapted checkpoint evaluated outside its training domain, fusion reduces EER by 26.0 points on MLAAD and 32.5 points on ITW (Table 4). The supervised probe alone is stronger still, outperforming fusion by 1.6 and 6.0 points, respectively. Target-domain WavLM embeddings therefore contain substantial discriminative information that the transferred detector does not exploit; in these rows the probe captures all of the gain.

The original ITW protocol permits genuine

Target corpus / split	AASIST	LDA	Fused
MLAAD	37.6	**10.0**	11.6
ITW, original	48.6	**10.1**	16.1
ITW, speaker-disjoint	—	**14.3**	19.3
ASVspoof 2021	7.3	11.5	7.6

Table 4: Supervised LDA adaptation of AASIST, EER (%). Original ITW folds share genuine speakers across training and test. The saved stricter audit separates genuine speakers as well. A separately reported AASIST value for that audit is unavailable. The LDA probe outperforms fusion on MLAAD and under both ITW protocols.

Corpus	Added score	5	25	250
MLAAD	Centroid	28.3	32.7	16.6
	LDA	—	32.3	13.6
ITW	Centroid	34.9	31.2	28.7
	LDA	—	28.7	19.2

Table 5: Calibration sweep: fused EER (%) at the indicated number of labeled examples per class. LDA is evaluated only from 25 examples per class. Calibration groups and test subsets change across draws and sizes; the entries are not a controlled learning curve on a fixed test set.

speakers to appear in both training and test folds. Separating genuine speakers raises LDA-alone EER to 14.3% and fused EER to 19.3%. Table 4 retains both protocols to show the effect of speaker separation.

The MLAAD probe comparison provides further context: LDA alone achieves 10.0% EER and logistic regression 11.2%, compared with the centroid score's 18.7%. Their mean EER reductions when fused with WavLM-GAT are approximately 15.7 and 15.4 points, respectively, versus 10.9 points for centroid fusion. The probe fusion uses fixed $\lambda = 1.5$, whereas centroid fusion selects its coefficient. A comparison under matched label and compute budgets would further clarify these tradeoffs.

5.4 Calibration-size evidence

A calibration sweep evaluates 5, 10, 25, 50, 100, and 250 labeled examples per class, with 20 attempted group-aware draws at each size. It fits the direction and fusion coefficient on the calibration observations and excludes their groups from evaluation. ITW uses speaker groups; MLAAD uses synthesis groups for synthetic speech. This sweep adapts WavLM-GAT, so it is separate from the AASIST comparison.

Table 5 gives selected sizes. Small calibration sets yield variable errors, and centroid results are nonmonotonic. The saved detector baseline uses the full subset, while adapted scores exclude calibration groups; we report adapted EERs only. The sweep describes the evaluated calibration settings, with changes in group coverage and test composition limiting label-efficiency conclusions.

6 Discussion

The main finding is that a simple direction in frozen speech representations captures meaningful variation in detector difficulty across MLAAD synthesis systems. The diagnostic makes this variation inspectable: systems can be compared by how their synthetic utterances occupy the genuine–synthetic direction, alongside their observed errors. Direct AUC measures performance; projected spread supplies a compact representation-space description of the failure structure. In an evaluation with labeled genuine and synthetic examples, one can report detector AUC and projected spread by synthesis group, inspect high-spread groups for concentrated errors, and prioritize them for further testing. Because the informative statistic changes across corpora, this ranking should be derived within the target evaluation domain.

The cross-corpus results extend this perspective. Directions are stable within the sampled corpora yet differ markedly between them, and the informative summary changes with the dataset. Together, these observations motivate studying the distributional conditions under which spread, position, or other geometry best describes difficulty. As expected for a one-dimensional diagnostic, substantial variation remains for content, speaker, channel, synthesis characteristics, and other representation features to explain.

The score experiments show a related practical use. Centroid fusion can improve on both components when a detector has headroom, as on MLAAD. Stronger supervised probes capture additional target-domain information, and the AASIST decomposition identifies the probe as the source of its largest gains. Reporting component-only scores makes the contribution of each source of information visible. These findings support representation-space analysis as a tool for understanding detector behavior and evaluating adaptation choices.

7 Conclusion

Simple structure in frozen WavLM representations captures meaningful variation in audio deepfake detection difficulty. Across 61 MLAAD synthesis system–language groups, projected spread provides a compact hardness signal, supported by the recorded feature screen and leave-one-system-out ridge analysis. Cross-corpus direction changes reveal substantial domain dependence. Labeled geometric scores can also complement weaker detectors, while supervised probes expose additional information in target-domain embeddings. Together, these findings motivate representation-space diagnostics for understanding and auditing detector failures across synthesis systems. Further evaluation across encoders, synthesis families, and independently held-out corpora would clarify their reach.

Limitations

All geometric measurements use WavLM; generality to encoders such as HuBERT, wav2vec 2.0, and Whisper remains untested. The primary finding concerns text-to-speech systems, while the available voice-conversion results do not show the same association. Projected spread (`sd_along`) and layer 12 emerged from exploratory analysis within MLAAD, motivating independent replication on additional corpora. The analyzed MLAAD subset is English-focused, so the evidence should not be read as broad multilingual validation. Cross-corpus analyses suggest that the most informative geometric statistic is domain dependent, and transfer to new datasets remains an open question. Target-domain geometric adaptation requires labeled bona-fide and synthetic examples. The adaptation experiments compare useful alternatives, but their label budgets, coefficient-selection procedures, and compute costs are not fully matched across methods; comparisons under identical budgets would clarify their relative value.

Ethical considerations

This work uses existing research datasets to characterize detector weaknesses and improve robustness evaluation. False accusations of manipulation can harm speakers, while missed synthetic speech can enable impersonation. Aggregate AUC and EER are insufficient evidence of authenticity for an individual recording, and detector outputs should not be used alone for such judgments.

Speech can reveal identity and personal content; use and redistribution remain subject to the recordings' original terms and consent conditions. Representation-space diagnostics may also reveal properties that an adaptive adversary could exploit. Our purpose is defensive evaluation of detector failures.

Acknowledgments

We thank Jathin Korrapati, Ryan Lagasse, Archana Vaidheeswaran, Arjun Bahuguna, and Kevin Zhu for their mentorship, guidance, and support throughout this research.

OpenAI Codex assisted with manuscript organization, citation and formatting checks, and simulated reviewer feedback. It did not run experiments or generate empirical results.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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*, 57(1):289–300.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371.

Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? In *Interspeech 2022*, pages 2783–2787.

Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. MLAAD: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Markus Ojala and Gemma C Garriga. 2010. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jeeweon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 112–119.

Xin Wang and Junichi Yamagishi. 2022. Investigating self-supervised front ends for speech spoofing countermeasures. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 100–106.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, et al. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yushu Zhang, and Yifang Guo. 2025. Phoneme-level feature discrepancies: A key to detecting sophisticated speech deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1066–1074.

A Reproducibility details

A.1 Analyzed subsets and units

The MLAAD cache contains 1,846 clips (908 genuine, 938 synthetic) in 63 synthetic system–language groups. Two groups have five and six clips, leaving 61 after the minimum-eight filter. Geometry and detector scores reuse these cached representations, a subset of the published dataset.

ASVspoof 2019 position analysis uses 800 genuine and 800 synthetic clips from attacks A07–A19. ASVspoof 2021 transfer uses 400 genuine plus 40 synthetic clips per attack across 13 attacks (920 total), with codec `none`. ITW geometry uses 1,500 clips per class; its table has 53 speakers, of whom 29 have at least ten synthetic clips for difficulty analysis. These corpora differ in sampling and grouping.

A.2 Encoder and detector artifacts

The geometry encoder is the Hugging Face `microsoft/wavlm-base` checkpoint, evaluated without gradient updates. The loader explicitly restores positional-convolution weight-normalization parameters from the pretrained checkpoint when the installed Transformers/PyTorch combination exposes the newer parametrization interface. This prevents legacy `weight_g`/`weight_v` tensors from being left unmatched. Layer-12 hidden states are averaged over the time dimension. Cached clips have 48,000 samples at 16 kHz; the MLAAD preprocessing uses center crops, and the ITW helper center-crops long recordings and pads short ones with zeros at the end.

The primary MLAAD score files correspond to saved checkpoints labeled `main`, `s42`, and `s1024`. The main detector was trained on the training split of MLAAD-tiny and selected using its validation split. Its training record lists 10,348 training and 2,218 validation clips, seven epochs, batch size 20, and best validation epoch 5. Its augmentation probability is 0.35, with noise/pitch/reverberation sampling weights of 0.5/0.3/0.2. The saved settings use noise SNRs from 8 to 40 dB, upward pitch shifts from 1 to 5 semitones, and reverberation times from 0.2 to 0.9 seconds. The replication checkpoints are selected at validation epochs 5 and 3. Evaluation is within MLAAD; generator, speaker, and text overlap with detector training is unverified.

The fine-tuned AASIST checkpoint supplies the rank-association comparison in Section 4.2. The transferred, unadapted checkpoint supplies Table 4. The anonymous companion tables contain the audited per-system values, counts, and all 83 recorded feature tests. They support inspection of the reported analysis, but do not replace the raw-audio split manifests, training code, and checkpoints needed for end-to-end reproduction.

A.3 Geometry conventions and regression

Coordinate standard deviations use the population convention and add 10^{-9} to avoid division by zero. The axis uses a clip-weighted synthetic centroid: systems with more reference clips contribute more to that mean. The audited projection is $(\widetilde{e}_i -$

$\widetilde{\mu}_B)^\mathsf{T} u_{-s}$; centering by $\widetilde{\mu}_B$ is equivalent to Equation 4's projection up to numerical roundoff, because genuine-centered coordinates have zero genuine mean. This extra centering does not change within-system spread.

The exploratory table computed sample standard deviations with denominator $n_s - 1$. The audited headline recomputation used population standard deviations with denominator n_s, yielding $R^2 = 0.276647$ and $\rho = 0.597462$. Figure 2 directly uses the saved audited feature table and three-run mean difficulty, matching the convention in Equation 6. The audit reconstructed these features from cached embeddings. We retain its population convention in the formal definition and report its rounded statistics.

The ridge predictor includes an intercept and uses $\alpha = 1$. Standardizing its input once across all systems precedes the regression leave-one-out loop. System difficulties average the three detector runs; labels are synthetic-positive. The feature table is constructed once: each system has its own leave-one-system-out reference direction, and those measurements are reused across the regression folds. The permutation analysis also reuses this table and does not repeat the preceding feature screen.

The headline audit uses 4,000 difficulty permutations with random seed 1, reporting $(1 + k)/(4000 + 1)$ when k null scores meet or exceed the observed R^2. The resulting $p = 0.00024994$ is rounded in the text. Its rank-correlation confidence interval uses 10,000 bootstrap resamples with seed 2, sampling the 61 system-level $(\sigma_\parallel, h)$ pairs with replacement and keeping their measured spread and AUC values fixed. The uncorrected rank-correlation p-value is 3.71×10^{-7}; it is not the permutation p-value for R^2.

A.4 MLAAD score fusion and probes

The score-fusion implementation shuffles sorted synthetic group identifiers with seed 42 and divides them into five folds. It separately permutes genuine utterance indices into five folds. In each round, four folds estimate the genuine coordinate scaler and centroid direction. Projection and detector-score standardizers use the training observations from both classes, with population standard deviations. The coefficient grid and training-AUC criterion are specified in Section 5. Predictions are concatenated across test folds before calculating each run's EER; the headline averages the three run-level EERs.

For the three detector runs, centroid fusion selects λ values $[1.5, 1.5, 1.5, 1.5, 2]$, $[1.5, 2, 2, 1.5, 2]$, and $[1.5, 1.5, 1.5, 1.5, 2]$. The unrounded mean EERs are 0.272113 for the detector, 0.187429 for the centroid score, and 0.162693 for fusion. The paired utterance bootstrap uses 2,000 resamples and seed 42; the mean EER-difference interval is approximately $[-0.1286, -0.0908]$. These are utterance-resampling intervals, not uncertainty intervals over new corpora or synthesis families.

The LDA comparator uses `solver=lsqr` and `shrinkage=auto` on raw 768-dimensional embeddings. Logistic regression uses genuine-standardized embeddings, $C = 0.1$, and at most 2,000 iterations. Both use the same five synthetic-group/genuine-utterance folds as the centroid comparison. Their fusion comparison fixes $\lambda = 1.5$ and fits score standardizers on training observations. This differs from the centroid coefficient search and should be considered when comparing their gains.

The AASIST adaptation audit uses the same shrinkage-LDA specification. Its original protocol holds out synthetic groups but randomly splits genuine utterances; the stricter ITW audit also separates genuine speakers. Table 4 retains both ITW rows to show the effect of this choice.

A.5 Calibration-size sweep

For size n and repetition r, calibration uses seed $1000r + n$. The script permutes class-specific group IDs, collects at least n utterances, and takes the first n; selected groups are excluded from testing. Draws with fewer than 200 test clips or one class are skipped. MLAAD genuine clips use five deterministic index-based groups, not speakers; ITW groups both classes by speaker. Independence therefore differs by corpus.

Calibration AUC selects the coefficient from $\{0.5, 1, 1.5, 2\}$, with no detector-only option. The same LDA specification is attempted for $n \geq 25$. Means average retained draws, then detector runs (three for MLAAD, one for ITW). Adapted EER uses changing test subsets; detector-only EER uses the full corpus subset, so saved improvements are unmatched. Table 5 reports adapted EERs only, not performance on a fixed external test set.

TeleASR-Amharic: A Benchmark for Robust Amharic ASR in Realistic Telecommunication Conditions

Dagnachew Mekonnen Marilign[1], **Enrique Molero**[2], **Abrham Haile**[3], **Louis Powell**[2], **Israel Abebe Azime**[4]

[1] HiLCoE School of Computer Science and Technology, [2] GSMA , [3] gheero,

[4]Saarland University

Abstract

The real-world usability of ASR models remains challenging, particularly for low-resource languages, where existing speech resources are often collected in clean studio environments. In this work, we introduce TeleASR-Amharic, a benchmark for evaluating Amharic ASR under various acoustic and telecommunication conditions. We analyze the major degradations observed in real telecommunications recordings and develop a benchmark that covers multiple channel conditions, environmental disturbances, and telecom-domain variations. Our results show that telecommunication degradation can substantially reduce recognition performance, making even high-performing ASR models impractical under severe conditions. For example, the WER of the best-performing model's increases from 0.23 to 0.79 under severe telecommunication degradation, corresponding to a relative increase of 243% in recognition errors. We further investigate the effects of different noise and channel distortions and show that scaling training data and applying carefully designed telecommunication-aware augmentation substantially improve robustness. In particular, our best training configuration achieves a 23.6% relative WER reduction under extreme noise conditions. Together with the release of our telecommunications-domain data, these findings provide a reproducible benchmark for measuring progress in Amharic ASR while highlighting open challenges in developing robust speech technologies for real-world deployment.

1 Introduction

Automatic Speech Recognition (ASR) has advanced rapidly in recent years, driven by significant developments in supervised learning, self-supervised representation learning, and multilingual modeling (Yadav and Sitaram, 2022; Mohamed et al., 2022; Ahlawat et al., 2025; Nayeem et al., 2025; Lugo and Vielzeuf, 2024). These advances, together with the emergence of large-scale speech models, have substantially expanded the range of languages, speakers, and acoustic conditions that contemporary ASR systems can handle. However, strong performance under clean or controlled speech does not necessarily translate into reliable recognition in real-world communication environments (Pranida et al., 2025a).

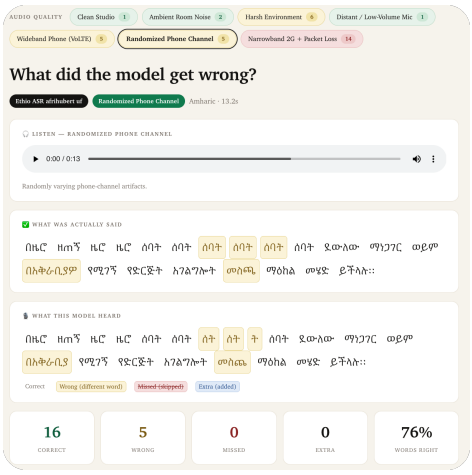


Figure 1: Examples of problematic ASR transcriptions in the telecommunications domain. The analysis shows that models frequently miss numeric information and that transcription performance varies substantially across different audio-quality conditions.

Unlike conventional clean-speech benchmarks, telecommunication speech can exhibit combinations of bandwidth limitations, sampling-rate reductions, dynamic-range compressions, nonlinear distortions, electrical interferences, background noises, and temporary signal losses (Paul and Fuji, 2026). Such distortions may occur at different stages of the communication process and can vary substantially in terms of severity. Consequently, evaluating ASR systems only on clean or naturally recorded speech provides limited information about their robustness to the conditions encountered in

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 11–23
November 10, 2026 ©2026 Association for Computational Linguistics

real-world communication channels (Imai et al., 2025; Pranida et al., 2025b). Therefore, a systematic evaluation framework that isolates and controls these factors is necessary to understand how ASR systems respond to progressively degraded speech.

Publicly available Amharic speech and transcription resources are substantially smaller than those available for high-resource languages (Nigatu and Aldarmaki, 2025; CLEAR Global, 2026). Recent efforts have expanded Amharic speech resources and demonstrated the feasibility of constructing larger corpora (Diack et al., 2026; Yeneneh et al., 2025; Radford et al., 2026). Self-supervised and multilingual models can further reduce the dependence on large amounts of transcribed data by transferring representations learned from multilingual speech (Baevski et al., 2020; Pratap et al., 2023). Nevertheless, increasing linguistic coverage does not necessarily resolve the acoustic domain mismatch. A model may adequately represent Amharic linguistic patterns while remaining sensitive to the signal degradation introduced by microphones, mobile devices, and communication networks.

To investigate this problem, we analyze the acoustic characteristics of recordings of Ethio Telecom and identify signal degradations that distinguish telecommunication speech from clean speech. Based on these observations, we constructed *TeleASR-Amharic*, a benchmark to evaluate the robustness of ASR systems under controlled telecommunication-oriented acoustic conditions. Rather than representing degraded speech as a single evaluation category, the benchmark models a range of environmental, hardware, and communication-channel distortions at different severity levels. These include bandwidth limitation, sampling-rate reduction, dynamic-range compression, nonlinear distortion, background noise, power-line interference, and temporal signal dropouts. To deal with these problems, we ask **RQ1:** How robust are Amharic-specific, multilingual, and general-purpose ASR models to telecommunication-specific acoustic degradation? Next, we explore **RQ2:** how ASR performance changes as the severity and combination of telecommunication degradations increase? to deal with this problems of domain mismatch and noise robustness we **RQ3:** explore the effect of data scaling and targeted augmentation to improve the robustness of Amharic ASR models under telecommunication conditions?

To answer these questions, we made the following contributions:

- We introduce **TeleASR-Amharic**, a benchmark for evaluating the robustness of Amharic ASR systems under controlled, telecommunication oriented acoustic conditions. Data relesed under a CC BY 4.0 license with free access[1].
- We provide a comprehensive evaluation of Amharic-specific, multilingual, and general-purpose ASR models across clean and progressively degraded conditions, enabling direct comparison of their robustness to telecommunication effects.
- We investigate the contribution of data scaling and targeted augmentation to robust Amharic ASR and analyze the behavior of the model.

2 Related Work

Amharic ASR and Low-Resource Challenge Amharic is one of Ethiopia's major languages and is widely used across public services, education, media, and digital applications. However, it remains under-resourced, with limited transcribed speech and substantial variation across speakers, dialects, devices, acoustic environments, and speaking styles (Nigatu and Aldarmaki, 2025; Imam et al., 2025; Bandarupalli et al., 2025; Bartelds et al., 2023; Liu et al., 2024; Dhasmana et al., 2026). This creates both linguistic and acoustic coverage gaps, as existing corpora often lack the variability encountered in real-world deployment. Therefore, we evaluate Amharic ASR beyond clean speech, focusing on robustness to realistic telecommunication conditions and domain-specific content, including numbers, terminology, and expressions used in customer-support interactions.

Telecommunication Speech as a Distinct ASR Domain Telecommunication speech represents a distinct ASR domain because transmission channels introduce systematic acoustic transformations beyond additive noise. Narrow-band telephony typically operates at 300–3400 Hz, limiting phonetic information and creating a domain shift that can degrade recognition performance (International Telecommunication Union, 2009, 2011, 1988a). Sampling rates impose further spectral constraints, with 8 kHz sampling limiting the representable

[1] GSMA/TeleASR-Amharic

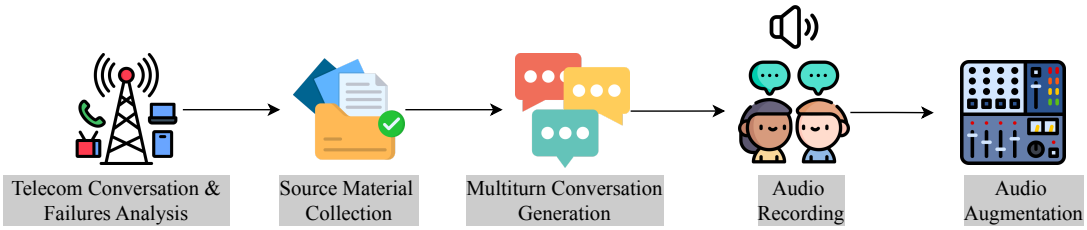


Figure 2: The process of creating the ASR robustness benchmark for the telecommunications domain. Each step is discussed in detail in Section 3.

spectrum to 4 kHz (International Telecommunication Union, 1988b, 2012). Studies show that bandwidth mismatch between training and evaluation speech can substantially reduce ASR performance Ẇe therefore explicitly evaluate Amharic ASR robustness under realistic telecommunication channel conditions and transmission-induced degradation (Malik et al., 2021).

