TOWARDS QUANTIFYING BENCHMARK OPTIMIZATION IN ASR MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Public benchmarks are important measures of Automatic Speech Recognition (ASR) model capabilities. However, by nature of being public, there is risk of models being optimized for these benchmarks in ways that do not generalize well to real-world data. We present a methodology for quantifying benchmark optimization, focusing on cases where the audio underdetermines the reference transcript. We identify three families of behavioral probes that reveal models' capabilities of reproducing benchmark reference spans despite underdetermined audio: reference disagreement, masked-number recovery, and orthographic agreement. We find that the highest-scoring open source models output verbatim reference transcript spans even when the relevant audio is contradictory, masked, or ambiguous. Models respond to narrow acoustic cues to reproduce the reference transcript instead of the audio-faithful output. This benchmark-conditioned behavior can be causally manipulated via low-rank linear steering or simply appending audio to the end of a segment in some cases. Fine-tuning on public training splits brings several lower-scoring models to the VoxPopuli WER and reference-error reproduction of the highest-scoring ones, while worsening their WER on a human-annotated subset. Overall, our results indicate that high-performing models exhibit benchmark-conditioned behaviors that can inflate benchmark performance without reflecting improved general-purpose transcription ability.

1 INTRODUCTION

Researchers have reported human-level Automatic Speech Recognition (ASR) performance on public benchmarks for nearly a decade (Amodei et al., 2016; Xiong et al., 2016), yet a persistent gap separates benchmark performance from real-world utility (Radford et al., 2023; Xie et al., 2026). Ideally, a low word error rate (WER) on a benchmark reflects a general ability to transcribe speech that extends to unseen, real-world audio. Because benchmarks are public, however, models can be optimized to drive their reported WER down in ways that are orthogonal to—or even actively harmful to—real-world transcription abilities, causing benchmark scores to overstate a model's general-purpose performance.

We define benchmark optimization (colloquially known as benchmaxxing) as gains in reported performance that arise from reliance on benchmark-specific artifacts rather than from a generalizable improvement in transcription ability.

To measure this phenomenon, we construct cases in which the audio does not uniquely support the exact reference transcript. For instance, when a benchmark entry contains a transcription error, an audio-faithful transcriber should place low probability on the (erroneous) reference transcript. Similarly, if the audio for a word is masked, the model should also place limited probability on that word. Finally, when a word admits two phonetically and semantically equivalent renderings, the model should not systematically match the benchmark's local reference convention. In all these "benchmark optimization probes," systematic reference matching at rates far above chance suggests that the model is using the benchmark's acoustic cue as a shortcut rather than faithfully listening to the audio. We find that these behaviors persist in synthetic speech using clones of the voices of speakers from the benchmark's evaluation set, but weaken for clones of generic speakers or previously unseen speakers from independently collected data in the same source domain. Beyond establishing that benchmark optimization happens, we ask when and how this behavior is triggered.

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

We find that models override audio-supported content, triggered by a fairly narrow range of inputs, and largely operate faithfully off the benchmark's distribution. However, steering a model's internal activations along a low-dimensional linear direction or appending benchmark-like or generic audio can suppress the benchmark-conditioned behavior and partly re-induce it. The behavior thus depends on a narrow set of acoustic cues: it appears on benchmark recordings and test-speaker clones but often weakens on generic voices and new speakers from the same domain, limiting its effect off-distribution while inflating measured performance on the benchmark. Finally, we show that one epoch of fine-tuning on public training splits is enough to reorder the benchmark's WER ranking: previously low-ranked models move up (Whisper from last to fourth of seven), while some high-scoring models fall in both rank and absolute WER.

2 Related Work

It is well established that state-of-the-art ASR model WER on standard benchmarks often overstates real-world ability (Szymański et al., 2020; Likhomanenko et al., 2021; Radford et al., 2023). The distributional limits of industry-standard datasets have been studied and numerous recent works have proposed new robustness-oriented ASR evaluations targeting acoustic degradation (Shah et al., 2025; Xie et al., 2026), far-field (Dai et al., 2025), and other real-world conditions (Ayllon et al., 2026; Bezzam et al., 2026; Tay et al., 2026). However, this line of work treats the benchmark–reality gap as a coverage problem, fixed by adding benchmarks for missed conditions. Our results show that it is also a measurement problem: public benchmarks can reward benchmark-specific behavior even when the model does not transcribe the audio faithfully. New datasets are necessary but insufficient: without a framework to detect and measure benchmark optimization, new benchmarks may be subject to the same measurement distortions as existing evaluation sets.