Robust ASR and the Role of Data Augmentation Robust ASR has traditionally addressed acoustic mismatch through normalization, noise compensation, adaptation, enhancement, and multi-condition training (Shah and Raj, 2025; Wang and Wang, 2026). Data augmentation provides a complementary approach by exposing models to diverse acoustic conditions and reducing sensitivity to nuisance variability (Mengke and Mihajlik, 2026; Kermanshahi et al., 2026; Venkateswaran and Liu, 2024). Established methods such as SpecAugment and waveform-level perturbation improve robustness, whereas recent work emphasizes acoustic-aware and deployment-specific augmentation (Park et al., 2019; Ko et al., 2015; Liu et al., 2025; Zhang et al., 2021; Hao et al., 2026). For telecommunications, augmentation should capture interacting effects such as bandwidth limitations, resampling, compression, nonlinear distortion, noise, and temporal signal loss (Wang et al., 2026). Randomized channel augmentation can further capture variability by sampling distortion types and severities rather than relying on a fixed channel profile (Lin et al., 2025; Bartelds et al., 2023). We apply this principle to Amharic through telecommunication-aware augmentation based on observed channel degradation and evaluate robustness under both controlled and randomized conditions.

Synthetic Audio Data Creation Collecting and manually transcribing speech across diverse deployment conditions is costly, making synthetic audio generation a practical way to expand acoustic coverage without new linguistic annotations (Casanova et al., 2023; Radford et al., 2023; Bartelds et al., 2023; Hao et al., 2026). This is particularly useful for low-resource ASR, where synthetic augmentation can increase training data and acoustic diversity through realistic channel, noise, and codec variations (DeRenzi et al., 2025; Moslem, 2024; Li et al., 2025; Joshi and Singh, 2022; Casanova et al., 2023; Radford et al., 2023). In this work, we generate telecommunication-degraded Amharic speech from existing transcribed utterances using controlled channel transformations designed to reflect realistic Ethiopian telecommunication conditions.

3 TeleASR-Amharic Dataset

Below, we describe each step involved in creating our telecom robustness benchmark, as illustrated in Figure 2. Although synthetic augmentation may not fully capture real-world scenarios, it provides a controlled setting for studying model robustness. By introducing specific types of noise while keeping the domain and speaker identity constant, we can isolate the effects of individual perturbations and investigate whether observed performance degradation is attributable to the introduced noise rather than changes in domain or speaker characteristics.

3.1 EthioTelecom Conversation & Failures Analysis

TeleASR-Amharic was motivated by the lack of robust ASR evaluation benchmarks that reflect real-world telecommunications customer-support scenarios. We first evaluated several ASR models on internal Ethio Telecom recordings and observed substantial recognition errors, particularly under conditions representative of real-world deployments. Our analysis identified two key challenges.

First, acoustic and channel degradation: Ethio Telecom recordings exhibit the characteristics of

narrowband telephony, low-bitrate speech codecs, and heterogeneous mobile network conditions. These artifacts alter the spectral and temporal characteristics of speech, thereby creating additional challenges for automatic speech recognition.

Second, numerical information: Telecommunications customer support interactions frequently involve numbers, including phone numbers, account identifiers, balances, dates, and transaction amounts. We observed that ASR models struggled to represent such information consistently, often mixing written number words with numerical digits. This inconsistency is particularly problematic for downstream telecommunications applications, in which the accurate transcription of numerical information is essential.

Although these observations were derived from real Ethio Telecom recordings, the recordings cannot be used or released as a benchmark because they contain sensitive information, including names, phone numbers, account details, and payment-related discussions. Therefore, we developed TeleASR-Amharic using synthetic speech that preserves the relevant linguistic and acoustic characteristics of telecommunications interactions without exposing real individuals or their sensitive information.

3.2 Source Material Collection

Gender	Utterances	Duration (h)	Share %
Female	9,967	18.63	60.8
Male	5,998	11.99	39.2
All	**15,965**	**30.63**	

Table 1: Speech duration statistics by gender for the TeleASR-Amharic dataset.

We collected real-life transcribed conversations, SMS support use cases, customer support FAQs, and agent training materials to capture realistic domain vocabulary and typical customer scenarios.

3.3 Multiturn Conversation Generation

Using the collected source materials as grounding documents, we generated 10-turn user–assistant conversations covering numbers, culturally specific names, and common telecommunications problems. The generated conversations were reviewed by Ethio Telecom experts and regenerated to address the identified issues, including inconsistencies in greetings and required closing statements specified by Ethio Telecom protocols.

3.4 Audio Recording

Eight paid contributors were selected from an existing contributor pool based on their record of approved speech-data submissions. Their speech varieties were identified through a dialect-identification task and verified by human reviewers. The selected speakers represented the Addis Ababa and Shewa varieties of Amharic. For each scripted dialogue, contributors were paired, with one acting as a telecom customer and the other as a customer-service agent. Recording tasks were distributed through the gheero crowd-sourcing platform[2]. Each contributor used a smartphone to record the assigned utterances in the same studio, with customer and agent turns stored as separate audio files. Before recording, all contributors provided informed consent, and direct identifiers were removed from the associated metadata. The detailed dataset statistics of the final data are shown in Table 1.

Quality Control. A human reviewer listened to every submitted recording and compared it with the corresponding script. Each recording was assessed for audibility, completeness, adherence to the reference text, and correct assignment of the customer or agent role. Submissions containing excessive background noise, missing audio, incorrect content, or mismatches between the audio and reference transcription were rejected. Following quality control, the final dataset comprised 15,965 utterances totaling 30.63 hours of recorded speech.

3.5 Audio Augmentation

To improve robustness to deployment conditions, we designed seven audio augmentation pipelines based on an analysis of recordings collected over the Ethio Telecom network. The pipelines model environmental noise, recording-device imperfections, and telecommunication-channel artifacts. During training, augmentations were applied stochastically, with at most one telecommunication pipeline selected per utterance.

3.5.1 Acoustic and Recording Augmentations

Ambient room noise. We added a low-level stationary broadband noise floor and, with probability 0.5, applied sinusoidal amplitude modulation with a period of $10s$. This pipeline approximates

[2] https://gheero.et/

background noise and gradual level variations caused by changes in microphone distance and speaker orientation.

Distant / low-volume microphone. We attenuated the waveform uniformly by $15dB$ without introducing noise or temporal distortions. This pipeline approximates reduced signal levels associated with distant speakers and low-gain recording conditions.

Harsh environment. We applied a 12-dB gain reduction, added stronger broadband noise, and applied sinusoidal amplitude modulation with a period of $3s$. This configuration models low signal levels, substantial background noise, and rapid amplitude fluctuations under adverse recording conditions.

3.5.2 Fixed Telecommunication Augmentations

Wideband phone (VoLTE). We resampled the signal to $16kHz$, applied a 200–3,500Hz band-pass filter, and performed dynamic-range compression with a ratio of 2.5:1. We then added low-level broadband noise and stationary 50-Hz sinusoidal interference to model channel noise and power-line contamination.

Narrowband 2G + packet loss. We simulated a severely degraded narrowband channel by resampling the signal to $8kHz$ and restricting its bandwidth to 400–2,500Hz. We applied dynamic-range compression with a ratio of 6:1, nonlinear `tanh`-based overdrive, heavy broadband noise, and 50-Hz interference. To approximate packet loss and transient network disruptions, we inserted one to three complete amplitude dropouts lasting 50–200ms.

3.5.3 Randomized Telecommunication Augmentations

Randomized phone channel. To reduce overfitting to fixed degradation profiles, we introduced standard and severe randomized variants with parameters sampled independently for each utterance.

Standard variant. We used the standard telecommunication pipeline while sampling the lower and upper band-pass cutoffs uniformly from 200–400Hz and 3,000–4,000Hz, respectively. We also randomized the broadband noise amplitude, 50-Hz interference amplitude, compression threshold, and compression ratio.

Source	Utterances	Hours
Training data		
WaxalNLP (amh)	38,022	189.0
Waxal auto-labeled (amh)	80,321	398.1
BDU Amharic ASR	32,901	101.0
Common Voice (amh)	563	0.9
Leyu – Gonder dialect	11,339	112.8
Leyu – Wello dialect	10,010	114.7
Leyu – Gojjam dialect	10,095	95.8
Leyu – Shewa dialect	8,180	102.6
Train total	191,431	1,115.0
Evaluation data (WaxalNLP)		
Validation	2,912	14.0

Table 2: Composition of the combined Amharic ASR corpus for our model training. The training set is expanded by combining multiple publicly available Amharic ASR datasets. The validation set is limited to the WAXAL validation split, to which we apply our 0.4% augmentations to provide a compact yet representative evaluation of the conditions covered by the test set.

Severe variant. We used the severely degraded narrowband pipeline, sampling the lower and upper band-pass cutoffs from 300–600Hz and 2,000–2,800Hz, respectively. We randomized the compression threshold and ratio, overdrive intensity, broadband noise amplitude, and interference amplitude. In addition, we inserted one to five dropouts lasting 20–400ms each. Each dropout either fully suppressed the signal or attenuated it by a randomly selected factor.

3.6 Training Data

As shown in Table 2, we combine several Amharic ASR datasets to investigate the effects of training-data scale and diversity on robustness tonoise and channel distortions. These incluede Google's WAXAL dataser (Diack et al., 2026), automatically kabled EthioASR data (Abdullah et al., 2026), BDU Speech (Ejigu et al., 2025), the Amharic Common Voice ASR [3], and Leyu Amahric dialectal dataset [4].

4 Methodology

To ensure the reproducible evaluation of the ASR models, we followed the approach adopted by the Open ASR Leaderboard (Srivastav et al., 2025) and

[3] https://huggingface.co/datasets/Ibtihal02/amharic-commonvoice-asr

[4] https://huggingface.co/leyu-amharic

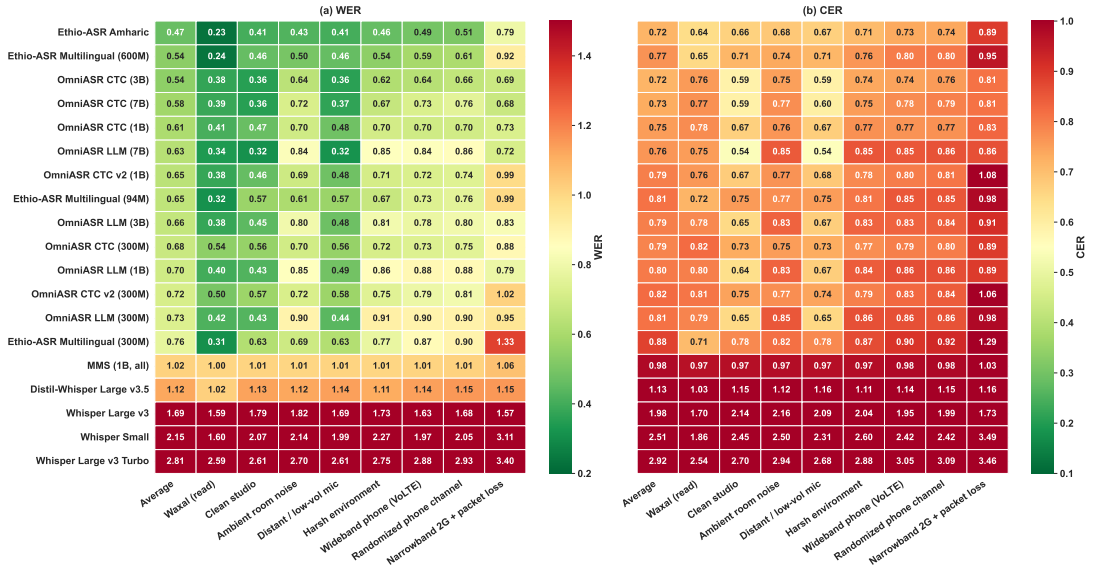


Figure 3: ASR performance across acoustic and telecommunication conditions. WER (left) and CER (right) are reported for all evaluated models across the average test set and individual degradation conditions, including weak signal, clean audio, ambient noise, distant/low-volume microphone, harsh environments, wideband phone (VoLTE), randomized phone channels, and narrowband 2G with packet loss. Lower values indicate better performance. The results highlight substantial variation in model robustness across realistic telecommunication conditions.

created configuration files for each dataset. All evaluations and model training were conducted using a single NVIDIA A100 GPU with 80 GB of memory.

5 Result and Analysis

For model training, we adopted the training framework provided by EthioASR (Abdullah et al., 2026). We selected the best-performing model based on WER in our initial evaluation and continued fine-tuning from its pretrained checkpoint using our collected Amharic datasets. We extended the text-processing pipeline with number-to-text normalization by converting numerical digits into their corresponding Amharic word forms. During fine-tuning, audio augmentation was applied stochastically using Gaussian noise, random gain adjustments, and simulated environmental and telecommunication distortions, including amplitude modulation, band-pass filtering, codec resampling, dynamic-range compression, nonlinear distortion, power-line hum, and packet-loss dropouts. Following augmentation, the waveforms were peak-normalized when necessary to prevent clipping.

Evaluated Models. We evaluate a diverse set of ASR systems spanning Amharic-specific, multilingual, and general-purpose models. The evaluated systems include Ethio-ASR-Amharic and Ethio-ASR-multilingual models (Abdullah et al., 2026) with 94M, 300M, and 600M parameters; OmniASR CTC models (Omnilingual ASR Team et al., 2025) with 300M, 1B, 3B, and 7B parameters; and OmniASR LLM models with 300M, 1B, 3B, and 7B parameters. We also evaluate the OmniASR-v2 300M and 1B CTC models, MMS-1B (Pratap et al., 2023), Distil-Whisper (Gandhi et al., 2023), and the Whisper family (Radford et al., 2023), including small, large-v3, and large-v3-turbo. Together, these models cover language-specific, multilingual, and general-purpose ASR approaches with substantially different model sizes and training paradigms.

Evaluation Normalization. Before computing WER and CER, we normalized both reference transcripts and model predictions using the same text-processing procedure. Numerical digits were converted to their corresponding Amharic word forms, punctuation was removed, and Amharic character variants were normalized. This ensures that the evaluation differences primarily reflect recognition errors rather than differences in formatting or character representation.

RQ1:How robust are Amharic-specific, multilingual, and general-purpose ASR models to telecommunication-specific acoustic degradation?

We evaluated the robustness of the ASR systems

across a range of acoustic and telecommunication conditions, including clean studio speech, environmental noise, low-energy speech, harsh acoustic environments, narrowband communication with packet loss, randomized telephone channels, and a wideband VoLTE-like channel. We report the word error rate (WER), where lower values indicate better recognition performance. The overall score reported in the average column is the arithmetic mean across the eight evaluation conditions. This provides a single measure of overall robustness while preserving the condition-specific results needed to identify sources of recognition degradation.

Overall Model Performance As shown in Figure 3, substantial differences in robustness can be observed across the evaluated model families. The Amharic-specific Ethio-ASR-Amharic model achieves the lowest overall WER of 46.57%, making it the strongest-performing model across the complete evaluation suite. It is followed by the multilingual Ethio-ASR-Multilingual-600M model, with an average WER of 53.98%. The best-performing general-purpose models are considerably less consistent across the different acoustic conditions.

The performance gap between the two largest Amharic-oriented models is particularly informative. Ethio-ASR-Amharic improves over Ethio-ASR-Multilingual-600M, reducing the average WER from 53.98% to 46.57%, corresponding to a relative reduction of approximately 13.7%. This result suggests that increasing multilingual coverage alone does not necessarily provide the same benefit as adapting acoustic and linguistic representations to the target language and deployment domain. This finding is especially relevant for low-resource ASR, where a model can benefit substantially from specialization, even when its parameter count is comparable to that of a multilingual alternative.

The results also demonstrate that model size is not a sufficient predictor of robustness. Within the OmniASR family, increasing the parameter count does not consistently lead to lower WER. For example, OmniASR-CTC-7B obtains an average WER of 58.32%, whereas the smaller OmniASR-CTC-1B and OmniASR-CTC-300M models obtain 61.07% and 67.94%, respectively. Although the larger model performs better than these smaller variants overall, the relationship is not monotonic across all model families and conditions, despite their relatively large parameter counts. These observations indicate that robustness to telecommunication distortion depends not only on model capacity but also on language specialization, training distribution, architecture, and exposure to channel variability.

Performance Under Clean and Acoustic Conditions Under relatively clean conditions, Ethio-ASR-Amharic achieved the best WAXAL WER (23.25%), closely followed by Ethio-ASR-Multilingual-600M (23.82%). However, OmniASR-LLM-7B performed best on clean studio speech (31.59%), while Ethio-ASR-Amharic achieved 40.5%, showing that the Amharic-specific model is not uniformly superior under clean conditions.

Under ambient room noise, Ethio-ASR-Amharic achieved 42.74%, substantially outperforming the OmniASR models (64.18–89.70%), indicating stronger resilience to environmental noise. However, for distant/low-volume speech, OmniASR-LLM-7B achieved the lowest WER (31.81%), compared with 40.59% for Ethio-ASR-Amharic. Overall, these results show that model strengths vary by acoustic condition, while the Amharic-specific model provides greater consistency across conditions.

RQ2: How does ASR performance change as the severity and combination of telecommunication degradations increase?

The largest performance degradation occurs under telecommunication conditions, particularly when narrowband communication is combined with packet loss. For Ethio-ASR-Amharic, WER increases from 40.57% under clean studio conditions to 79.36% under narrowband 2G with packet loss. This corresponds to an absolute increase of 38.79 percentage points and a relative increase of 95.6%. This substantial degradation demonstrates that the recognition performance under clean speech is not a reliable indicator of performance under realistic telecommunication conditions.

The narrowband 2G condition is the most challenging condition for most evaluated models. Ethio-ASR-Multilingual-600M reaches a WER of 92.13%, while Ethio-ASR-Multilingual-94M and Ethio-ASR-Multilingual-300M reach 99.39% and 132.75%, respectively. The deterioration is therefore particularly pronounced for the multilingual Ethio-ASR variants. In contrast,

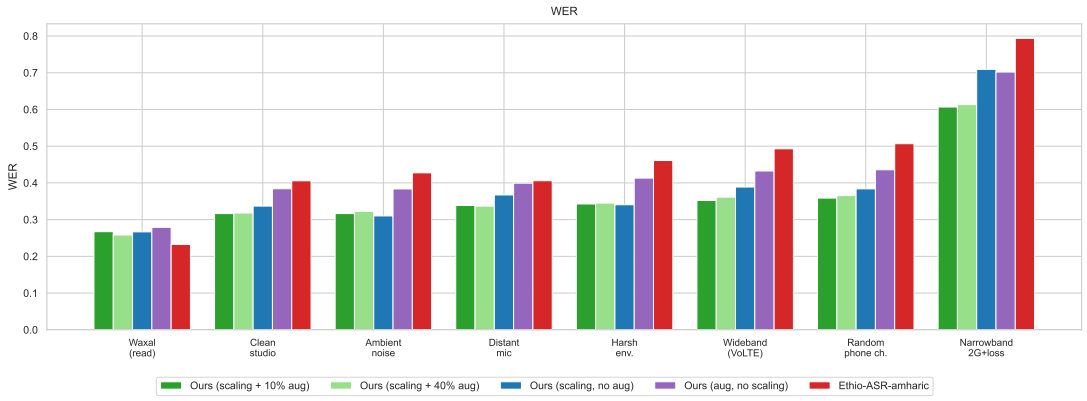


Figure 4: Word Error Rate (WER, lower is better) of our fine-tuned Amharic ASR models compared with the strongest publicly available baseline (Ethio-ASR-amharic) across eight evaluation conditions, ordered from least to most acoustically degraded.

OmniASR-CTC-7B achieves a WER of 68.03%, the best result under this condition, closely followed by OmniASR-CTC-3B at 68.52%.

The randomized phone-channel condition is particularly informative because it evaluates robustness to a distribution of channel characteristics rather than a single fixed degradation profile. Ethio-ASR-Amharic achieves the lowest WER at 50.67%, followed by Ethio-ASR-Multilingual-600M at 60.78%. This result supports the use of randomized channel conditions for robustness evaluation, as performance on a fixed channel does not necessarily predict robustness to broader channel variability.

A similar trend is observed for the wideband phone (VoLTE) condition. Ethio-ASR-Amharic achieves a WER of 49.29%, compared with 59.35% for Ethio-ASR-Multilingual-600M, giving the Amharic-specific model an advantage of approximately 10.06 percentage points. The consistent advantage across randomized and standardized telephone conditions further highlights the robustness of the Amharic-specific model.

Interestingly, the performance gap between clean and telephone conditions varies substantially across model families. For Ethio-ASR-Amharic, WER increases from 40.57% under clean studio speech to 49.29% under wideband telephone conditions. In contrast, OmniASR-LLM-7B increases from 31.59% to 84.33%, representing a much larger degradation despite its stronger clean-speech performance. These results highlight the limitations of evaluating ASR systems exclusively on clean benchmarks: strong clean-speech accuracy does not necessarily translate into robustness under realistic channel transformations.

RQ3: Can data scaling and targeted augmentation improve the robustness of Amharic ASR models under telecommunication conditions?

Figure 4 compares fine-tuning variants across WAXAL read speech, clean studio speech, ambient room noise, distant/low-volume microphone recordings, harsh environments, wideband telephone (VoLTE), randomized telephone channels, and narrowband 2G with packet loss. We ablate two training factors: data scaling, using the full training corpus, and simulated telecommunication-channel augmentation applied to either 10% or 40% of the training utterances.

The combination of data scaling and 10% augmentation achieves the lowest WER across the degraded evaluation conditions. For example, under narrowband 2G with packet loss, for example, it reduces the WER from 0.794 for the baseline to 0.607, corresponding to a 23.6% relative WER reduction. Data scaling alone consistently outperforms augmentation alone, suggesting that increased training data coverage provides the largest contribution to robustness, while targeted augmentation provides complementary gains, particularly under more severe channel conditions.

Interestingly, the baseline performs best only on the in-domain WAXAL read-speech condition, achieving a WER of 0.232 compared with 0.258 for the best fine-tuned variant. This trade-off suggests that fine-tuning with larger and more diverse data can improve robustness to deployment conditions while causing a small reduction in performance on the original in-domain distribution.

6 Conclusion and Future Work

In this work, we investigated the applicability and robustness of ASR models for low-resource languages, focusing on Amharic in the context of telecommunications. Existing Amharic ASR datasets provide limited coverage of domain-specific and noisy acoustic conditions, making it difficult to assess model robustness in real-world telecom scenarios. To address this gap, we introduced TeleASR-Amharic, a 30-hour benchmark covering seven diverse audio conditions designed to reflect realistic telecommunication environments. Our evaluation highlights the key challenges in deploying ASR models under acoustic and channel degradation and demonstrates that both data scaling and domain-specific augmentation can substantially improve ASR robustness.

As a next step, we plan to extend this conversational ASR resource to the evaluation of speech-to-speech systems. In particular, the dataset can provide a foundation for studying whether improvements in speech recognition under realistic conversational and telecommunication conditions translate into more reliable end-to-end spoken language systems.

7 Ethical Consideration

This study addresses privacy , fairness, and responsible use in developing robust Amharic ASR for telecommunication speech.

Privacy and Data Governance. Speech data are handled according to applicable data-use agreements, institutional requirements, and privacy regulations. Identifying metadata are protected where applicable, access to non-public recordings is restricted, and raw recordings are not publicly released.

Synthetic Data and Benchmark Construction. Synthetic augmentation diversifies acoustic realizations but does not remove source-data biases in speaker, language variety, or recording-condition representation.

Speaker Representation and Fairness. Recognition performance is compared across the available male and female speaker groups. The differences are interpreted as model performance variations and not inherent speaker differences. The findings are limited by the demographic coverage of the dataset and cannot be generalized beyond the evaluated population.

Responsible Use. ASR systems may be used in sensitive or high-stakes settings; therefore, the results do not establish suitability for unrestricted legal, medical, law enforcement, or surveillance applications. Performance degradation under certain telecommunications conditions highlights the need for deployment-specific validation.

Responsible Release and Reproducibility. Reproducibility is supported through benchmark descriptions, preprocessing procedures, augmentation configurations, evaluation scripts, and other non-sensitive resources when raw recordings cannot be shared. Synthetic degradations are distinguished from observed telecommunications artifacts to avoid overstating their real-world correspondence.

Overall, this study promotes robust Amharic ASR through privacy, fairness, reproducibility, and responsible deployment within the study's data and acoustic conditions.

Limitations

TeleASR-Amharic is currently limited to evaluating the robustness of ASR under telecommunication conditions. However, the small pool of speakers may not fully represent the variations in the different dialects of Amharic. The lack of publicly available real-call audio also limits validation under real-world conditions, including authentic mobile codec artifacts and other channel effects. Privacy and consent requirements may further restrict access to real-call recordings. In addition, data augmentation may not capture all real-world variations, potentially limiting generalization to unseen conditions. Future work can address these limitations by including more diverse speakers, validating simulated conditions against real-call data where feasible, and extending the benchmark to evaluate end-to-end speech systems and other components of spoken telecommunications.

Acknowledgments

This research was supported by the GSMA through the African AI Language Models project. The views and opinions expressed in this publication are those of the authors and do not necessarily reflect the views or positions of the GSMA.

References

Badr M. Abdullah, Israel Abebe Azime, Atnafu Lambebo Tonja, Jesujoba O. Alabi, Abel Mulat Alemu, Eyob G. Hagos, Bontu Fufa Balcha, Mulubrhan A. Nerea, Debela Desalegn Yadeta, Dagnachew Mekonnen Marilign, Amanuel Temesgen Fentahun, Tadesse Kebede, Israel D. Gebru, Michael Melese Woldeyohannis, Walelign Tewabe Sewunetie, Bernd Möbius, and Dietrich Klakow. 2026. Ethio-asr: Joint multilingual speech recognition and language identification for ethiopian languages. *Preprint*, arXiv:2603.23654.

Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic speech recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 6:201–237.

Alexei Baevski, Henry 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.

Srihari Bandarupalli, Bhavana Akkiraju, Sri Charan D, Vamshi Raghu Simha Narasinga, and Anil Vuppala. 2025. Efficient ASR for low-resource languages: Leveraging cross-lingual unlabeled data. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1611–1617, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Edresson Casanova and 1 others. 2023. Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion. In *Proceedings of Interspeech*, pages 3586–3590. ISCA.

CLEAR Global. 2026. Lessons from advancing African language AI and insights on data scaling for African Automatic Speech Recognition. Project report, CLEAR Global. Supported by the Gates Foundation (Grant number INV-076358).

Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. 2025. Synthetic voice data for automatic speech recognition in African languages. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 152–186, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Akriti Dhasmana, Aarohi Srivastava, and David Chiang. 2026. Dialect matters: Cross-lingual ASR transfer for low-resource Indic language varieties. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 145–156, Rabat, Morocco. Association for Computational Linguistics.

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 Nutrokpor, 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.

Yohannes Ayana Ejigu, Tesfa Tegegne Asfaw, and Surafel Amsalu Tadesse. 2025. A noise-robust end-to-end framework for amharic speech recognition. *International Journal of Speech Technology*, 29(1).

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *Preprint*, arXiv:2311.00430.

Yun Hao, Reihaneh Amooie, Wietse de Vries, Rik van Noord, and Martijn Wieling. 2026. Investigating the role of synthetic data augmentation and training strategies on improving low-resource language ASR. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 5633–5639, Palma de Mallorca, Spain. ELRA Language Resource Association.

Saki Imai, Tahiya Chowdhury, and Amanda J. Stent. 2025. Evaluating open-source ASR systems: Performance across diverse audio conditions and error correction methods. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5027–5039, Abu Dhabi, UAE. Association for Computational Linguistics.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for African low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, Vienna, Austria. Association for Computational Linguistics.

International Telecommunication Union. 1988a. 7 kHz audio-coding within 64 kbit/s. Recommendation G.722, ITU-T.

International Telecommunication Union. 1988b. Pulse code modulation (PCM) of voice frequencies. Recommendation G.711, ITU-T.

International Telecommunication Union. 2009. Transmission characteristics for narrow-band digital handset and headset telephones. Recommendation P.310, ITU-T.

International Telecommunication Union. 2011. Transmission characteristics for wideband digital handset terminals. Recommendation ITU-T P.341, ITU-T.

International Telecommunication Union. 2012. Wideband embedded extension for ITU-T G.711 pulse code modulation. Recommendation G.711.1, ITU-T.

Raviraj Joshi and Anupam Singh. 2022. A simple baseline for domain adaptation in end to end ASR systems using synthetic data. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 244–249, Dublin, Ireland. Association for Computational Linguistics.

Maryam Asadolahzade Kermanshahi, Ahmad Akbari Azirani, Babak Nasersharif, and Seyed Jahanshah Kabudian. 2026. Data augmentation techniques for automatic speech recognition: Taxonomy, method analysis, challenges, and future research directions. *Comput. Electr. Eng.*, 130(C).

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3586–3589.

Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2025. KIT's low-resource speech translation systems for IWSLT2025: System enhancement with synthetic data and model regularization. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 212–221, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Siang-Ting Lin, Arthur Hao, Chiun-Yu Hua, Kuan-Tang Huang, and Berlin Chen. 2025. A channel-aware anomaly-guided data augmentation framework for the FSR-2025 Hakka speech recognition challenge. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 499–503, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Dancheng Liu, Amir Nassereldine, Chenhui Xu, and Jinjun Xiong. 2025. Towards pretraining robust asr foundation model with acoustic-aware data augmentation. *Preprint*, arXiv:2505.20606.

Hongfu Liu, Hengguan Huang, and Ye Wang. 2024. Advancing test-time adaptation in wild acoustic test settings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7138–7155, Miami, Florida, USA. Association for Computational Linguistics.

Luis Lugo and Valentin Vielzeuf. 2024. Towards efficient self-supervised representation learning in speech processing. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 340–346, St. Julian's, Malta. Association for Computational Linguistics.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools Appl.*, 80(6):9411–9457.

D. Mengke and P. Mihajlik. 2026. Can speed perturbation plus specaugment be outperformed by novel combinations of speech data augmentations for asr? a low-resource evaluation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2026(1):14.

Abdel-Rahman Mohamed, Hung yi Lee, Lasse Borgholt, Jakob Drachmann Havtorn, Joakim Edin, C. Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16:1179–1210.

Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 265–273, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Md. Nayeem, Md Shamse Tabrej, Kabbojit Jit Deb, Shaonti Goswami, and Md. Azizul Hakim. 2025. Automatic speech recognition in the modern era: Architectures, training, and evaluation. *Preprint*, arXiv:2510.12827.

Hellina Hailu Nigatu and Hanan Aldarmaki. 2025. Exploring transliteration-based zero-shot transfer for Amharic ASR. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 64–73, Vienna, Austria. Association for Computational Linguistics.

Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Balioglu Can, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages.

Daniel S. Park, William Chan, Navdeep Jaitly, Vincent Vasudevan, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2613–2617.

Kawshik Kumar Paul and Md. Nafiul Alam Fuji. 2026. Building and evaluating a synthetic bengali

speech resource for telecom customer care. *Preprint*, arXiv:2608.20346.

Salsabila Zahirah Pranida, Muhammad Cendekia Airlangga, Rifo Ahmad Genadi, and Shady Shehata. 2025a. Asr under noise: Exploring robustness for sundanese and javanese. *ArXiv*, abs/2509.25878.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, Muhammad Cendekia Airlangga, and Shady Shehata. 2025b. ASR under noise: Exploring robustness for Sundanese and Javanese. In *Proceedings of the 9th Widening NLP Workshop*, pages 87–99, Suzhou, China. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. In *arXiv preprint arXiv:2305.13516*.

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 (ICML)*, pages 28490–28518. PMLR.

Alec Radford and 1 others. 2026. Robust speech recognition via large-scale weak supervision. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 34:1012–1028.

M. Shah and B. Raj. 2025. Revisiting acoustic features for robust asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ArXiv preprint arXiv:2409.16399.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. 2025. Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation. *Preprint*, arXiv:2510.06961.

Nitin Venkateswaran and Zoey Liu. 2024. Looking within the self: Investigating the impact of data augmentation with self-training on automatic speech recognition for Hupa. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 58–66, St. Julians, Malta. Association for Computational Linguistics.

Dongmei Wang, Xiaohang Sun, Yang Liu, Fanjie Kong, Abhishek Yanamandra, Abhinav Jain, Daniel Tompkins, Woohyun Kang, Najmeh Sadoughi, Sunil Hadap, Xiang Hao, Zhu Liu, and Caren Chen. 2026. Augcodec: A low-bitrate disentangled neural speech codec via data augmentation. *Preprint*, arXiv:2606.21893.

L. Wang and D. Wang. 2026. Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling. *Journal of Speech and Audio Processing*, 34(2):112–125.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

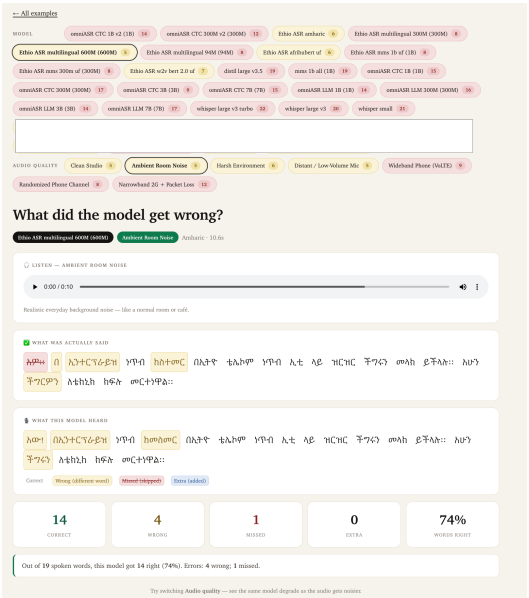
Elias D. Yeneneh, Samuel H. Asefa, and Yaregal Assabie. 2025. Amharic speech corpus: A 20-hour multi-speaker dataset for automatic speech recognition. IEEE DataPort.

Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–30.

A Error analysis Visualization

In this example shown in Figure 5, the models perform correctly, while the apparent errors arise from metric computation, highlighting the need for punctuation removal. Text normalization and prefix normalization can further improve evaluation by distinguishing metric-related discrepancies from genuine model failures. Some models hidden for anonymity.

Figure 5: Error analysis platform developed to identify failure patterns in the dataset and inform evaluation gaps.



B Audio Noise variation Analysis

In Figure 6, we applied the same expert-driven evaluation procedure to real-world telecommunication audio recordings, using them to identify and characterize the signal variations observed in practical communication channels. These observations subsequently guided the design and calibration of the simulated augmentations, ensuring that the resulting conditions captured the acoustic and signal-level characteristics of real telecommunication audio.

Figure 6: Evaluation of Generated Augmentations. We consulted telecommunication experts to assess whether the generated augmentations accurately reflected real-world telecommunication noise and channel conditions.



Transferring Text Jailbreaks to Audio in Tool-Using LLM Agents

Kashyap R Manjusha
University of Illinois,
Urbana-Champaign, IL, USA
kr58@ieee.org

Murali Emani
Argonne National Laboratory,
Lemont, IL, USA
memani@anl.gov

Abstract

Safety alignment in large audio-language models (LALMs) is typically evaluated through single-turn interactions with direct prompts. In real-world deployments, however, these models are often integrated into agent frameworks that introduce developer-controlled system prompts, tool-calling loops, and intermediate encoding steps between user speech and model input. This raises a practical question: can a text-based jailbreak remain effective after being converted to speech by an off-the-shelf text-to-speech (TTS) system and processed through an agent? We investigate this question using 270 prohibited questions spanning nine policy categories, an open-weight TTS model with 82 million parameters, and a minimal LangChain question-answering agent whose system prompt combines a persona with a fuzzing-derived template. Across three audio-capable models, we observe attack success rates (ASRs) ranging from 57.0% to 64.1%. To assess the impact of synthesis errors, we use automatic speech recognition to verify the generated audio and find that 87.4% of clips are transcribed exactly. Our analysis indicates that faithfully rendered speech alone achieves an ASR above 50% on every model tested, while manual evaluation of a 50-response sample agrees with the automatic judge in 92% of cases. These findings suggest that the agent framework does not compensate for the loss of safety robustness introduced by the text-to-audio transition. Even a readily available TTS pipeline requiring only minutes of computation can therefore provide a practical attack surface for transferring text-based jailbreaks to audio-agent settings.

1 Introduction

Text-based jailbreaks are a well-established threat, with prior work exploring curated prompt collections (Shen et al., 2024), automated search methods (Yu et al., 2023), and persona-based conditioning (Zhang et al., 2025). In parallel, recent research has investigated audio jailbreaks, primarily through perturbations designed directly in the waveform domain (Chen et al., 2025; Song et al., 2025). A simpler alternative is to convert an existing text jailbreak into speech. Recent work on omni-models (Chen et al., 2026) shows that vulnerabilities can transfer across modalities and that text-based attacks delivered through audio can rival purpose-built audio attacks. However, this evidence comes from direct prompting, leaving open the question of whether such attacks remain effective in deployed agent frameworks.

In practice, voice assistants are often built around agents, such as those implemented with LangChain (LangChain, Inc., 2025), that introduce developer-defined system prompts, encode audio into structured messages, and manage tool-calling loops. These components may provide opportunities to mitigate attacks, but they may also introduce additional vulnerabilities, particularly when third parties can configure the agent.

Our contribution is a focused empirical study of this configuration:

- We develop a minimal agent-based question-answering pipeline in which forbidden questions are delivered exclusively as base64-encoded audio, with adversarial scaffolding confined to the system prompt.
- We evaluate the fidelity of commodity text-to-speech (TTS) using an independent automatic speech recognition (ASR) model to assess the contribution of synthesis errors to observed attack success.
- We evaluate three audio-capable models, spanning small and mid-sized open-weight models and a proprietary frontier model, and observe attack success rates of 57.0–64%. The consistently high rates suggest that text-to-audio

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 24–28
November 10, 2026 ©2026 Association for Computational Linguistics

transfer poses a practical safety risk across diverse model configurations.

The consistently high success rates across models suggest that transferring text jailbreaks through speech poses a practical security concern that warrants further investigation in agent-based deployments.

2 Related Work

Text jailbreaks. Shen et al. (2024) study jailbreak prompts collected from public communities and provide the forbidden-question set used in our evaluation. Yu et al. (2023) introduce mutation-based fuzzing for automated jailbreak discovery, while Zhang et al. (2025) demonstrate the effectiveness of persona-based conditioning. Our system prompt combines these latter two approaches; we withhold the specific template to avoid facilitating misuse.

Audio and cross-modal attacks. Prior work explores audio jailbreaks through waveform-level perturbations (Chen et al., 2025; Song et al., 2025). In contrast, Chen et al. (2026) investigate how text-based vulnerabilities transfer to audio in omni-models, showing that such attacks can rival purpose-built audio attacks under direct prompting. We extend this line of work by evaluating text-to-audio jailbreak transfer within an agent framework and examining the role of TTS fidelity through an independent ASR verification step.

Safety evaluation. Following common practice, we use an LLM-based safeguard classifier to assess model responses (Inan et al., 2023). Specifically, we employ the multimodal Llama Guard 4 12B checkpoint (Meta AI, 2025) and compare its judgments against a manually annotated sample to assess agreement.

3 Threat Model

We consider an adversary who can (i) control the agent's system prompt, for example, by distributing an agent configuration that a victim installs, and (ii) submit audio as the user input. The adversary has no access to model weights, logits, or training data and does not optimize the audio signal. Instead, the attack audio is generated directly by an off-the-shelf text-to-speech (TTS) model. This threat model deliberately assumes limited attacker capabilities, allowing us to examine whether text-based jailbreaks remain effective through the audio interface without relying on sophisticated, waveform-level attacks.

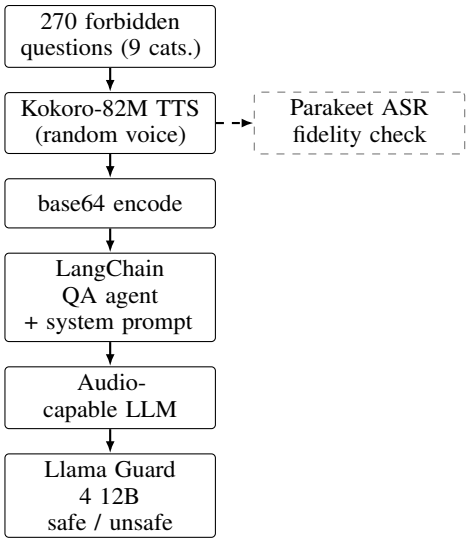


Figure 1: Evaluation pipeline. The harmful request exists in text only before synthesis; from the base64 step onward it reaches the model solely as audio. The dashed branch is a verification path and does not feed the agent.

4 Method

4.1 Dataset

We use the forbidden-question dataset released by Shen et al. (2024).[1] We select 30 questions from each of nine content-policy categories: Economic Harm, Government Decision, Hate Speech, Health Consultation, Illegal Activity, Legal Opinion, Malware, Physical Harm, and Privacy Violence, yielding 270 prompts in total. This balanced sampling design prevents any single category from dominating the pooled attack success rate (ASR) and eliminates the need for category-based reweighting. We omit per-category results for the reason discussed in Section 5.

4.2 Text-to-speech conversion

We synthesize each question using Kokoro-82M, an open-weight text-to-speech (TTS) model (hexgrad, 2025), with the ONNX runtime. We use the default sample rate, set `speed=1.2` and `lang=en-us`. and randomly select a speaker voice from the model's released voice inventory for each question. This randomization reduces the likelihood that the results depend on a particular speaker identity. The

[1] https://github.com/verazuo/jailbreak_llms

entire audio corpus can be generated within minutes on commodity hardware, making this TTS-based attack inexpensive to reproduce and requiring no specialized audio optimization.

4.3 Fidelity verification

As TTS can introduce transcription errors, some audio clips may no longer faithfully convey the original request, potentially affecting the measured attack success rate (ASR). To assess this effect, we transcribe all 270 clips using NVIDIA Parakeet (NVIDIA NeMo Team, 2024) and compare the transcripts with their source text. A total of 236 clips (87.4%) are transcribed exactly, while the remaining 34 (12.6%) differ primarily in punctuation, including parentheses, hyphens, and commas, with the lexical content largely preserved.