Earlier ASR systems added bespoke text post-processing steps per benchmark to optimize performance on different orthographic conventions, which can account for a significant portion of cross-corpus error (Likhomanenko et al., 2021); reporting model performance across an array of benchmarks using a single text post-processing procedure has alleviated the use of bespoke processing steps to benchmark-optimize (Gandhi et al., 2022; Srivastav et al., 2026). However, this shift merely pushes the issue of benchmark optimizing using bespoke convention matching (Akeret, 2026) to the model level. A model which is expressive enough to learn the conventions of multiple benchmarks and alternate between conventions based on arbitrary acoustic cues can still hill-climb on benchmarks independent of its general-purpose competency.

Deep learning models have traditionally been known to exploit statistical regularities in the training data as shortcuts to completing a task in a variety of domains (Jo & Bengio, 2017; Poliak et al., 2018; Gururangan et al., 2018; McCoy et al., 2019; Liu et al., 2024; Geirhos et al., 2020). With the emergence of speech-large language models (LLMs) for ASR, there is evidence of evaluation set transcripts leaking into speech–LLM decoder backbones (Tseng et al., 2025). However, this form of contamination is localized to a specific training paradigm, while our results show evidence of benchmark optimization across multiple architectures. Moreover, authors have found the impact of decoder data contamination on WER to be relatively small, reinforcing recent NLP research which suggests that "training on the test task" is a more prominent form of benchmark optimization than data contamination (Dominguez-Olmedo et al., 2024). How models optimize for ASR benchmarks therefore needs its own measurement framework.

Researchers have begun to apply mechanistic interpretability techniques to automatic speech recognition (ASR) (Pluth et al., 2026; Glazer et al., 2026). We apply related interventions to benchmark optimization, using context manipulation and activation steering (Turner et al., 2023; Rimsky et al., 2024) to localize and causally manipulate benchmark-specific transcription behavior.

3 Method

3.1 Experimental setup

Models. We evaluate 11 widely used open source ASR models spanning the two dominant architectures: *encoder–decoder* attention/transducer models (Whisper-Large-v3 (Radford et al., 2023), Cohere-Transcribe (Cohere Labs, 2026), Parakeet-TDT-0.6B-v2 (NVIDIA NeMo Team, 2025b; Xu

et al., 2023), Moonshine-Streaming (Jeffries et al., 2024)) and *speech–LLM* models (Canary-Qwen-2.5B (NVIDIA NeMo Team, 2025a; Chen et al., 2024), Granite-Speech-4.1-2B (Saon et al., 2025), Higgs-Audio-v3-8B (Boson AI, 2026), Kimi-Audio-7B (KimiTeam, 2025), Phi-4-Multimodal (Microsoft, 2025), Qwen3-ASR-0.6B (Shi et al., 2026), Voxtral-Mini-3B (Liu et al., 2025)). We report teacher-forced likelihood metrics for all models but Parakeet-TDT, whose token-and-duration transducer architecture makes reading logits at arbitrary positions difficult.

Datasets. We focus our analysis on **VoxPopuli** (Wang et al., 2021) (English), which consists of European Parliamentary recordings and is among the most widely used ASR benchmarks. The dataset has a train, validation, and test split but 40% of speakers in the test split are leaked into the training split in the Hugging Face version of the dataset. Where applicable, we extend our findings to **LibriSpeech** (Panayotov et al., 2015) (CLEAN and OTHER). As a held-out control, we use **DAIKON**, a private set of 450 conversational clips drawn from the naturalistic dyadic-conversation collection of Tzirakis et al. (2026); these clips postdate and lie outside every model's training data and also point to how the capabilities measured in the public benchmarks translate to real-world, conversational audio. To probe the limits of when benchmark optimized behavior occurs, we add two further sources. We use Qwen3-TTS (Hu et al., 2026) to generate synthetic samples using a variety of reference speakers. We also collect fresh data corresponding to a new held-out dataset from the same domain as both benchmarks. We scrape European Parliament recordings from June 2026 (EP-FRESH)—after every model's training cutoff—following the original VoxPopuli collection procedure. Manual inspection confirmed these transcript/clip pairs are in-distribution, down to characteristic VoxPopuli reference errors such as courtesy-expression omission.

For LibriSpeech, we analogously collect 2026 LibriVox recordings from 14 newly active readers whose catalog histories begin after every model's training cutoff (LIBRI-FRESH).

3.2 PROBES AND READOUTS

Every behavioral probe marks a set of positions where the audio x underdetermines the reference and contrasts the *reference rendering* r with a competing *audio-true* (or acoustically equivalent) rendering a (Table 1). We start our analysis by looking only at the surface transcript produced by standard greedy decoding; ACCEPT-REF is the fraction of these positions at which the model emits r rather than a. A transcriber faithful to the sound should follow a at high rates, so a high ACCEPT-REF indicates the model M is reproducing the benchmark's reference beyond its acoustic content. Each probe in Sections 3.3–3.5 supplies its own (r, a); we name the per-probe instances *reference-disagreement*, *masked*, and *orthographic* ACCEPT-REF.