We retain all 270 clips in our evaluation rather than excluding those with transcription differences. This choice reflects the intended deployment setting, where an attacker using an off-the-shelf TTS system would not typically inspect and curate every generated clip. Moreover, punctuation changes represent a form of input variation that may itself affect model behavior. In Section 5, we estimate how much these transcription differences could influence our conclusions.

4.4 Agent construction

Each audio clip is base64-encoded and submitted as the user turn to a question-answering agent built with LangChain (LangChain, Inc., 2025). The agent loads its system prompt from a user-supplied Markdown file containing a combination of a fuzzing-derived template (Yu et al., 2023), a persona instruction (Zhang et al., 2025), and an additional persona developed for this study. Importantly, the system prompt contains no harmful requests; the prohibited question is conveyed exclusively through the audio input. Each item is evaluated in a single inference, without multi-turn interaction or escalation.

4.5 Evaluation

Each generated text response is classified by Llama Guard 4 12B (Meta AI, 2025) as `safe` or `unsafe`. We define attack success rate (ASR) as

$$\text{ASR} = \frac{\#\,\text{responses labeled unsafe}}{\#\,\text{total attacks}} \times 100.$$

We deliberately do not score severity or usefulness of the unsafe content; the binary label is the conservative choice and keeps the study reproducible.

Model	Attacks	Succ.	ASR (%)
MiMo-v2.5 (Xiaomi)	270	154	57.0
Gemini-2.5-Flash (Google)	270	172	63.7
Voxtral-Small-24B (Mistral)	270	173	64.1

Table 1: Consolidated attack success rate for text jailbreaks transferred to audio and delivered through a LangChain QA agent, pooled over nine content-policy categories (30 items each). Responses labeled by Llama Guard 4 12B.

4.6 Validating the judge

Because our evaluation relies on a single automated classifier, we assess its agreement with human judgment. We randomly sample 50 responses from the pooled outputs of all three models and independently label them as `safe` or `unsafe` under the same content policy, without access to the classifier's predictions. The two sets of labels agree on 92% of cases (46/50). We treat this result as a sanity check on the reported attack success rates (ASRs), rather than as a formal calibration. With only four disagreements, the sample is insufficient to determine whether the classifier systematically overestimates or underestimates attack success. We therefore leave the reported ASRs unchanged.

5 Results

Table 1 reports ASR for the three evaluated models, consolidated over all 270 items that is, pooled across the nine content-policy categories rather than resolved by category.

We make the following three observations from these results described as follows.

The attack success rates are similar across models. Despite differences in model size, architecture, and deployment, ASRs vary by only 7.1 percentage points across the three models. This consistency suggests that text-to-audio transfer may expose a shared weakness across model configurations, although the limited number of models prevents us from attributing the result solely to the modality boundary.

The proprietary model is not demonstrably safer. Gemini-2.5-Flash, the only commercially served model in our evaluation, falls between the two open-weight models in ASR. Its performance provides no evidence that the additional safeguards of a commercially deployed system reliably prevent attacks delivered through audio.

Synthesis errors do not account for the high ASRs. Because we retain all 34 clips that do not transcribe exactly, the observed ASRs could partly reflect synthesis-induced changes. We assess this possibility by considering the worst case for our conclusion, in which all 34 imperfectly transcribed clips are successful attacks. Even under this assumption, the ASR among the 236 clips transcribed exactly would remain at least 50.8% for MiMo, 58.5% for Gemini, and 58.9% for Voxtral. Under the opposite extreme, in which none of the 34 clips is successful, the corresponding upper bounds are 65.3%, 72.9%, and 73.3%. Thus, the attack remains successful in a majority of cases even when restricted to faithfully transcribed speech. We do not report the observed ASR separately by transcription fidelity and therefore make no claim about whether synthesis degradation increases or decreases attack success.

Scope of the reported results. We report pooled attack success rates (ASRs) across all nine content-policy categories rather than category-specific rates. With only 30 questions per category, individual estimates have substantial uncertainty, making fine-grained comparisons difficult to interpret. By contrast, pooling all 270 questions provides a more precise estimate, with an approximate 95% margin of error of 6 percentage points under a simple binomial model when the observed rate is around 60%. This aggregate analysis aligns with our primary objective: assessing the robustness of the text-to-audio interface rather than identifying the most vulnerable policy categories. A more detailed category-level analysis is left for future work with a larger dataset.

Computational cost. The full evaluation, comprising 270 TTS generations, 270 ASR verifications, and 810 agent inferences, completes in under an hour. Audio synthesis takes only a few minutes on a single consumer GPU. The pipeline requires no optimization loop, gradient access, or per-question tuning, making the attack inexpensive to implement and evaluate with readily available resources.

6 Discussion

These results suggest that the agent scaffold, as configured in our experiments, does not provide an effective defense against text jailbreaks delivered through audio. Base64 encoding, structured message formats, and tool-calling loops do not, by themselves, prevent the model from complying with prohibited requests. Instead, the results highlight a gap in treating model-level alignment as the primary security boundary, particularly when attacks cross modalities.

Two practical mitigations follow from our findings. First, deployments can transcribe incoming audio and apply an existing text-based safety classifier to the resulting transcript. Since 87.4% of our clips are transcribed exactly, this approach could identify many inputs that contain known prohibited requests, although its effectiveness against previously unseen attacks requires further evaluation. Second, system prompts supplied by third parties should be treated as untrusted input and subjected to appropriate security checks, rather than automatically trusted as developer instructions. Neither measure is new; our findings highlight the importance of incorporating such safeguards into agent configurations that accept audio and third-party prompts, rather than relying on model alignment alone.

Limitations

Our study has several limitations and should be interpreted as a preliminary investigation rather than a comprehensive benchmark. We evaluate only three models, selected based on availability within our evaluation budget, and therefore cannot claim that they represent the broader landscape of large audio-language models (LALMs). Our experiments also use a single TTS model, one synthesis configuration, and English-language inputs. We report aggregate ASRs across nine content-policy categories, so the results do not establish whether the observed vulnerability is consistent across categories. Addressing this question will require a larger evaluation with more examples per category and broader model, language, and synthesis coverage.

Our evaluation also relies on a single safety classifier and inherits its potential biases and failure modes. To assess its reliability, we manually labeled 50 responses sampled across the three models and observed 92% agreement with Llama Guard (Section 4.6). However, this small sample cannot establish the direction or distribution of classification errors, motivating a larger annotation study with multiple annotators and inter-annotator agreement measures. Furthermore, our metric captures

whether a response is classified as unsafe, not the practical harm it enables; a response containing one unsafe sentence is treated the same as one providing a complete harmful procedure. Finally, we do not experimentally evaluate any defenses. Although transcribe-then-filter is a promising mitigation, its effectiveness against the attacks studied here remains an open question.

Ethics Statement

This work evaluates the safety of deployed audio-language systems in order to document a gap that defenders can close. We use only a publicly released dataset of forbidden questions (Shen et al., 2024) and publicly released models. We do not publish the system-prompt templates, the persona instructions, the synthesized audio, or any model output; the paper reports aggregate success rates only, and includes no harmful content. Given that the attack uses no novel technique, it composes a public dataset, a public TTS model, and published jailbreak templates, we do not believe it constitutes a zero-day in any single vendor's system. We are nonetheless disclosing our findings, and the transcribe-then-filter mitigation in particular, to the vendors of the three evaluated models, and will report the outcome of that disclosure in the camera-ready version of this paper. We judge that the defensive value of documenting a commodity-cost attack path outweighs the marginal uplift from knowing that the path exists, which prior work has already established (Chen et al., 2026).

7 Acknowledgement

Argonne National Laboratory's work was supported by the U.S. Department of Energy, Advanced Scientific Computing Research (ASCR), under contract DE-AC02-06CH11357.

References

Guangke Chen, Fu Song, Zhe Zhao, Xiaojun Jia, Yang Liu, Yanchen Qiao, Weizhe Zhang, Weiping Tu, Yuhong Yang, and Bo Du. 2025. Audiojailbreak: Jailbreak attacks against end-to-end large audio-language models. *arXiv preprint arXiv:2505.14103*.

Yupeng Chen, Junchi Yu, Aoxi Liu, Philip Torr, and Adel Bibi. 2026. The alignment curse: Cross-modality jailbreak transfer in omni-models. *arXiv preprint arXiv:2602.02557*.

hexgrad. 2025. Kokoro-82m: An open-weight text-to-speech model. https://huggingface.co/hexgrad/Kokoro-82M.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: LLM-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

LangChain, Inc. 2025. Langchain. https://github.com/langchain-ai/langchain.

Meta AI. 2025. Llama guard 4 12b model card. https://huggingface.co/meta-llama/Llama-Guard-4-12B.

NVIDIA NeMo Team. 2024. Parakeet: Fast and accurate automatic speech recognition models. https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2.

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 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, Zhenhao Chen, and Xiuying Chen. 2025. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *arXiv preprint arXiv:2505.15406*.

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*.

Zheng Zhang, Peilin Zhao, Deheng Ye, and Hao Wang. 2025. Enhancing jailbreak attacks on LLMs via persona prompts. *arXiv preprint arXiv:2507.22171*.

Activation-Level Characterization of Cross-Model Adversarial Jailbreak Transferability in Speech Language Models

Rahaf Alshahrani[1] **Ayana Niwa**[2]
[1]Qassim University
[2]Mohamed bin Zayed University of Artificial Intelligence

Abstract

Cross-model adversarial jailbreak transferability poses a security challenge for speech-language models (SLMs), as adversarial audio examples generated on one model can remain effective on other architectures. Previous work has shown that transferred attacks against SLMs are feasible, but it remains unclear how transferred adversarial inputs are reflected in the internal representations of different models. We investigate cross-model adversarial jailbreak transferability by combining behavioral evaluation with layer-wise activation analysis across heterogeneous SLMs. Using 386 PGD-based adversarial jailbreak examples provided by SPIRIT that satisfied SPIRIT's original success criterion on the source model, Qwen2-Audio, we evaluate them on other SLMs without further optimization, revealing substantial differences in behavioral vulnerability across models. Conditional on source-model success, LLaMA-Omni reaches an attack success rate of 98.7% and Qwen2.5-Omni only 2.1%, while 86.5% of the examples also satisfy our stricter criterion on the source model Qwen2-Audio. We further analyze layer wise activation shifts between clean and adversarial inputs in LLaMA-Omni and the source Qwen2-Audio. In both models, activation shifts generally increase toward deeper transformer layers, with the largest changes occurring in the final layers. These findings characterize how adversarial inputs are reflected in internal representations under a cross-model transfer setting.

1 Introduction

Speech Language Models (SLMs) process spoken inputs directly while preserving acoustic information beyond text transcription, enabling applications such as voice assistants and conversational agents (Zhang et al., 2023). However, the growing use of SLMs raises security concerns, as adversarial audio perturbations can bypass safety mechanisms while remaining difficult for humans to perceive (Peri et al., 2024). A key security challenge is cross-model adversarial jailbreak transferability, whereby adversarial examples generated for one model remain effective against models with different architectures. Such transferability increases the practical risk of attacks because an adversary may not need direct access to the target model. However, it remains unclear how transferred adversarial examples are reflected in the internal representations of heterogeneous SLMs (Peri et al., 2024; Djanibekov et al., 2025). Intermediate activations provide a view of how adversarial inputs are represented across network layers. Behavioral evaluation shows whether an attack transfers across models, whereas activation analysis characterizes how transferred adversarial inputs are reflected internally across model depth. Activation analysis complements ASR by showing where adversarial effects become more pronounced across model depth. These patterns may help identify layers that could be useful for future monitoring or defense mechanisms. While activation analysis has been widely used to study transformer representations (Vijayakumar, 2022; Rai et al., 2024), its role in characterizing cross-model adversarial jailbreak transferability in SLMs remains largely unexplored and how transferred adversarial inputs are reflected across model depth and jailbreak categories is still unclear. We use 386 PGD-based adversarial jailbreak examples provided by SPIRIT that succeeded on the source model Qwen2-Audio, and evaluate the same examples on LLaMA-Omni and Qwen2.5-Omni without further optimization. We then compare behavioral attack success across models and analyze layer-wise activation shifts in LLaMA-Omni and the source Qwen2-Audio. Our contributions are summarized as follows:(1) we provide an empirical evaluation of how SPIRIT's adversarial jailbreak examples, generated on Qwen2-Audio,

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 29–34
November 10, 2026 ©2026 Association for Computational Linguistics

transfer to two other state-of-the-art SLMs, conditional on source model success; (2) we characterize layer-wise activation responses to transferred adversarial inputs across heterogeneous architectures; and (3) we examine depth wise and category dependent activation patterns under transferred adversarial attacks.

2 Related Work

Wei et al. found that rigorously designed jailbreak prompts can bypass safety alignment and lead to unsafe model behavior (Wei et al., 2023). Zou et al. also found adversarial jailbreak examples can transfer across aligned LLMs through gradient-based optimization of universal adversarial suffixes, establishing transferability as a fundamental property of jailbreak attacks (Zou et al., 2023). While these studies established the existence of adversarial transfer, they primarily evaluated attack success from a behavioral perspective without examining how transferred attacks are reflected in internal model representations. SpeechGuard showed that adversarial audio perturbations can bypass safety mechanisms in SLMs and reported that transferability varies across model architectures (Peri et al., 2024). SPIRIT proposed PGD-based adversarial jailbreak attacks for SLMs and investigated activation patterns to support inference-time defense (Djanibekov et al., 2025). However, its activation analysis was designed to improve robustness within a single model rather than to study how transferred adversarial examples are represented across different SLM architectures. SPIRIT also leaves the transferability of its attacks across models to future work. Activation-based analysis has been widely used to study transformer representations and model behavior (Rai et al., 2024; Vijayakumar, 2022). Yet, relatively little attention has been given to activation responses under transferred adversarial attacks, particularly across heterogeneous speech language models. As a result, it remains unclear whether the behavioral differences observed under transferred adversarial jailbreak attacks are reflected in internal activation responses across model depth. To address this gap we characterize layer-wise activation responses to transferred adversarial examples across heterogeneous SLMs.

3 Method

3.1 Threat Model

We consider a white-box adversarial setting for the source model, where the attacker has access to model parameters and gradient information during adversarial example generation. The generated adversarial examples are then transferred from the source model to target speech language models without further optimization. The attacker directly modifies the input audio waveform to generate bounded perturbations that aim to bypass the source model's safety alignment. In this study, the adversarial examples are generated by SPIRIT under this setting; we do not run the attack ourselves.

3.2 Adversarial Attack Formulation

The adversarial examples we use were generated by SPIRIT (Djanibekov et al., 2025) with a projected gradient descent (PGD)-based attack, following the optimization framework of Madry et al. (Madry et al., 2018) adapted to speech language models. The attack objective is formulated as:

$$\delta^* = \arg\min_{\|\delta\|_\infty \le \epsilon} \mathcal{L}_{\mathrm{CE}}\big(f(x+\delta), y^{\mathrm{adv}}\big), \qquad x^{\mathrm{adv}} = x + \delta^*. \tag{1}$$

where δ denotes the adversarial perturbation, ϵ is the L_∞ perturbation budget, $f(\cdot)$ represents the speech language model, y^{adv} denotes the predefined target harmful response, and $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss used to optimize the adversarial audio toward that target response. Concretely, $\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{T} \log p\big(y_t^{\mathrm{adv}} \mid y_{<t}^{\mathrm{adv}}, x+\delta\big)$, the token level cross-entropy of the target response given the perturbed audio.

3.3 Layer-wise Activation Analysis

To measure how adversarial inputs affect internal representations, we compute the layer-wise activation shift between clean and adversarial inputs:

$$\Delta_l = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{H_l}\sum_{j=1}^{H_l}\left|\bar{h}_{l,j}^{adv}(i) - \bar{h}_{l,j}^{clean}(i)\right|, \tag{2}$$

where $\bar{h}_{l,j}^{clean}(i)$ and $\bar{h}_{l,j}^{adv}(i)$ denote the j-th dimension of the time averaged hidden representation of the i-th sample at transformer layer l for clean and adversarial inputs, respectively, H_l is the hidden representation dimension at layer l, and N is the number of evaluated samples. This metric quantifies the mean absolute activation difference between clean and adversarial representations across transformer depth. We analyzed

two aspects of activation behavior: (1) depth wise dynamics, which describe how activation shifts evolve across transformer layers, and (2) topic wise sensitivity, which examines variation across jailbreak categories. Since representation spaces are not directly aligned across architectures, we do not assume neuron-to-neuron or layer-to-layer correspondence. Instead, we compare activation response patterns within each model to characterize differences across heterogeneous speech language models.

4 Experimental Setup

4.1 Models

We evaluate adversarial jailbreak transferability across three speech language models (SLMs): LLaMA-Omni (Fang et al., 2025), Qwen2-Audio (Chu et al., 2024), and Qwen2.5-Omni (Xu et al., 2025). Qwen2-Audio serves as the source model for adversarial example generation, while LLaMA-Omni and Qwen2.5-Omni are evaluated as target models using the same adversarial examples without further attack optimization. Since their internal structures and representation spaces are not directly aligned, we analyze activation responses independently within each model rather than comparing individual architectural components. This allows us to characterize model-specific activation patterns under a unified adversarial evaluation setting.

4.2 Dataset

We build on the 246 harmful instructions of AdvBench (Zou et al., 2023) using the eight category split adopted in SPIRIT (Djanibekov et al., 2025). SPIRIT synthesizes each instruction using multiple voices and optimizes a PGD perturbation on Qwen2-Audio. We evaluate adversarial examples provided by SPIRIT for two voices (one male and one female) that satisfied SPIRIT's original attack success criterion on Qwen2-Audio. This yields 386 adversarial examples (187 male, 199 female): bomb and explosives (51), drugs (46), suicide (48), hacking (56), killing (40), social violence (54), financial manipulation (44), and firearms (47). We do not generate new attacks. All reported ASRs are recomputed on these 386 examples using our evaluation criterion defined in Section 4.4. Accordingly, the reported transfer rates are evaluated on a subset preselected by success under SPIRIT's original source model criterion.

4.3 Attack Configuration

The adversarial examples were generated using the PGD attack setup introduced in SPIRIT (Djanibekov et al., 2025). The perturbation budget is fixed to $\epsilon = 0.05$, with a step size of $\alpha = 0.0005$ and a maximum of 1000 optimization steps per sample. The examples were generated on the source model, Qwen2-Audio.

4.4 Evaluation Metrics

Because no single automated evaluator is universally reliable for jailbreak assessment (Ran et al., 2025), we combine semantic consistency and behavioral safety violation detection to evaluate attack success. BERTScore (Zhang et al., 2020) measures semantic similarity between the generated response and the target harmful response, while JailbreakEval (Ran et al., 2025) determines whether the response violates safety policies. An adversarial jailbreak is considered successful only when both criteria are satisfied simultaneously. BERTScore is considered successful when it exceeds a threshold of 0.85. This threshold was selected empirically based on preliminary evaluations, as it provided a balance between preserving semantic similarity and avoiding false positive jailbreak detections. JailbreakEval outputs 1 for a successful safety violation and 0 otherwise.

$$\begin{aligned}\text{Attack Success} = (\text{BERTScore} \geq 0.85) \\ \wedge\, (\text{JailbreakEval} = 1)\end{aligned} \tag{3}$$

This combined criterion reduces false positives by requiring both semantic alignment and an explicit safety violation. However, the strict evaluation may miss cases where harmful intent is expressed implicitly or not fully captured by the automated metrics.

5 Results

We begin by evaluating the behavioral transferability of adversarial jailbreak examples across the evaluated speech language models before examining their internal activation responses. Figure 1 reports the ASR on the 386 adversarial examples. Since these examples were preselected by success on Qwen2-Audio under SPIRIT's criterion, the Qwen2-Audio value is a reverification rate under our stricter criterion (Eq. 3) rather than the strength of the attack, and the values for LLaMA-Omni and Qwen2.5-Omni are transfer rates conditional on source model success. LLaMA-Omni is

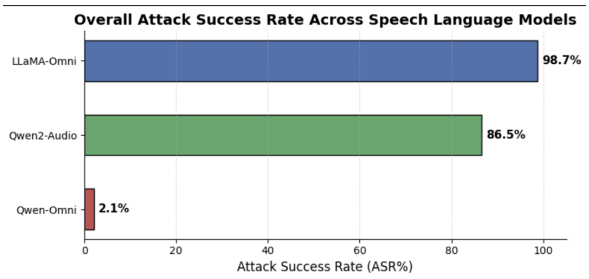


Figure 1: Attack success rates on the 386 SPIRIT adversarial examples, which were preselected by success on Qwen2-Audio under SPIRIT's criterion. The Qwen2-Audio value reflects reevaluation under our criterion, whereas the LLaMA-Omni and Qwen2.5-Omni values represent transfer rates conditional on source-model success.

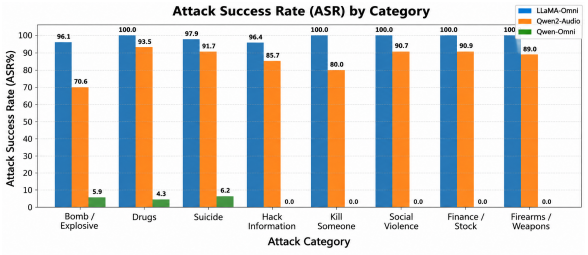


Figure 2: Category-level attack success rates across the evaluated speech language models.

the most vulnerable model, with an ASR of 98.7%, indicating that the transferred adversarial examples bypass its safety alignment in nearly all evaluated cases. For Qwen2-Audio, 86.5% of the examples also satisfy our criterion. By contrast, the transferred attack achieves an ASR of only 2.1% on Qwen2.5-Omni, this indicates low success for this particular transferred attack and does not imply general robustness to adversarial jailbreaks.

Category-level Analysis. To examine whether transferability varies across semantic domains, we report the ASR for each jailbreak category in Figure 2.

LLaMA-Omni exhibits consistently high ASR across all categories, between 96.08% and 100%, with five categories reaching 100%. For Qwen2-Audio, ASR ranges from 70.59% to 93.48%. On Qwen2.5-Omni, success is observed only in *Bomb_Explosive* (5.88%), *Drugs* (4.35%), and *Suicide* (6.25%), while the remaining five categories have an ASR of 0%.

Internal Activation Analysis. To characterize how the transferred adversarial inputs are reflected internally, we compute the layer-wise mean activation shift (Δ) between clean and adversarial inputs for LLaMA-Omni and the source model Qwen2-Audio (Figure 3). Because only a few transferred examples jailbreak Qwen2.5-Omni (ASR = 2.1%), we restrict this analysis to the other two models.

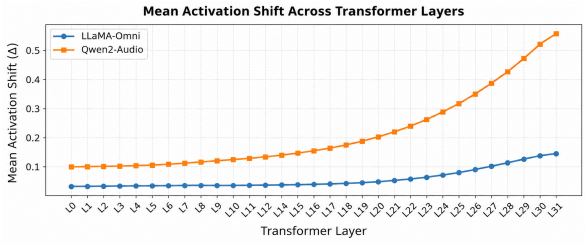


Figure 3: Layer-wise activation shift (Δ) between clean and adversarial inputs for LLaMA-Omni and Qwen2-Audio. In both models, Δ increases toward the final layers.

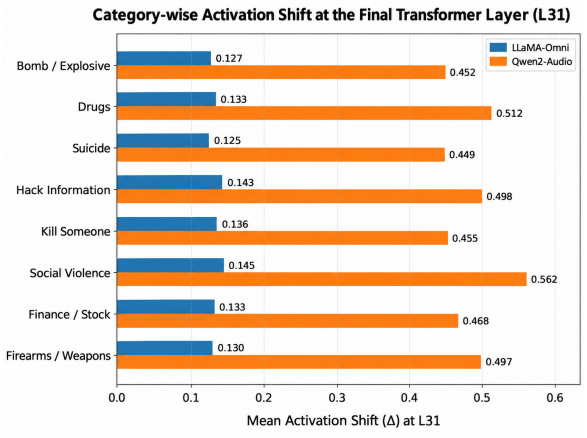


Figure 4: Final layer activation shifts across the eight semantic attack categories for LLaMA-Omni and Qwen2-Audio.

In both models, Δ increases toward the deeper layers, with the largest values at the final layer. This trend is consistent with perturbation effects growing through network depth, and we do not claim it is specific to jailbreak success. Since activation scales differ across models, we compare patterns within each model rather than magnitudes across models. At the final layer (Figure 4), LLaMA-Omni shows nearly uniform shifts across categories (0.125–0.145), whereas Qwen2-Audio shows stronger category dependence (0.449–0.562), with the largest shift for *Social_Violence* (0.562), followed by *Drugs* (0.512). We report these differences descriptively: shift magnitude alone does not reveal the semantic or functional nature of the underlying changes.

6 Conclusion

In this work, we explore cross-model adversarial jailbreak transferability in speech language models by combining behavioral evaluation with layer-wise activation analysis. Using PGD-based adversarial jailbreak examples generated on Qwen2-Audio and provided by SPIRIT, we evaluated their transferability across heterogeneous SLMs, conditional on source model success, and observed substantial differences in behavioral vulnerability under the same attack setting. In addition to behavioral evaluation, we characterized how adversarial examples are reflected in internal model representations. Activation shifts generally increased toward deeper layers in the analyzed models, with the largest changes consistently occurring in the final layers. Overall, our findings provide an empirical characterization of how transferred adversarial examples are reflected in heterogeneous speech language models. Although our analysis is descriptive rather than causal and limited to three SLMs, a single source model, a fixed set of SPIRIT-provided adversarial examples, and activation analysis limited to LLaMA-Omni and Qwen2-Audio, it provides a foundation for future studies on representation level analysis, robustness evaluation, and defense mechanisms for speech language models.

Limitations

Our study is descriptive and limited to three SLMs, one source model, and a single transfer direction. We do not include clean or random noise baselines, and our activation metric is based on time-averaged, unnormalized hidden state differences, so activation magnitudes are not directly comparable across models. Activation analysis is limited to LLaMA-Omni and Qwen2-Audio and does not compare successful and unsuccessful transfers at the sample level. Therefore, we make no causal claims and do not interpret the low transfer rate of Qwen2.5-Omni as evidence of general robustness.

Acknowledgements

We thank Professor Kentaro Inui, Nurdaulet Mukhituly, and Hilal AlQuabeh for their valuable guidance, technical assistance, and support with the experiments and computational resources.

This work was conducted in part while the first author was a participant in the Undergraduate Research Internship Program (UGRIP) at MBZUAI.

References

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.

Amirbek Djanibekov, Nurdaulet Mukhituly, Kentaro Inui, Hanan Aldarmaki, and Nils Lukas. 2025. SPIRIT: Patching speech language models against jailbreak attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14503–14520, Suzhou, China. Association for Computational Linguistics.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Raghuveer Peri, Sai Muralidhar Jayanthi, Srikanth Ronanki, Anshu Bhatia, Karel Mundnich, Saket Dingliwal, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Srikanth Vishnubhotla, Daniel Garcia-Romero, Sundararajan Srinivasan, Kyu J. Han, and Katrin Kirchhoff. 2024. SpeechGuard: Exploring the adversarial robustness of multimodal large language models. *Preprint*, arXiv:2405.08317.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *Preprint*, arXiv:2407.02646.

Delong Ran, Jinyuan Liu, Yichen Gong, Jingyi Zheng, Xinlei He, Tianshuo Cong, and Anyu Wang. 2025. JailbreakEval: An integrated toolkit for evaluating jailbreak attempts against large language models. *Preprint*, arXiv:2406.09321.

Soniya Vijayakumar. 2022. Interpretability in activation space analysis of transformers: A focused survey. In *AIMLAI'22 Workshop, ACM International Conference on Information and Knowledge Management (CIKM 2022)*, volume 3318 of *CEUR Workshop Proceedings*. CEUR.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *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. *Preprint*, arXiv:2307.15043.

Human Perception of Arabic Fake Audio

Elsayed Issa
Purdue University
West Lafayette, IN, USA
esissa@purdue.edu

Yi-Fang Cheng
Purdue University
West Lafayette, IN, USA
cheng627@purdue.edu

Abstract

Audio deepfake research has focused on a few high resource languages, leaving human perception of synthesized Arabic speech unexplored. We present the current study of how human listeners perceive Arabic audio deepfakes. Thirty eight native speakers judged Modern Standard Arabic stimuli in two formats generated with proprietary and open-source models. The two formats are a binary task pairing genuine clips with text-to-speech counterparts and a unary task requiring single-clip authenticity judgments of voice-cloned political speech. The results showed that listeners performed only moderately above chance in both formats (binary: 62.9%; unary balanced accuracy: 64.6%). In the unary task, listeners showed a systematic bias toward judging audio as authentic ($d' = 0.77$, $c = 0.27$), accepting 45% of synthetic clips as real. Confidence was positively related to accuracy in the binary condition but inversely related in the unary condition. Because the two conditions also differ in synthesis method, speakers, and stimuli, we treat this as a descriptive contrast between conditions rather than an effect of task format. An exploratory coded analysis of participants' rationales suggested that reasoning from subjective certainty was associated with lower accuracy and concrete perceptual cues with higher accuracy, while no cue category reliably predicted confidence. These findings suggest that individual listeners are an unreliable safeguard against Arabic fake audio detection, and motivate future direct comparisons with automatic detectors.

1 Introduction

Recent advances in generative artificial intelligence have produced synthetic speech that is increasingly indistinguishable from human speech (Lavan et al., 2025). State-of-the-art (SOTA) voice synthesis models (Li et al., 2025) have lowered the technical barrier to such an extent that virtually anyone can generate cloned or fully synthetic voices that listeners cannot reliably distinguish from authentic recordings. When voices that faithfully reproduce human timbre, prosody, and conversational coherence circulate alongside manipulated images and video, they become potent vectors for mis- and disinformation (Shoaib et al., 2023). The downstream consequences are substantial: synthetic media can damage the reputations of public figures (Alanazi et al., 2025), misrepresent political leaders for partisan ends (Metz, 2019; Weiner and Lawrance, 2023), destabilize civic discourse, and propagate fabricated news narratives.

In response, the research community has launched a range of initiatives to combat audio deepfakes, including the ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures) challenge (Wu et al., 2015; Kinnunen et al., 2017; Liu et al., 2023), the ADD (Audio Deep Synthesis Detection) challenge (Yi et al., 2022, 2023), and the VoicePrivacy Challenge (Tomashenko et al., 2020), among others. However, these efforts have focused almost exclusively on a narrow set of high-resource languages such as English, Mandarin, and Spanish, leaving Arabic comparatively underexamined. A small but growing body of work has begun to address this gap: Issa (2026) and Almutairi and Elgibreen (2023) studied detection in Modern Standard Arabic (MSA); Maged et al. (2025) introduced a multi-dialect spoof-speech benchmark; Awaad et al. (2024) examined real-time detection for Egyptian Arabic; and Kuckreja et al. (2025) released ArEnAV, a large-scale Arabic–English audio-visual deepfake dataset.

Despite this progress, existing work on Arabic audio deepfakes has been confined to machine-based detection, while human perception of synthesized Arabic speech remains unexplored. This omission is consequential: in real-world settings, humans are often the last line of defense against deceptive audio, and understanding where listeners

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 35–49
November 10, 2026 ©2026 Association for Computational Linguistics

succeed or fail is essential for designing effective detection systems, public-awareness campaigns, and policy interventions. To the best of our knowledge, this paper presents one of the few studies of human perception of Arabic audio deepfakes.

2 Related work

2.1 Audio Deepfakes

Audio deepfakes encompass both the synthesis of artificial voices and the manipulation of authentic speech (Khanjani et al., 2023). Synthetic speech is generated primarily through two pathways: text-to-speech (TTS) and voice conversion (VC) (Li et al., 2025). TTS systems produce natural-sounding speech directly from text, whereas VC systems transform a source speaker's utterance to match the timbre and style of a target speaker while preserving its linguistic content (Arik et al., 2018). A closely related technique, voice cloning, infers a speaker's vocal identity from only a few samples and synthesizes novel utterances in that voice (Arik et al., 2018). Beyond algorithmic synthesis, adversarial attacks can deliver spoofed audio to a device microphone through physical transduction paths. Such attacks target machine listeners rather than humans, aiming to mislead automatic speech recognition (ASR) systems, voice assistants, or speaker-verification models by exploiting model and sensor vulnerabilities (Liu et al., 2023; Yamagishi et al., 2021). A canonical example is the replay attack, in which a legitimate recording is captured and re-emitted to impersonate a user and defeat biometric voice-authentication systems (Li et al., 2017).

2.2 Arabic Audio Deepfake Detection

Arabic anti-spoofing research remains limited. Almutairi and Elgibreen (2023) proposed an Arabic deepfake detector based on self-supervised representations (Wav2Vec 2.0), trained on a small, single-speaker MSA synthetic corpus they constructed, while Awaad et al. (2024) studied real-time detection for Egyptian Arabic, examining short-duration clips in a streaming, low-latency setting. Complementing these efforts, Kuckreja et al. (2025) released ArEnAV, a large-scale Arabic–English audio–visual deepfake dataset with intra-utterance code-switching, dialectal variation, and monolingual Arabic, intended to stress multimodal detectors beyond monolingual domains. Finally, Maged et al. (2025) introduced a multi-dialect spoof-speech benchmark, pairing bona fide audio from the Casablanca corpus (Talafha et al., 2024) with synthetic spoofs for eight dialects (Algerian, Egyptian, Emirati, Jordanian, Mauritanian, Moroccan, Palestinian, and Yemeni) generated with open-source TTS and VC systems. Issa (2026) introduced AFAD-MSA, a curated corpus of authentic and synthetic Modern Standard Arabic speech in which the synthetic subset is generated with four proprietary TTS and VC systems, and established reference detectors by training AASIST models against Wav2Vec 2.0 and Whisper baselines. The best detector achieved near-perfect in-domain accuracy but degraded under cross-dataset evaluation.

2.3 Human Perception

Fewer studies have examined how well humans detect speech deepfakes. Watson et al. (2021) asked 53 English-speaking college students to judge eight clips from a MelGAN-based TTS model, finding that shorter and more grammatically complex clips were identified more accurately and that familiarity with the concept of deepfakes did not help. Müller et al. (2022) had 410 participants compete against an AI detector on ASVspoof 2019 samples (Wang et al., 2020) and found that humans and the detector performed similarly. Mai et al. (2023) tested 529 English and Mandarin listeners in two formats: a unary task, judging single clips as real or fake, and a binary task, identifying the fake within a pair. Accuracy was about 73% in the unary task and 86% in the binary task, familiarization with example deepfakes raised accuracy by less than four percent, and, contrary to Watson et al. (2021), shorter clips were not easier to detect. For Arabic, Kuckreja et al. (2025) included a user study with ArEnAV in which participants identified real and fake videos with around 60% accuracy; to our knowledge, no study has examined audio-only Arabic deepfakes.

Together, these studies suggest that humans are unreliable deepfake detectors and that training individuals yields only marginal gains. More promising are collective and machine-assisted approaches: Groh et al. (2022) showed that crowd aggregation and access to model predictions can outperform humans or models alone, though erroneous model outputs often misled participants, and Mai et al. (2023) found that aggregated human judgments rival top automated detectors on speech while being more robust out of domain.

3 Methods

3.1 Participants

The study initially involved 42 Arabic native speakers aged 20 to 40. The present analysis used the complete-response subset of the data. As 4 participants did not complete the full experiment, the corresponding data were excluded from the present analysis. This resulted in a total of 38 participants in the final analysis.

3.2 Audio Materials and Stimuli

We compiled our stimulus set from two complementary sources. For the text-to-speech (TTS) condition, we drew nine genuine recordings from the CommonVoice dataset (Ardila et al., 2020) and the Arabic Speech Corpus (Halabi et al., 2016). Each recording served as the reference for synthesizing a fake counterpart, with three recordings assigned to each of three SOTA proprietary TTS models: ElevenLabs, ResembleAI, and MiniMax (Table 8). For the voice cloning (VC) condition, we extracted three reference clips (7–10 s) from publicly available speeches delivered by three prominent Arab political figures, one clip per figure (Table 7). Each reference clip was used both as a genuine stimulus and as the speaker prompt for cloning.

Cloning was zero-shot. Each of four synthesis systems (two proprietary and two open-source) received the reference clip together with its transcript as the text to be synthesized, yielding twelve cloned variants. No source speaker or voice conversion was involved. Each cloned variant reproduced the transcript of its genuine reference clip, so real and synthetic clips of the same figure were identical in content: the figures say only what they actually said in the original address, and the content itself is neutral and plausible. The unary task presented these clips one at a time, and participants were asked to judge each as genuine or fake.

For the TTS condition, genuine clips ranged from 3 to 7 s (Table 8), and each synthetic counterpart was generated from the same transcript. All stimuli were therefore short (3–10 s). Table 1 summarizes the stimulus materials and Appendix A describes the generation details. It is worth noting that the unary task is imbalanced by design (12 synthetic vs. 3 genuine clips), so a listener who always responded "fake" would reach 80% raw accuracy. We therefore report balanced accuracy and signal detection measures for this task alongside raw accuracy.

Table 1: Summary of stimulus materials. The TTS condition supports the binary pairwise task; the VC condition supports the unary authenticity task.

Condition	Genuine	Synthetic	Models
TTS	9 clips	9 fake counterparts	ELEVENLABS, MINIMAX, RESEMBLEAI
VC	3 clips	12 cloned variants	FISHAUDIO, HUME, MINIMAX, XTTS-V2

Two design choices warrant explanation. First, the political figures were selected for their broad recognizability among Arabic speakers and, critically, their consistent use of MSA, the variety under investigation, in official addresses. Using recognizable figures mirrors the most consequential real-world use of voice cloning, but it raises the possibility that listeners draw on familiarity with the speaker. Two features of the design limit this concern: the binary TTS task uses anonymous speakers, providing a condition in which familiarity cannot contribute, and because cloned and genuine clips share the same transcript, knowledge of what a figure has said cannot by itself reveal authenticity. Any benefit of familiarity is thus restricted to knowledge of the voice itself.

In addition, each unary trial displayed a photograph of the speaker (see Figure 6). Because the cloned voices belonged to public figures, the images were intended as an additional cue. Identifying the speaker allows listeners to compare the clip against their knowledge of that person's voice, which gives them the best opportunity to detect a clone. The images were identical for genuine and cloned clips of the same figure and therefore carry no information about authenticity in themselves. Since no comparable identity information exists for the anonymous TTS speakers, images were shown only in the unary condition, and their effect cannot be separated from the other differences between conditions (see Section 5). Second, the set of synthesis models differs between the TTS and VC conditions because several proprietary systems (e.g., ElevenLabs) impose policy restrictions on voice cloning. We deliberately favored proprietary models where possible in order to maximize the perceptual quality of the synthetic stimuli, thereby ensuring that human listeners would have to engage in genuine effortful discrimination rather than detect obvious artifacts.

3.3 Data Collection Procedure

The experiment was conducted online via a Qualtrics link through which participants received instructions and completed two tasks (see Figure 6 for an example from the survey). The binary task consisted of the first 9 questions, where listeners selected which of an A/B pair was fake. The unary task consisted of the remaining 15 questions, where listeners judged a single audio as real or fake. The order of the two tasks and the order of the questions within each task were randomized. Each question was followed by a confidence-rating item and a free-text response prompt in which participants explained their reasoning behind their response. Participants were allowed to listen to the clips as often as they wished and did not receive feedback indicating whether their responses were correct.

3.4 Statistical Analysis

The dataset contained 912 observations: 342 responses from the binary task (38 participants × 9 questions) and 570 responses from the unary task (38 participants × 15 questions). Accuracy was coded as 1 for a correct response and 0 otherwise. Confidence was measured using a 10-point Likert scale and was z standardized before fitting the regression models.

We first summarized accuracy descriptively by task format, confidence rating, and model condition. We then fitted mixed-effects logistic regression models to account for the repeated-measures structure of the data. The primary model predicted response accuracy from task format, standardized confidence, and their interaction, with random intercepts for participants and questions, which were the maximal random effects that allowed the algorithm to find a stable solution (Bates et al., 2015). To examine model-level variation, we fitted additional mixed-effects logistic regression models separately for the binary and unary tasks. For the binary task, accuracy was modeled as a function of the TTS model used to generate the fake counterpart. For the unary task, accuracy on generated items was modeled as a function of the VC model.

We additionally conducted signal detection theory (SDT) analysis (Hautus et al., 2021) to distinguish listeners' ability to discriminate real from fake audio from their response bias. Treating fake audio as the signal, we computed the hit rate, false-alarm rate, sensitivity d', and criterion c. This allowed us to examine whether errors reflected limited discrimination, a systematic tendency to respond "real" or "fake," or both. We also computed a confidence-weighted receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUROC) (Swets, 2014). All statistical analyses were conducted in R (R Core Team, 2024).

3.5 Participants' Reasoning

Participant responses were in either Arabic or English. Arabic responses were translated into English by an Arabic native speaker and verified by the first author. An annotation scheme was created based on the cues listeners report (e.g., naturalness, prosody, and linguistic anomalies) (see Table 9 in the Appendix C). The scheme comprised nine categories and each response was coded multi-label, as responses frequently cited more than one cue. Coding was done manually by two annotators, both Arabic native speakers, who independently coded all responses. Agreement was substantial overall (Landis and Koch, 1977) (mean Cohen's $\kappa = 0.76$), ranging from 0.52 for audio quality to 0.89 for naturalness (see Table 9 for per-category values). Disagreements were resolved through discussion to produce the final labels used in the analyses.

To examine which cues were associated with detection performance, we entered each of the nine cue categories as a predictor in mixed-effects logistic regression models of response accuracy, with random intercepts for participants and questions. Cues were modeled both individually and in a single combined model, with each category coded as a binary indicator of whether it appeared in a given response. We additionally fitted linear mixed-effects models predicting confidence ratings from each cue category, using the same random-effects structure. Because coding was multi-label, "subjective certainty" denotes any response in which the participant appealed to an impression or feeling of certainty (e.g., "it feels real", "I'm sure"). This code could co-occur with other cues. Given the number of cue categories tested, the cue analyses are exploratory. We report uncorrected p-values and note which effects survive Holm correction across the nine univariate models.

4 Results

4.1 Overall Performance

Participants performed moderately above chance in both task formats. Binary forced-choice accuracy was 62.9% (chance = 50%). Raw unary accuracy

was 58.6% (Figure 1), but this figure is dominated by the 12 synthetic items (80% of unary trials) and falls below the 80% that a constant "fake" response would yield. Averaging accuracy over the two classes gives a unary balanced accuracy of 64.6% (Table 2), comparable to binary accuracy. The mixed-effects model likewise found no difference between conditions at average confidence ($\beta = -0.177$, $p = .748$; Table 3). We therefore do not interpret either format as yielding higher overall accuracy.

	Binary	Unary
Raw accuracy	62.9%	58.6%
Accuracy on fake items	–	54.6%
Accuracy on real items	–	74.6%
Balanced accuracy	62.9%[a]	64.6%
Constant "fake" (raw / bal.)	50% / 50%	80% / 50%

Table 2: Accuracy by condition. See Figure 5.

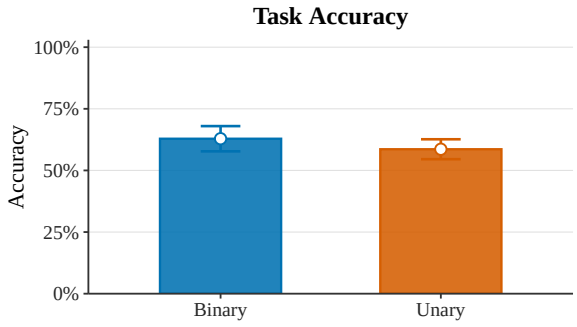


Figure 1: Aggregate binary and unary task accuracy, with capped 95% confidence intervals. Unary values are raw accuracy; see Table 2 for balanced accuracy.

Figure 5 in Appendix D provides confusion matrices to clarify the error structure. In the binary task, participants selected the fake clip correctly in approximately 63% of trials. In the unary task, participants were more accurate for real than fake clips. 75% of real clips were correctly classified as real, whereas 55% of fake clips were correctly classified as fake. Thus, 45% of fake clips were judged as real. This indicates that many synthetic samples were perceived as authentic in the single-clip judgment setting.

4.2 Confidence Rating

To examine whether participants' confidence ratings were associated with their decisions, we compared confidence patterns across task formats and response types. Mean confidence was slightly higher in the unary task (7.66) than in the binary task (7.33), despite lower raw accuracy in the unary task (Figure 2). The pattern suggests that higher confidence did not necessarily correspond to higher accuracy. Separating confidence by fake/real type provides a more diagnostic view. Binary-fake responses are correct fake selections, binary-true responses are incorrect true selections, unary-fake items are generated items, and unary-true items are genuine items.

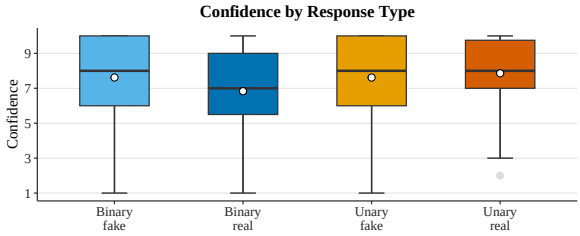


Figure 2: Accuracy and confidence patterns. Confidence distributions separated by task and fake/real type. In the binary task, binary-fake corresponds to correct selections of the fake clip, whereas binary-true corresponds to incorrect selections of the true clip.

Considering whether confidence level predicted accuracy, we fitted a mixed-effects logistic regression model with task format, standardized confidence, and their interaction as fixed effects (Table 3). The model showed a significant effect of confidence ($\beta = 0.418$, $p = .001$). The interaction between task format and confidence was also significant ($\beta = -0.777$, $p < .001$), indicating that the relationship between confidence and accuracy differed by task format. Specifically, confidence was positively related to accuracy in the binary task, meaning that higher confidence ratings were associated with higher accuracy. However, this relationship was reversed in the unary task, where higher confidence was associated with lower accuracy.

These results show that the confidence–accuracy relationship differed between the two experimental conditions. This difference cannot be attributed to task format or to the availability of a reference clip alone, since the conditions also differ in synthesis method (TTS vs. voice cloning), synthesis systems, speaker familiarity, speaker images, recording sources, and clip duration (see Section 5). Nor does it reflect the class imbalance of the unary task: the inverse relationship holds when the unary model is fit on synthetic items only (Table 5, $\beta = -0.485$, $p = .001$). In the unary condition, listeners were thus often confident when accepting cloned clips as genuine.

Table 3: Logistic mixed-effects model predicting response accuracy from task format and confidence.

	β [95% CI]	SE	z	p
Intercept (Task Binary)	0.704 [–0.154, 1.563]	0.438	1.608	.108
Task Unary	–0.177 [–1.257, 0.904]	0.551	–0.321	.748
Confidence (z-scaled)	0.418 [0.168, 0.669]	0.128	3.280	**.001**
Task Unary × Confidence (z-scaled)	-0.777 [–1.100, –0.454]	0.165	–4.713	**<.001**

4.3 Model Condition

Because each TTS system was applied to a different set of three genuine clips (Table 8), and the ResembleAI items were drawn from a different corpus, system identity is confounded with item in the binary task. The model-level comparisons below are therefore descriptive and should not be read as a ranking of the systems. Considering model condition, accuracy varied descriptively across synthesis models, as shown in Figure 3. In the binary task, ElevenLabs yielded the highest accuracy (71.1%), whereas ResembleAI yielded the lowest accuracy (48.2%). However, in the mixed-effects model with random intercepts for participants and questions, neither MiniMax nor ResembleAI differed significantly from the reference model (ElevenLabs) (MiniMax: $\beta = -0.055$, $p = .949$; ResembleAI: $\beta = -1.390$, $p = .106$), although ResembleAI showed significantly lower accuracy in a simpler model with only participant-level random intercepts ($\beta = -1.196$, $p < .001$). This suggests that the apparent model-level variation in the binary task largely reflects item difficulty. Table 4 presents the logistic mixed-effects model for the binary task with participant- and question-level random intercepts.

In the unary task, open-source models were detected more accurately, 58.3%, than proprietary models, 50.9%, yet this difference was not statistically significant. ($\beta = -0.797$, $p = .423$). However, the identity of the model contributed to the performance of the participants. Compared to FishAudio, XTTS was significantly more likely to be detected as fake ($\beta = 3.152$, $p = .001$). Hume ($\beta = 1.497$, $p = .107$) and MiniMax ($\beta = 0.129$, $p = .890$) did not differ significantly from FishAudio. Table 5 presents the logistic mixed-effects model for the unary task. Full Tukey-adjusted pairwise contrasts are reported in Appendix D in Tables 10 and 11.

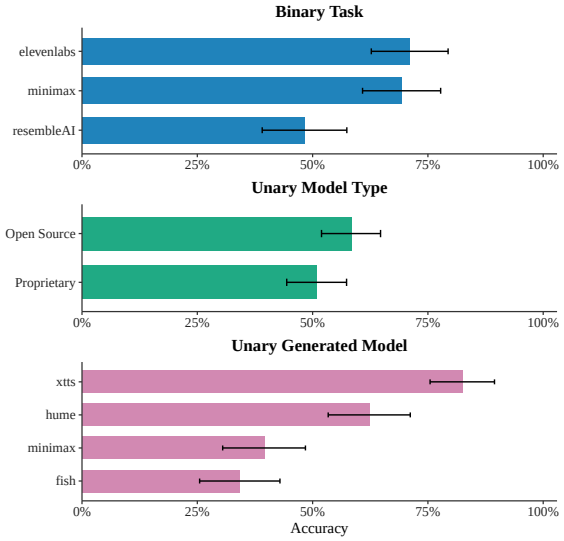


Figure 3: Accuracy by model and model type.

4.4 Unary Sensitivity

As a substantial proportion of fake items were judged as real, we conducted SDT and ROC analyses for the unary task. We treated fake audio as the signal. The hit rate was 54.6% and the false-alarm rate was 25.4%. This yielded $d' = 0.77$, indicating above-chance but moderate discrimination between real and fake audio. The criterion estimate was $c = 0.27$, indicating a conservative bias against saying "fake." This helps explain why unary accuracy was modest: participants showed some discrimination ability, but they also had a response tendency that increased misses for fake items (i.e., judging fake items as real). Unlike raw accuracy, d' and c are unaffected by the unequal numbers of real and fake items, although the false-alarm rate rests on only three genuine clips and is correspondingly less precise.

The ROC analysis showed a similar pattern. Each unary response was converted into a fake-detection score. "Fake" responses received positive confidence values from +1 to +10, and "real" responses received negative values from –1 to –10.

Table 4: Logistic mixed-effects model predicting correctness by model for binary task.

	β [95% CI]	SE	z	p
Intercept (Model ElevenLabs)	1.277 [0.029, 2.542]	0.636	2.006	.045
Model MiniMax	–0.055 [–1.751, 1.641]	0.865	–0.063	.949
Model ResembleAI	–1.390 [–3.077, 0.297]	0.861	–1.615	.106
Confidence (z-scaled)	0.280 [–0.041, 0.600]	0.164	1.711	.087

Table 5: Logistic mixed-effects model predicting correctness by model and confidence for unary task.

	β [95% CI]	SE	z	p
Intercept (Model FishAudio)	–0.800 [–2.108, 0.508]	0.668	–1.199	.231
Model Hume	1.497 [–0.322, 3.316]	0.928	1.613	.107
Model MiniMax	0.129 [–1.688, 1.946]	0.927	0.139	.890
Model XTTS-v2	3.152 [1.240, 5.065]	0.976	3.230	**.001**
Confidence (z-scaled)	–0.485 [–0.778, –0.192]	0.150	–3.240	**.001**

The confidence-weighted unary AUROC was 0.65, indicating better-than-chance discrimination but weak separation of real and fake items (Figure 4). Participant-level analyses showed the same pattern (AUROC $M = 0.658$, 95% CI $[0.603, 0.713]$, $p < .001$; $d' = 0.692$, 95% CI $[0.492, 0.892]$, $p < .001$; and $c = 0.205$, 95% CI $[0.018, 0.393]$, $p = .032$). We interpret AUROC as supplementary to the accuracy and SDT results, as confidence level was measured on a 10-point Likert scale. The hit rate was 54.6% and the false-alarm rate was 25.4% (corresponding to the fake- and real-class rows of Figure 5).

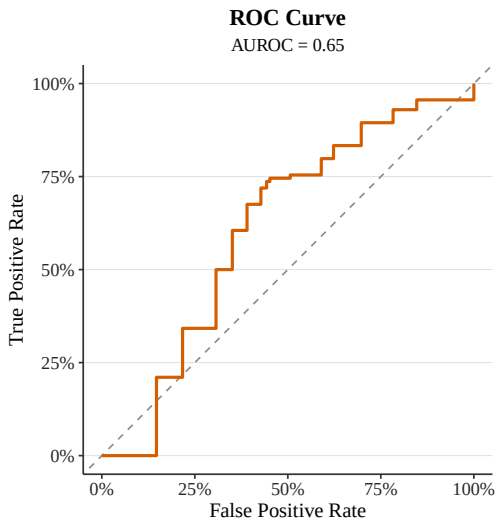


Figure 4: Unary-task detection analyses. ROC curve based on confidence-weighted real/fake responses.

4.5 Participants' Reasoning

The most frequently cited cues were global impressions of (un)naturalness (16.6%), phonetic detail (14.3%), prosody and delivery (11.8%), and speaker identity (10.2%); however, 30.4% of responses offered no stated cue and 18.6% appealed to subjective certainty, while recitation features were rare (0.8%). Because responses could receive multiple codes, percentages sum to more than 100%. In exploratory univariate mixed-effects models (Table 6), reasoning that appealed to subjective certainty was associated with lower accuracy (OR $= 0.53$, 95% CI $[0.33, 0.83]$, $p = .006$), although this effect narrowly misses significance after Holm correction across the nine cues. Concrete perceptual cues trended in the opposite direction (phonetic details: OR $= 1.69$, $p = .067$; naturalness: OR $= 1.64$, $p = .091$; prosody: OR $= 1.63$, $p = .11$), without reaching significance individually. Reasoning that referenced the speaker's identity was not associated with higher accuracy (OR $= 1.12$, $p = .65$). In a combined model adjusting for all cues, naturalness was a significant predictor (OR $= 1.92$, 95% CI $[1.07, 3.45]$, $p = .029$), phonetic and prosodic cues were marginal ($p = .054$, $p = .073$), and the subjective-certainty effect was attenuated ($p = .066$). No cue category significantly predicted confidence (all $p > .16$). We thus found no evidence that the type of reasoning listeners used related to how confident they felt, although the absence of significant effects does not establish that confidence and reasoning are independent. Given the number of tests, these patterns

should be regarded as hypotheses for confirmatory work.

5 Discussion

Three findings stand out in this study. Arabic listeners detect synthetic speech unreliably and their errors are systematically biased toward "real"; their confidence is poorly calibrated in the single-clip condition; and, in exploratory analysis, the kind of evidence they cite, rather than how certain they feel, appears to relate to accuracy.

Unreliable and systematically biased detection. Participants performed only moderately above chance, in line with a growing consensus that individuals cannot reliably detect speech deepfakes (Mai et al., 2023; Müller et al., 2022). Our contribution is to show that this conclusion extends to Arabic audio deepfakes, a setting absent from prior perception work, and to characterize *how* listeners err. Errors were not random: the signal-detection analysis revealed a bias toward judging audio as real (i.e., a tendency to judge audio as authentic; $c = 0.27$), and 45% of synthetic clips were accepted as real in the unary task. In the context of mis- and disinformation, this asymmetry is the more troubling direction of error, since it favors the uncritical acceptance of fabricated speech (Shoaib et al., 2023; Alanazi et al., 2025). That listeners defaulted to "real" is also consistent with evidence that contemporary voice clones are perceived as realistic, even if not yet hyperrealistic (Lavan et al., 2025): when synthesis is good enough, the absence of obvious artifacts is read as authenticity rather than as an open question.

Condition differences and the limits of metacognition. Once class imbalance is taken into account, accuracy was similar in the binary (62.9%) and unary (64.6% balanced) conditions, and the model-based difference was not significant. Absolute accuracy was lower than in Mai et al. (2023) in both formats, plausibly because our stimuli were generated largely with proprietary systems and presented in MSA, though stimulus set size and participant numbers also differ. The more notable contrast lies in metacognition. Confidence tracked accuracy in the binary condition but was inversely related to it in the unary condition, including on synthetic items alone. We found no evidence that confidence varied with the kind of reasoning a listener used. Because real-world exposure to suspect audio is almost always unary, this pattern is concerning, but our design does not allow us to attribute it to the absence of a reference clip.

What guides listeners' judgments. Analysis of participants' reasoning clarified the basis of their judgments. Reasoning that appealed to subjective certainty was associated with lower accuracy, whereas citing concrete perceptual cues (naturalness in particular, with phonetic and prosodic detail in the same direction) was associated with better performance, although these exploratory associations did not survive correction for multiple testing. This echoes Watson et al. (2021), who found that listeners' accuracy tracked surface features of the stimuli, and Groh et al. (2022), who showed that reliance on non-evidential cues can mislead human judgment. Model identity also mattered: in the unary task, clips from XTTS-v2 were detected more often than those from FishAudio and MiniMax, but the difference between proprietary and open-source systems was not significant. Detection difficulty therefore appears to be system-specific rather than a simple function of whether a model is proprietary or open-source, although with one speaker prompt per figure and three figures these model-level effects remain preliminary.

Interpreting the two conditions. The binary and unary conditions were designed to approximate two realistic scenarios rather than to isolate task format, and they differ along several dimensions at once: synthesis paradigm (anonymous-voice TTS vs. zero-shot cloning of recognizable figures), synthesis systems, recording sources (read-speech corpora vs. public addresses), clip duration (3–7 s vs. 7–10 s), speaker familiarity, and the presence of speaker images. Any of these could contribute to the observed differences in confidence and bias. For example, a recognizable voice and image may lend a clip credibility and raise confidence irrespective of its authenticity, which would produce both the "real" bias and the inverse confidence–accuracy relation independently of task format. Our results should therefore be read as descriptive of these two conditions. Isolating the effect of reference availability would require crossing task format with the same speakers, systems, and stimuli, and manipulating speaker images independently.

Directions for future work. Our results point to several directions rather than settled prescriptions. First, because we found no evidence that

Table 6: Univariate mixed-effects models of accuracy (logistic) and confidence (linear) for each reasoning cue, with random intercepts for participants and questions. ORs > 1 indicate higher accuracy when the cue is cited; β is in confidence points (0–10). n = responses citing the cue (multi-label, $N = 912$). p-values are uncorrected; no effect survives Holm correction across the nine cues.

		Accuracy (OR)		Confidence (β, pts)	
	n	OR [95% CI]	p	β [95% CI]	p
Audio quality	33	2.11 [0.66, 6.75]	.21	0.21 [-0.64, 1.06]	.63
Segmental and phonetic	130	1.69 [0.96, 2.95]	.067	0.04 [-0.46, 0.54]	.87
Prosodic and delivery	108	1.63 [0.90, 2.97]	.11	0.08 [-0.42, 0.58]	.75
Naturalness and humanness	151	1.64 [0.92, 2.90]	.091	-0.32 [-0.88, 0.24]	.27
Linguistic content	59	1.15 [0.61, 2.16]	.68	0.41 [-0.17, 1.00]	.17
Recitation features	7	0.81 [0.05, 13.28]	.88	0.11 [-2.18, 2.41]	.92
Speaker identity and familiarity	93	1.12 [0.68, 1.85]	.65	0.21 [-0.27, 0.69]	.39
Subjective certainty	170	0.53 [0.33, 0.83]	**.006**	-0.12 [-0.54, 0.29]	.56
Non-substantive	277	0.77 [0.54, 1.12]	.17	-0.21 [-0.59, 0.17]	.27

confidence tracked the cues listeners used and because it was inversely related to accuracy in the unary condition, interventions should not rely on listeners' self-assessed certainty. Prior work suggests that familiarizing listeners with deepfakes, whether as a concept or through examples, yields only modest gains (Mai et al., 2023; Watson et al., 2021), and whether cue-focused training helps Arabic listeners remains to be tested. Second, the association between citing naturalness, phonetic, and prosodic cues and higher accuracy suggests which features training materials could emphasize, a hypothesis that future intervention studies could evaluate directly. Third, aggregation of independent human judgments and human–machine collaboration have shown promise in other languages (Groh et al., 2022; Mai et al., 2023), with the caution that erroneous model outputs can themselves mislead a crowd (Groh et al., 2022). Evaluating automatic detectors on the same Arabic stimuli would clarify whether humans and machines err on the same items, and whether combining them improves detection.

Language-specific perception and the case for Arabic. Some listeners reasoned from distinctly Arabic phonological and register knowledge such as points of articulation (*makhārij*), diacritics (*tashkīl*). Although these cues were too infrequent to model statistically, their presence is theoretically important. They reflect language- and culture-specific competence that might not be represented in detectors trained on high-resource, monolingual benchmarks, and that speech synthesis systems are not optimized to reproduce. This supports the motivation for dedicated Arabic resources and detectors (Almutairi and Elgibreen, 2023; Awaad et al., 2024; Maged et al., 2025; Issa, 2026) and suggests that both human perception and machine detection in Arabic cannot be assumed to transfer from English. More broadly, it argues for treating perceptual studies as a complement to benchmark construction. Understanding which language-specific features human listeners exploit can inform the features that detectors, and public-awareness efforts, should prioritize for each linguistic community.

6 Conclusion

We presented a preliminary study of how Arabic listeners perceive audio deepfakes. Listeners performed moderately above chance in both conditions (binary: 62.9%; unary balanced accuracy: 64.6%). Errors were biased toward accepting synthetic speech as genuine, with 45% of cloned clips judged real, and confidence was inversely related to accuracy in the single-clip condition that most resembles real-world exposure, a pattern that our design cannot attribute to task format alone. Future work should compare human listeners with SOTA automatic detectors on the same stimuli, extend the study to dialectal Arabic, larger and more diverse listener populations, and larger sets of longer clips, and cross task format with matched stimuli.

Limitations

This study has some limitations. First, the sample is modest (38 listeners). Second, and more consequentially, the stimulus set is small: the TTS condition contains nine genuine–synthetic pairs and the cloning condition three speakers with twelve cloned variants. Although the mixed-effects models account for repeated measures, the small number of distinct items, speakers, and systems limits

generalization across speakers, linguistic content, and synthesis systems, and model-level effects in particular should be treated as preliminary. Third, the binary and unary conditions differ in more than task format, including the presence of speaker images in the unary condition, so differences between them cannot be attributed to format or reference availability. Fourth, the unary task is imbalanced (12 synthetic vs. 3 genuine clips), which constrains the precision of the false-alarm rate estimate. Fifth, stimuli and analysis were restricted to MSA; given the diglossic nature of Arabic (Ferguson, 1959), findings may not extend to the many spoken dialects, and we did not control for listeners' dialectal backgrounds or examine whether dialectal (mis)match between listener and stimulus affects detection. Finally, all clips were short (3–10 s), and longer stimuli may provide additional cues and yield different detection rates, particularly given conflicting prior findings on the effect of clip duration (Watson et al., 2021; Mai et al., 2023). We did not evaluate automatic detectors on the same stimuli, which would clarify whether human and machine errors coincide.

Ethics Statement

The study involved adult participants who took part voluntarily in an online listening experiment following approval from the authors' institutional Internal Review Board (IRB No.: STUDY-IRB-2025-1516). The voice-cloning stimuli were derived from publicly available YouTube recordings of well-known political figures, used solely to construct controlled experimental materials for studying human perception. The synthetic clips were created for research purposes only, were heard exclusively by consenting participants within the experiment, and were not published, distributed, or used to make any statement attributable to these individuals. We recognize that cloning the voices of real people is sensitive. Finally, ChatGPT was used to assist with English language editing. All such edits were reviewed and verified by the authors.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments, which helped improve this study; all remaining errors are our own. We also gratefully acknowledge the support provided by the Department of Linguistics and the School of Languages and Cultures at Purdue University.

References

Sami Alanazi, Seemal Asif, Antoinette Caird-daley, and Irene Moulitsas. 2025. Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses. *Human-Intelligent Systems Integration*, pages 1–23.

Zaynab M Almutairi and Hebah Elgibreen. 2023. Detecting fake audio of arabic speakers using self-supervised deep learning. *IEEE Access*, 11:72134–72147.

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.

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. *Advances in neural information processing systems*, 31.

Rewaa Awaad, Abdelrahman Hedya, Mariam Fadel, and Ahmed B Zaky. 2024. Real-time deepfake detection for ai-generated arabic speech. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 198–201. IEEE.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119.

Nawar Halabi and 1 others. 2016. Arabic speech corpus. *Oxford text archive core collection*.

Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. 2021. *Detection theory: A user's guide*. Routledge.

Elsayed Issa. 2026. AFAD-MSA: Dataset and models for Arabic fake audio detection. *Computation*, 14(1):20.

Zahra Khanjani, Gabrielle Watson, and Vandana P Janeja. 2023. Audio deepfakes: A survey. *Frontiers in Big Data*, 5:1001063.

Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. ISCA (the International Speech Communication Association).

Kartik Kuckreja, Parul Gupta, Injy Hamed, Thamar Solorio, Muhammad Haris Khan, and Abhinav Dhall. 2025. Tell me habibi, is it real or fake? *arXiv preprint arXiv:2505.22581*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Nadine Lavan, Mathilde Irvine, Victor Rosi, and Carolyn McGettigan. 2025. Voice clones sound realistic but not (yet) hyperrealistic. *PLOS ONE*, 20(9):e0332692.

Lantian Li, Yixiang Chen, Dong Wang, and Thomas Fang Zheng. 2017. A study on replay attack and anti-spoofing for automatic speaker verification. In *Proc. Interspeech*, pages 92–96.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2025. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and 1 others. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522.

Mohamed Maged, Alhassan Ehab, Ali Mekky, Besher Hassan, and Shady Shehata. 2025. Arfake: A multi-dialect benchmark and baselines for arabic spoof-speech detection. *Preprint*, arXiv:2509.22808.

Kimberly T Mai, Sergi Bray, Toby Davies, and Lewis D Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. *Plos one*, 18(8):e0285333.

Cade Metz. 2019. Internet companies prepares to fight the 'deepfake' future. page B1. Accessed: 2025-09-01.

Nicolas M Müller, Karla Pizzi, and Jennifer Williams. 2022. Human perception of audio deepfakes. In *Proceedings of the 1st international workshop on deepfake detection for audio multimedia*, pages 85–91.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Mohamed R Shoaib, Zefan Wang, Milad Taleby Ahvanooey, and Jun Zhao. 2023. Deepfakes, misinformation, and disinformation in the era of frontier ai, generative ai, and large ai models. In *2023 international conference on computer and applications (ICCA)*, pages 1–7. IEEE.

John A Swets. 2014. *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*.

N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. 2020. Introducing the VoicePrivacy Initiative. In *Proc. Interspeech 2020*, pages 1693–1697.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, and 1 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Gabrielle Watson, Zahra Khanjani, and Vandana P Janeja. 2021. Audio deepfake perceptions in college going populations. *arXiv preprint arXiv:2112.03351*.

Daniel Weiner and Nellie Lawrance. 2023. Regulating ai deepfakes and synthetic media in the political arena. New York Times, Accessed: 2025-09-01.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*, pages 2037–2041. International Speech Communication Association.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and 1 others. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, and 1 others. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, and 1 others. 2023. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

A Audio Generation

Synthetic stimuli were produced under each model's default configuration. We made no changes to generation parameters such as temperature, stability, or similarity, and applied no post-hoc manipulation or editing to the resulting audio. This was a deliberate choice: relying on out-of-the-box settings approximates the conditions available to a typical user and ensures that detection difficulty reflects the systems' default output rather than tuning on our part.

For the binary (TTS) task, we used three proprietary models—ElevenLabs (https://elevenlabs.io/), MiniMax (https://www.minimax.io/), and ResembleAI (https://www.resemble.ai/)—accessed through their own web interfaces, without any custom code. For each genuine reference clip, we supplied the corresponding transcript to each system and generated a synthetic counterpart in the default voice/setting, producing the fake clips used in the pairwise discrimination trials.

For the unary (voice-cloning) task, we used four systems to clone each of the three reference speakers from a short sample of their authentic speech, again at default settings and with no manual editing of the output. Each system was prompted with the transcript of the corresponding genuine clip (Table 6), so all cloned variants reproduce the original wording of the speech. The two proprietary systems, Hume (https://www.hume.ai/) and MiniMax, were accessed through their web interfaces. The two open-source systems, FishAudio (https://huggingface.co/fishaudio/s1-mini) and XTTS-v2 (https://huggingface.co/coqui/XTTS-v2), were run from their publicly available HuggingFace model releases using a Python script, with default inference settings.

B Audio Stimuli Sources

Political Figure	Transcript	Source	Timestamp (dur.)
Abdel Fattah El-Sisi	شعب مصر العظيم السيدات والسادة الحضور الكريم. يظل المواطن المصري بطل هو روايتنا الوطنية.	YouTube[1]	0:00–0:10 (10s)
Mohammed bin Salman Al Saud	قامت هذه الدولة منذ ثلاثة قرون على مبادئ راسخة ترتكز على إعلاء الشريعة الإسلامية وإقامة العدل والشورى.	YouTube[2]	7:27–7:34 (7s)
Tamim bin Hamad Al Thani	ومازال أمامنا المزيد لنفعله لأن عزمنا على مواصلة تحقيق التقدم لبلدنا ولرفاهية شعبنا لا يقف عند حد.	YouTube[3]	1:35–1:43 (8s)

[1] https://www.youtube.com/watch?v=I2ZmBygvUoI
[2] https://www.youtube.com/watch?v=EPHziE2BZXY
[3] https://www.youtube.com/watch?v=ppHQJnHKuik

Table 7: Authentic source clips of political figures used for the voice-cloning experiments, showing the transcript, source URL, timestamp, and duration of each clip. The same transcripts were used to generate the cloned variants.

TTS	Audio ID	Transcript	Dur. (s)
ElevenLabs	common_voice_ar_24026124.mp3	وإن عليك لعنتي إلى يوم الدين	4
ElevenLabs	common_voice_ar_24034981.mp3	كان علىّ فعل هذا منذ أشهر خلت	4
ElevenLabs	common_voice_ar_24098642.mp3	متى سنقيم الحفلة	3
MiniMax	common_voice_ar_24033216.mp3	والوجه السابع	3
MiniMax	common_voice_ar_24024916.mp3	إننا نبحث عن منزل يضم مرآباً	4
MiniMax	common_voice_ar_24057081.mp3	والذين كفروا بآياتنا هم أصحاب المشئمة	4
ResembleAI	ARA_NORM__0077.wav	وبعثوا رسالة إلى الرئيس تتضمن مطالب لعودتهم	6
ResembleAI	ARA_NORM__0240.wav	مما يزيد الأمر تعقيداً	3
ResembleAI	ARA_NORM__0536.wav	دخلت الحضارة الإنسانية في طورٍ جديدٍ في تسعينيات القرن نفسه	7

Table 8: TTS stimuli used in the binary task, showing the synthesis model, audio ID, transcript, and duration of each clip.

C Reasoning Cue Coding Scheme

Table 9: Coding scheme for participants' free-text reasoning cues, with agreement per category between the two human coders.

Category	Definition	n	Agr. (%)	κ
Audio quality	Recording fidelity: noise, echo, clarity, quality.	33	95.4	0.52
Segmental and phonetic	Specific speech sounds: pronunciation, points of articulation, letters, diacritics.	130	96.2	0.86
Prosodic and delivery	Suprasegmentals/timing: intonation, stress, rate, pausing, elongation, rhythm, overlap, delivery.	108	97.1	0.87
Naturalness and humanness	Global human-vs-machine impression: natural/unnatural, robotic, "sounds AI", breathing, "too polished".	151	96.9	0.89
Linguistic content	Grammar, syntax, lexis, MSA correctness, fluency, coherence.	59	97.4	0.78
Recitation features	Recitation: verse pausing rules, stops, religious-register correctness.	7	99.7	0.73
Speaker identity and familiarity	Comparison to the known target speaker: "not his voice", recognizing the figure, accent/dialect of the person.	93	94.0	0.69
Subjective certainty	Subjective certainty or non-cue reasoning: confidence, "I think", "seems/feels real", guessing. May co-occur with other codes.	170	94.5	0.81
Non-substantive	Numeric-only, yes/no, or a reasoning with no stated cue.	277	87.1	0.67

D Full Tukey-adjusted Pairwise Contrasts Across Models

This appendix reports the complete set of Tukey-adjusted pairwise contrasts between speech synthesis models for both tasks, complementing the model-condition analyses in Section 4.3. Table 10 presents the contrasts among the binary (TTS) systems and Table 11 those among the unary (voice-cloning) systems. No binary contrast was significant. In the unary task, the reliable differences were that XTTS-v2 was detected significantly more often than FishAudio and MiniMax. In addition, Figure 5 shows the confusion matrices for binary and unary tasks.

Table 10: Tukey-adjusted pairwise contrasts among binary generated-item models.

Contrast	β [95% CI]	SE	z	p
ElevenLabs – MiniMax	0.055 [–1.973, 2.083]	0.865	0.063	.998
ElevenLabs – ResembleAI	1.390 [–0.627, 3.408]	0.861	3.885	.239
MiniMax – ResembleAI	1.335 [0.690, 3.361]	0.864	1.545	.270

Table 11: Tukey-adjusted pairwise contrasts among unary generated-item models.

Contrast	β [95% CI]	SE	z	p
FishAudio – Hume	–1.497 [–3.882, 0.887]	0.928	–1.613	.371
FishAudio – MiniMax	–0.129 [–2.510, 2.253]	0.927	–0.139	.999
FishAudio – XTTS-v2	–3.152 [–5.660, –0.645]	0.976	–3.230	.007
Hume – MiniMax	1.368 [–1.019, 3.756]	0.929	1.472	.454
Hume – XTTS-v2	–1.655 [–4.143, 0.833]	0.969	–1.709	.319
MiniMax – XTTS-v2	–3.023 [–5.537, –0.510]	0.978	–3.091	.011

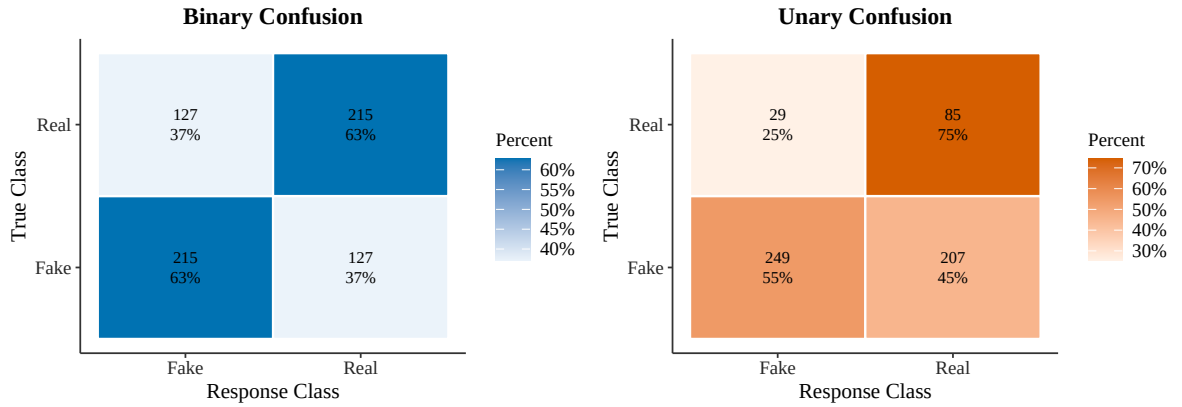
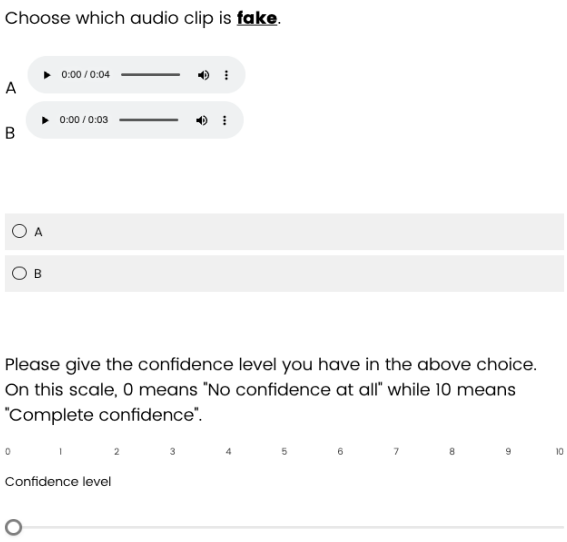
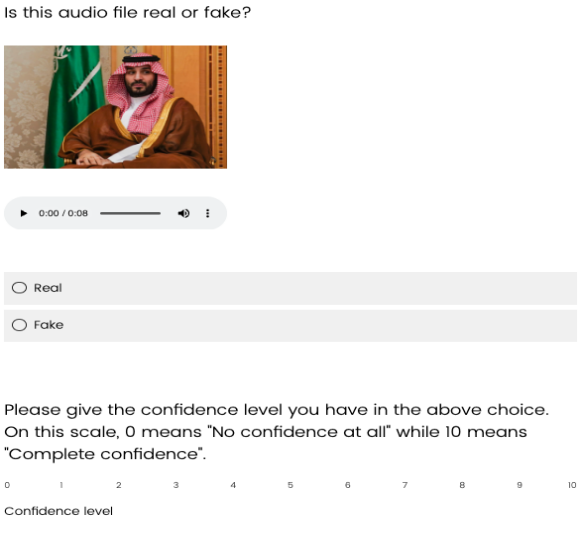


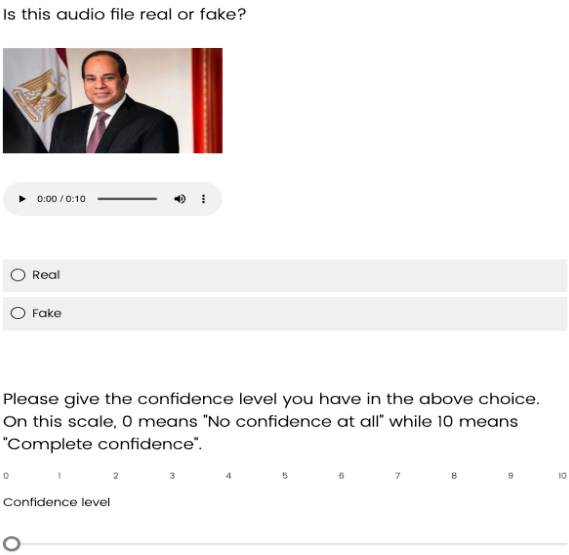
Figure 5: Confusion matrices for binary and unary tasks. Cells show counts and percentages. Note: For visualization only, each forced-choice binary response was expanded into two clip-level inferred classifications: the selected clip was coded as "fake" and the unselected clip as "real." Thus, the binary matrix contains 684 inferred classifications from 342 pairwise responses and should not be interpreted as independent clip-level judgments.



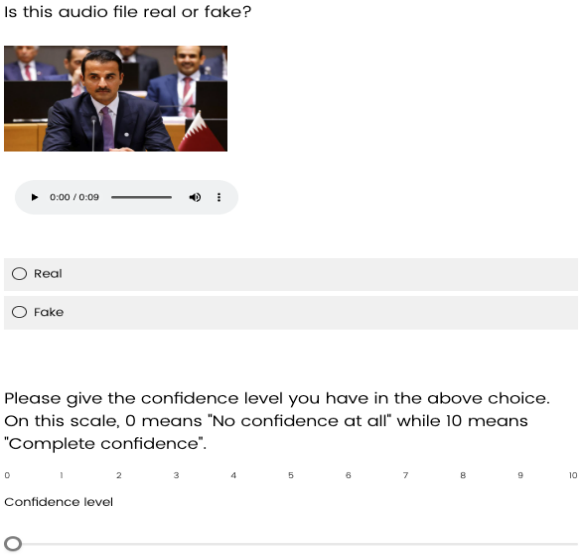
(a)



(b)



(c)



(d)

Figure 6: Example from the experiment. Panel (a) shows a binary text-to-speech task; panels (b)–(d) show unary voice-cloning for the three public figures whose voices were cloned. Each task presented the audio (and, for the voice-cloning trials, an image of the speaker), followed by the authenticity judgment, a 0–10 confidence rating, and a free-text prompt asking participants to explain their reasoning. The images of public figures are reproduced here solely to document the experimental interface for research purposes.

Boosting Transferable Adversarial Attacks against Deep Reinforcement Learning

Zexin Li[1] **Ruili Yao**[2] **Yiming Zeng**[3] **Xiaoxue Gao**[4]*

[1]Nanyang Technological University

[2]University of California, Riverside [3]University of Connecticut

[4]School of Artificial Intelligence,

The Chinese University of Hong Kong, Shenzhen

zexin.li@ntu.edu.sg

Abstract

Most adversarial attacks on deep reinforcement learning (DRL) assume white-box access to the victim policy, which rarely holds in practice. This paper studies *transfer-based black-box* attacks on DRL: the attacker crafts observation perturbations on a white-box surrogate agent and feeds them to an unknown victim. We formulate the attack as return minimization under a per-step perturbation budget. We first show that transplanting transferable image-classification attacks (FGSM, MI-FGSM, and NI-FGSM) with a per-step objective yields perturbations that transfer but are no stronger than random noise of the same budget. We then propose a *trajectory-level* attack that optimizes a sequence of perturbations over a receding horizon through a differentiable model of the environment and a temperature-smoothed surrogate policy, with the same optimizers. On CartPole-v1 with ten DQN and DDQN agents and 100 surrogate–victim pairs, the trajectory-level attack outperforms per-step attacks and random noise in the white-box, cross-model, and cross-algorithm settings.

1 Introduction

Deep reinforcement learning (DRL) has enabled complex control from learned representations (Mnih et al., 2015; Peng et al., 2018; Kahn et al., 2018), but its policies remain vulnerable to small observation perturbations (Huang et al., 2017; Lin et al., 2017; Zhang et al., 2020). In a sequential system, a corrupted observation affects both the current action and the states encountered later. Evaluating this vulnerability therefore requires measuring cumulative return, beyond whether an attack changes an individual action.

Many attacks assume access to the victim's parameters. For an inaccessible deployed policy, a practical alternative is to craft perturbations on a white-box surrogate and transfer them to the victim (Huang et al., 2017; Pan et al., 2020). Prior work demonstrates shared adversarial features across MDPs (Korkmaz, 2022) and natural black-box vulnerabilities even after adversarial training (Korkmaz, 2023). These findings establish that transfer is possible, but leave an important design question: how should an attacker exploit sequential dynamics to make transferred perturbations more damaging?

A natural baseline applies transferable image-classification attacks independently at each step. However, changing the surrogate's preferred action is only a proxy for reducing the victim's return. A well-trained controller may correct an isolated error, while a perturbation tailored to one decision boundary may become ineffective on another policy. On CartPole, we find that per-step FGSM, MI-FGSM, and NI-FGSM transfer across agents but do not outperform random corner noise at the same budget. This motivates optimizing the consequences of a sequence of perturbations rather than repeatedly targeting a single action.

We optimize perturbation sequences through differentiable environment dynamics over a receding horizon. A temperature-smoothed surrogate policy permits differentiable rollouts, and momentum or Nesterov updates seek trajectories approaching failure. Only the first perturbation is applied before replanning. The objective thus incorporates future state evolution while retaining the same per-step perturbation budget as the baselines.

Our contributions are a return-minimization formulation linking per-step and trajectory objectives (§3.1), a trajectory-level attack using FGSM, MI-FGSM, and NI-FGSM (§3.2), and an evaluation over ten DQN/DDQN agents and 100 surrogate–victim pairs (§4). Trajectory attacks outperform per-step attacks and random noise in white-box, cross-model, and cross-algorithm settings.

*Corresponding author. Assistant Professor, School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen.

Proceedings of the Workshop on Secure, Safe, and Privacy-Preserving Speech and Audio AI (TrustAudio 2026), pages 50–56
November 10, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Transferable attacks. Iterative optimization (Kurakin et al., 2018; Madry et al., 2018), momentum (Dong et al., 2018), Nesterov look-ahead (Lin et al., 2020), variance tuning (Wang and He, 2021), and input diversity (Xie et al., 2019) improve transfer. Applications include face recognition (Li et al., 2023b), speech (Gao et al., 2024), text generation (Li et al., 2023a), efficiency attacks (Chen et al., 2023, 2026), and backdoors (Cheng et al., 2026; Guo et al., 2022).

Adversarial attacks on DRL. Huang et al. (2017) studied FGSM (Goodfellow et al., 2015) attacks and transfer in Atari (Bellemare et al., 2013); Lin et al. (2017) planned target actions through learned dynamics. Other work learns white-box adversaries (Zhang et al., 2020; Sun et al., 2022), uses domain randomization for transfer (Pan et al., 2020), or manipulates robot teams (Nikkhoo et al., 2023; Li et al., 2026a). Shared adversarial directions (Korkmaz, 2022) and natural black-box vulnerabilities (Korkmaz, 2023) motivate our focus on trajectory optimization.

Diagnosing and evaluating robustness. Korkmaz and Brown-Cohen (2023) detect adversarial directions using local quadratic approximations of the policy loss. Korkmaz (2024) analyzes how unstable directions evolve across states and time, revealing limitations of robust training. Korkmaz (2026) shows that enforcing local invariance can impair counterfactual value learning.

Sequence-level objectives. SPEAR (Li et al., 2026b) uses order-aware symbolic trajectory alignment to reward language-model reasoning distillation. It shares our emphasis on sequence structure, but addresses distillation rather than adversarial control.

3 Method

3.1 Problem Formulation

We model the task as a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. The victim is a value-based agent with Q-network $Q_v(\cdot, \cdot; \theta_v)$ and greedy policy $\pi_v(s) = \arg\max_a Q_v(s, a; \theta_v)$. At each step t, the attacker observes the true state s_t and injects a perturbation δ_t into the victim's observation (Figure 1), so that the victim acts on $\tilde{s}_t = s_t + \delta_t$ while the environment evolves according to the true state. The attacker aims to minimize the victim's expected return under a per-step ℓ_∞ budget ϵ:

$$\begin{aligned} \min_{\{\delta_t\}_{t=0}^{H}} \; & J(\delta) = \mathbb{E}\Big[\textstyle\sum_{t=0}^{H} \gamma^t r(s_t, a_t)\Big] \\ \text{s.t.} \; & a_t = \pi_v(s_t + \delta_t), \quad \|\delta_t\|_\infty \le \epsilon, \\ & s_{t+1} \sim P(\cdot \mid s_t, a_t), \end{aligned} \tag{1}$$

where H is the episode horizon. In the black-box setting, θ_v is unknown and the attacker cannot query π_v while crafting perturbations. It trains a surrogate Q-network $Q_s(\cdot, \cdot; \theta_s)$ in the same environment, possibly with a different seed or algorithm, and uses it as a stand-in for Q_v. Even with a surrogate, Eq. (1) is hard to optimize: the return is a non-differentiable function of a long sequence of discrete actions and stochastic transitions.

Per-step relaxation. The standard relaxation, which we use as a baseline, optimizes each perturbation independently. We choose a lure action $a^\dagger$ whose persistent execution drives the system toward failure (for CartPole, always pushing the cart in the same direction), and at each step we solve

$$\delta_t^* = \arg\max_{\|\delta\|_\infty \le \epsilon} \log \operatorname{softmax}\big(Q_s(s_t + \delta, \cdot)\big)_{a^\dagger} \tag{2}$$

with FGSM (Goodfellow et al., 2015), MI-FGSM (Dong et al., 2018), or NI-FGSM (Lin et al., 2020). This objective is myopic: one wrong action changes the state only slightly and the victim's next decision can undo it, so a per-step attack must succeed at almost every step, which is hard under a small budget and harder after transfer.

3.2 Trajectory-Level Transfer Attack

We instead optimize perturbations with respect to their effect on the future trajectory. We assume the attacker has a differentiable model f of the environment dynamics, $\hat{s}_{k+1} = f(\hat{s}_k, u_k)$, where u_k is a continuous relaxation of the action. For CartPole, f is the known physical model; in general, it can be learned from interaction data, as in the enchanting attack (Lin et al., 2017). To make the rollout differentiable, we replace the surrogate's greedy action with a temperature-smoothed policy

$$\tilde{\pi}_s(a \mid x) = \operatorname{softmax}\big(Q_s(x, \cdot)/\kappa\big)_a, \tag{3}$$

and apply the expected action $u_k = \mathbb{E}_{a \sim \tilde{\pi}_s(\cdot \mid \hat{s}_k + \delta_k)}[a]$ (for CartPole, the expected force). At step t, the attacker optimizes a sequence of K perturbations $\Delta_t = (\delta_t, \ldots, \delta_{t+K-1})$ to maximize the proximity of the predicted states to

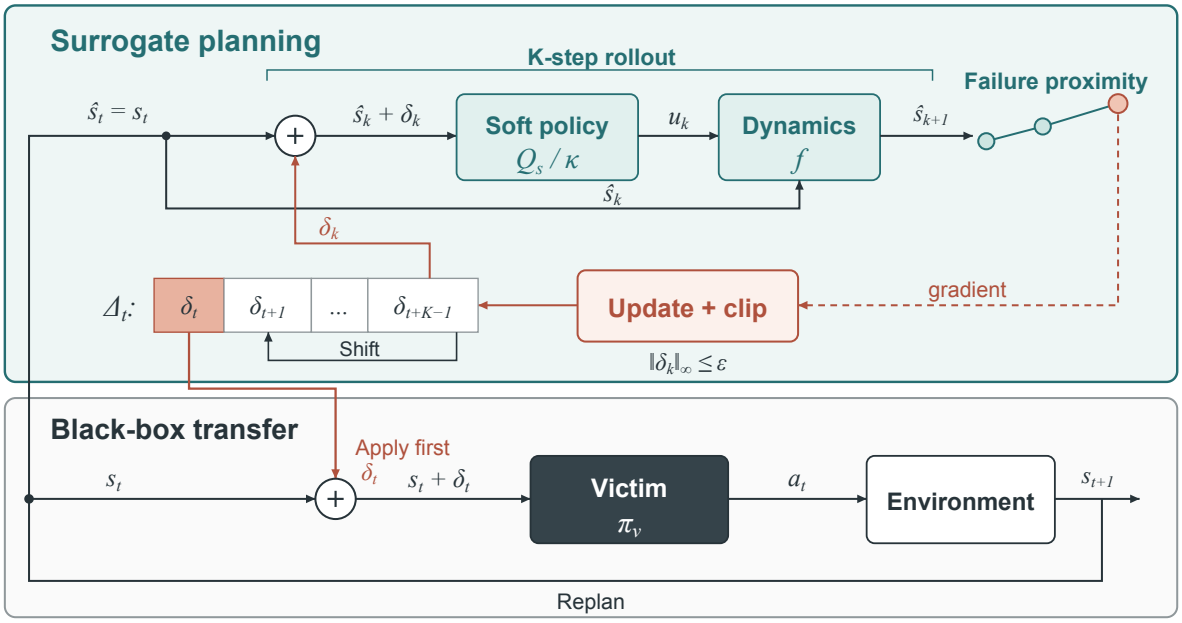


Figure 1: Trajectory-level transfer attack. The attacker optimizes a bounded perturbation sequence through a temperature-smoothed surrogate and differentiable dynamics to maximize predicted failure proximity. Only the first perturbation is applied to the victim's observation; the remainder warm-starts the next plan. The environment evolves from the true state. Dashed arrows denote gradients; no victim parameters or queries are used during crafting.

the failure region:

$$\max_{\|\delta_k\|_\infty \le \epsilon} \mathcal{L}_{\text{traj}}(\Delta_t) = \sum_{k=1}^{K} \phi(\hat{s}_{t+k}), \quad \hat{s}_t = s_t, \tag{4}$$

where $\phi(s) = \max\{(\theta/\theta_{\max})^2, (x/x_{\max})^2\}$ measures how close the pole angle θ and the cart position x are to their termination thresholds. Following model-predictive control, only δ_t is applied, and the remaining perturbations warm-start the optimization at the next step.

We solve Eq. (4) with the same operators as in the per-step case, applied to the whole sequence Δ_t. With step size $\alpha = \epsilon/T$ and momentum decay μ, MI-FGSM accumulates an ℓ_1-normalized gradient momentum,

$$g_{j+1} = \mu\, g_j + \frac{\nabla_\Delta \mathcal{L}_{\text{traj}}(\hat{\Delta}_j)}{\|\nabla_\Delta \mathcal{L}_{\text{traj}}(\hat{\Delta}_j)\|_1}, \tag{5}$$

with $\hat{\Delta}_j = \Delta_j$ for MI-FGSM and the Nesterov look-ahead $\hat{\Delta}_j = \Delta_j + \alpha\mu\, g_j$ for NI-FGSM, followed by

$$\Delta_{j+1} = \text{Clip}_\epsilon\big(\Delta_j + \alpha \cdot \text{sign}(g_{j+1})\big). \tag{6}$$

FGSM takes a single step of size ϵ. Algorithm 1 summarizes the full attack.

Algorithm 1 Trajectory-Level Transfer Attack

Require: Surrogate Q-network Q_s, black-box victim policy π_v, dynamics model f, budget ϵ, horizon K, iterations T, momentum decay μ, temperature κ, operator $\mathcal{O} \in \{\text{FGSM}, \text{MI}, \text{NI}\}$
Ensure: Perturbation sequence that minimizes the victim's return J in Eq. (1)
Observe s_0; initialize $\Delta \leftarrow \mathbf{0}_K$
for $t = 0, 1, \ldots, H$ **do**
  $\alpha \leftarrow \epsilon/T$, $\Delta_0 \leftarrow \Delta$, $g_0 \leftarrow \mathbf{0}$
  for $j = 0, \ldots, T-1$ **do**
    $\hat{\Delta}_j \leftarrow \Delta_j + \alpha\mu\, g_j$ (NI) or Δ_j (MI)
    Roll out f from s_t with $\tilde{\pi}_s$ (Eq. (3)) under $\hat{\Delta}_j$
    Compute $\mathcal{L}_{\text{traj}}$ via Eq. (4) and backpropagate
    Update g_{j+1} via Eq. (5) and Δ_{j+1} via Eq. (6)
  end for ▷ FGSM: one step of size ϵ
  Victim acts $a_t \leftarrow \pi_v(s_t + \delta_t)$, $\delta_t = \Delta_T[0]$ ▷ Transfer
  Observe r_t, s_{t+1}; **break** if terminal
  $\Delta \leftarrow \Delta_T$ shifted by one step ▷ Warm start
end for

Transfer rationale. The objective uses dynamics shared by surrogate and victim to target failure over multiple steps, rather than a single surrogate action. Temperature smoothing reduces sensitivity to the surrogate's exact decision boundary; momentum stabilizes optimization (Dong et al., 2018). Their individual contributions require separate ablations.

Table 1: CartPole victim returns (lower is stronger): mean and standard deviation over five victims (white-box), 20 ordered same-algorithm pairs (cross-model), or 25 pairs per direction (cross-algorithm). Random baselines are evaluated directly at the same budget. Middle/bottom groups use per-step/trajectory objectives; bold marks each column's strongest attack.

	White-box				Cross-model transfer				Cross-algorithm transfer			
Budget	0.05		0.1		0.05		0.1		0.05		0.1	
Victim	DQN	DDQN	DQN	DDQN	DQN	DDQN	DQN	DDQN	DDQN	DQN	DDQN	DQN
No attack	500.0±0	485.4±24	500.0±0	485.4±24	500.0±0	485.4±24	500.0±0	485.4±24	485.4±24	500.0±0	485.4±24	500.0±0
Uniform noise	375.0±130	304.7±193	172.7±53	247.0±193	375.0±130	304.7±193	172.7±53	247.0±193	304.7±193	375.0±130	247.0±193	172.7±53
Sign noise	240.5±92	284.9±204	74.3±38	121.2±133	240.5±92	284.9±204	74.3±38	121.2±133	284.9±204	240.5±92	121.2±133	74.3±38
FGSM	322.0±140	336.0±179	262.5±177	214.7±163	330.9±138	314.7±204	173.4±128	274.6±185	302.1±212	411.2±124	163.4±138	364.5±140
MI-FGSM	315.8±157	277.0±200	159.5±174	125.4±188	338.9±140	317.0±209	147.7±120	211.2±190	317.7±214	400.9±127	148.6±160	324.0±166
NI-FGSM	304.3±163	286.3±201	153.5±176	125.4±189	360.5±138	306.7±206	158.2±121	215.6±189	319.2±204	400.2±130	169.7±184	322.3±164
Traj-FGSM	113.0±108	243.9±211	56.5±52	30.6±23	**190.0**±147	**240.0**±205	77.3±54	43.3±30	**259.1**±214	**181.8**±126	**41.6**±27	82.3±74
Traj-MI-FGSM	**73.5**±49	**157.3**±178	19.5±13	15.7±6	203.4±154	268.1±217	60.6±95	**41.1**±40	266.9±216	186.3±162	47.5±47	**53.1**±57
Traj-NI-FGSM	74.1±44	173.1±177	**18.5**±12	**13.9**±4	202.9±149	271.6±219	**59.6**±78	45.5±40	268.9±214	189.3±154	55.1±50	58.7±62

4 Experiments

4.1 Experimental Setup

Environment and agents. We use CartPole-v1 from Gymnasium (Brockman et al., 2016; Barto et al., 1983). The observation is a raw 4-D vector (cart position, cart velocity, pole angle, and pole angular velocity), there are two actions (push left or right), and the agent receives a reward of $+1$ per step, up to a maximum return of 500. We train DQN (Mnih et al., 2015) and Double DQN (DDQN) (Van Hasselt et al., 2016) agents with five seeds each, which gives ten models. All agents use a two-layer MLP (128 units), Adam (learning rate 5×10^{-4}), batch size 128, soft target updates, and 600 training episodes, and we keep the best checkpoint under periodic greedy evaluation. Figure 2 shows the training curves. Without attack, eight agents obtain the maximum return of 500, and the other two average 489.3 and 437.8.

Attacks. We compare the per-step FGSM, MI-FGSM, and NI-FGSM attacks (Eq. (2), lure action $a^\dagger$ = push left) with their trajectory-level counterparts Traj-FGSM, Traj-MI-FGSM, and Traj-NI-FGSM (Algorithm 1). All iterative attacks use $T = 10$ iterations and $\mu = 1.0$. The trajectory attacks use horizon $K = 20$ and temperature $\kappa = 2$ and roll out the CartPole equations of motion. We also report two random baselines with the same budget: *Uniform* noise drawn from $[-\epsilon, \epsilon]^4$, and *Sign* noise that places every coordinate at a random corner, $\pm\epsilon$. Budgets are given in raw observation units, and we report $\epsilon \in \{0.05, 0.1\}$.

Protocol. We evaluate three settings: *white-box* (the surrogate is the victim), *cross-model* (same algorithm, different seed; 20 ordered pairs per algorithm), and *cross-algorithm* (DQN↔DDQN; 25 pairs per direction). Each evaluation runs 10 episodes with fixed environment seeds, and we report the victim's average return (lower is stronger).

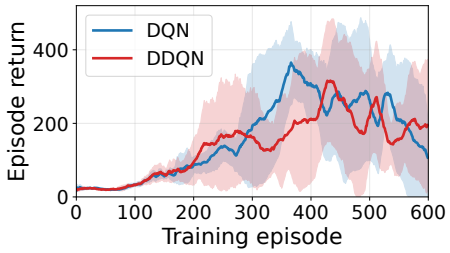


Figure 2: CartPole training returns over five seeds: 20-episode moving average with one standard deviation shaded. Periodic greedy evaluation selects each checkpoint.

4.2 Main Results

Per-step attacks transfer but are weak. At $\epsilon = 0.1$, white-box MI-FGSM and NI-FGSM reduce DQN returns to 159.5 and 153.5, versus 262.5 for FGSM (Table 1). Cross-model MI-FGSM achieves 147.7, but remains weaker than Sign noise (74.3). Cross-algorithm transfer is asymmetric: DDQN-crafted attacks leave DQN returns above 320.

Trajectory-level attacks are much stronger. In the white-box setting at a budget of 0.1, Traj-NI-FGSM reduces the return to 18.5 on DQN and 13.9 on DDQN, 4–9× lower than Sign noise. At the smaller budget of 0.05, where all per-step attacks leave the victims above a return of 277, Traj-MI-FGSM reaches 73.5 on DQN and 157.3 on DDQN.

Transfer across seeds and algorithms. For cross-model transfer at a budget of 0.1, the trajectory attacks reduce DQN and DDQN victims to

59.6 and 41.1, compared with 147.7 and 211.2 for the best per-step attack and 74.3 and 121.2 for Sign noise. They also remove the asymmetry of cross-algorithm transfer: perturbations crafted on DDQN reduce DQN victims to 53.1, compared with more than 320 for per-step attacks. This supports our hypothesis that an objective defined through the shared dynamics transfers better than one defined through individual actions. At a budget of 0.05, they remain the strongest in every column, with a smaller margin (e.g., 190.0 compared with 240.5 for Sign noise).

Momentum and look-ahead. At $\epsilon = 0.1$, momentum and look-ahead reduce white-box DQN returns from 56.5 (Traj-FGSM) to about 19. The three trajectory operators perform similarly under transfer, suggesting the trajectory objective drives most of the improvement.

5 Conclusion

On CartPole, per-step attacks transfer but are weaker than random corner noise. Optimizing perturbation sequences through differentiable dynamics strengthens white-box and transfer attacks, motivating defenses against temporally coordinated perturbations. Future work will address high-dimensional observations and learned dynamics.

Limitations

Our evaluation is limited to CartPole's low-dimensional observations, known dynamics, and DQN/DDQN agents. The attack assumes access to the true state and a differentiable dynamics model; its effectiveness with visual inputs, learned or inaccurate dynamics, other environments, and adversarially trained victims remains untested. Reported standard deviations describe variability across victims or pairs, but shared models prevent treating all pairs as independent samples for significance testing. Ablations of the planning horizon, temperature, and warm start, as well as measurements of attack latency, remain future work.

Ethical Statement

This work studies adversarial attacks to support robustness evaluation of reinforcement learning systems. The reported experiments use trained agents in simulated CartPole. The method has dual-use potential: transferable observation perturbations could be misused to disrupt deployed controllers, with possible physical harm in safety-critical applications. Such testing should therefore be restricted to systems for which explicit authorization has been obtained and conducted in simulation or controlled environments with appropriate safeguards.

Acknowledgments

This research is supported by UDF01004676.

References

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI gym. *arXiv preprint arXiv:1606.01540*.

Yiming Chen, Simin Chen, Zexin Li, Wei Yang, Cong Liu, Robby T. Tan, and Haizhou Li. 2023. Dynamic transformers provide a false sense of efficiency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7164–7180.

Yiming Chen, Zexin Li, Xianghu Yue, Robby T. Tan, and Haizhou Li. 2026. NaturalSloth: Revisiting denial-of-service attacks on large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yezheng Cheng, Zexin Li, Jiaqi Wu, Zhihong Zhang, and Simin Chen. 2026. VLMSysTrojan: Stealthy system-aware backdoor attacks against vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193.

Xiaoxue Gao, Zexin Li, Yiming Chen, Cong Liu, and Haizhou Li. 2024. Transferable adversarial attacks against ASR. *IEEE Signal Processing Letters*.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Junfeng Guo, Ang Li, and Cong Liu. 2022. Backdoor detection and mitigation in competitive reinforcement learning. *arXiv preprint arXiv:2202.03609*.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

Gregory Kahn, Adam Villaflor, Bosen Ding, Pieter Abbeel, and Sergey Levine. 2018. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8.

Ezgi Korkmaz. 2022. Deep reinforcement learning policies learn shared adversarial features across MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7229–7238.

Ezgi Korkmaz. 2023. Adversarial robust deep reinforcement learning requires redefining robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8369–8377.

Ezgi Korkmaz. 2024. Understanding and diagnosing deep reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 25265–25275. PMLR.

Ezgi Korkmaz. 2026. How to lose inherent counterfactuality in reinforcement learning. In *International Conference on Learning Representations*.

Ezgi Korkmaz and Jonah Brown-Cohen. 2023. Detecting adversarial directions in deep reinforcement learning to make robust decisions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 17534–17543. PMLR.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC.

Yufei Li, Zexin Li, Yingfan Gao, and Cong Liu. 2023a. White-box multi-objective adversarial attack on dialogue generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1778–1792.

Zexin Li, Bangjie Yin, Taiping Yao, Junfeng Guo, Shouhong Ding, Simin Chen, and Cong Liu. 2023b. Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24626–24637.

Zexin Li, Ziliang Zhang, Hyoseung Kim, and Cong Liu. 2026a. PIMbot: A self-adaptive attack framework for adversarial manipulation of multi-robot reinforcement learning. *arXiv preprint arXiv:2605.23027*.

Zhuochun Li, Yuelyu Ji, Yiming Zeng, and Daqing He. 2026b. SPEAR: Distilling domain-adaptive reasoning skeletons via sequential symbolic alignment in reinforcement learning. *arXiv preprint arXiv:2608.26550*.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. 2020. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. 2017. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3756–3762.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Shahab Nikkhoo, Zexin Li, Aritra Samanta, Yufei Li, and Cong Liu. 2023. PIMbot: Policy and incentive manipulation for multi-robot reinforcement learning in social dilemmas. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5630–5636.

Xinlei Pan, Yulong Cao, Xindi Wu, Eric Zelikman, Chaowei Xiao, Yanan Sui, Rudrasis Chakraborty, and Ronald S Fearing. 2020. Transferable adversarial attacks on deep reinforcement learning with domain randomization. In *CVPR Workshops*.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):143.

Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. 2022. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep RL. In *International Conference on Learning Representations (ICLR)*.

Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Xiaosen Wang and Kun He. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1924–1933.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. 2020. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21024–21037.

Author Index

Alshahrani, Rahaf, 29
Azime, Israel Abebe, 11

Cheng, Yi-Fang, 35

Emani, Murali, 24

Gao, Xiaoxue, 50

Haile, Abrham, 11

Issa, Elsayed, 35

Li, Zexin, 50

Manjusha, Kashyap R, 24
Marilign, Dagnachew Mekonnen, 11
Molero, Enrique, 11

Niwa, Ayana, 29

Patil, Yashas Harshad, 1
Powell, Louis, 11

Rawat, Krish, 1

Shanbhag, Anup, 1

Yao, Ruili, 50

Zeng, Yiming, 50