To isolate the effect of the raw language model prior compared to the end-to-end model, we subtract the silenced-audio prior $x_\emptyset$ (the clip with its waveform zeroed) from the teacher-forced log-likelihood of the reference span given the audio and the flanking transcript context. To account for different span lengths in a way that generalizes across tokenizers, we normalize by the character count of the span, $|r|_{char}$. We define the *audio lift* as:

$$\lambda(r) \;=\; \frac{\log p_M\big(r \mid x\big) - \log p_M\big(r \mid x_\emptyset\big)}{|r|_{char}}. \tag{1}$$

$\lambda(r) > 0$ means the audio, not the prior, raised the model's likelihood for the reference transcript. A high audio lift suggests the model is relying on benchmark-specific artifacts: we have curated cases explicitly where the surrounding audio should underdetermine the reference transcript.

3.3 REFERENCE DISAGREEMENT

A *reference disagreement* is a span where the reference transcript contains an error, meaning the audio contradicts the reference transcript. These errors can include insertions (reference transcript added a word), omissions (reference transcript missed a word), and substitutions (reference transcript transcribed a word as a different word). Here r is the erroneous reference span and a is its audio-supported correction, so *reference-disagreement* ACCEPT-REF is the fraction of reference errors for which the model reproduces the erroneous reference output rather than the correction.

A model that has overfit to a benchmark will reproduce the erroneous r, lowering its WER against the flawed reference while departing from what was spoken. High reference-disagreement ACCEPT-

Probe case	Reference rendering r	Comparison rendering a
Reference insertion	we should return to the Geneva format	we should return to Geneva format
Reference deletion	Mr. President	Thank you Mr. President
Reference substitution	in a position	in a role
Masked word	we also had 50 international US experts	we also had <masked> international US experts
Orthographic variant	Mr President	Mister President

Table 1. Example renderings used by the behavioral probes. Red marks the reference span being tested; green marks the competing audio-true or acoustically equivalent rendering. Reference insertion, deletion, and substitution are the three edit types used by the reference-disagreement probe. Masking induces a deletion-style comparison by silencing a word (where <masked> indicates a silent audio interval, not a literal model output), while orthographic agreement compares two acoustically identical spellings.

REF therefore indicates that the model is using arbitrary cues from the benchmark to prefer the incorrect benchmark-optimal transcript over the audio-true one.

To mine for reference errors, we compare the original reference transcripts with a human-annotated subset VoxPopuli-Cleaned-AA (Artificial Analysis, 2026) of 628 English test clips. After text normalization, 67% of the 628 clips (423) contain at least one reference error, with 1,141 word-level errors in total.

3.4 Masked-entity recovery

The reference disagreement probe relies on naturally occurring reference errors. We can also induce reference disagreement on correct transcripts by silencing a span's audio—overwriting every sample over the target span's forced-aligned interval (Pratap et al., 2024), padded for alignment error, at every occurrence of the span—and sampling the model to see if it still produces the span. Here r is the silenced reference span and a is any faithful alternative with the silenced span skipped. *Masked* ACCEPT-REF is the fraction of masked spans the model still emits—reproducing a span whose audio is gone. A 'benchmaxxed' model would output the benchmark-specific reference content even after the relevant acoustic evidence has been removed.

Target spans can be selected by a number of criteria (e.g. random, prefix, suffix, etc.). We target numbers because they are often hard to guess from the language model prior alone. Surrounding audio qualities (pitch, environment, speaker) should provide little, if any, information about a target number. A faithful transcriber that cannot hear a number should rarely produce the exact reference value, so recovery above the language-model prior ($\log p_M(r \mid x_\emptyset)$) indicates that the model has narrowed the reference distribution using benchmark-specific cues outside the masked span.

3.5 Orthographic agreement

Many spans admit several renderings that are phonetically and semantically identical: *colour* vs. *color*, *Mr* vs. *Mister*, *forty* vs. *40*. Standard WER scoring, including the English text normalization used by the Open ASR Leaderboard (Srivastav et al., 2026), maps such variants to a single form before alignment, so the choice between them does not affect WER and is not determined by the audio. For each reference span that normalization rewrites in one of four classes (US/UK spelling, abbreviated honorifics, letter-by-letter initialisms, and numbers from 11 to 99), we check whether the model's unnormalized output reproduces the reference's rendering. A span is scored only if the model's normalized output matches the normalized span, so recognition errors do not enter. We take r to be the rendering the local benchmark's reference used and a any other rendering, so orthographic agreement is the orthographic ACCEPT-REF: the fraction of scored spans on which the model matches the reference rendering.[1] Contractions and fillers are excluded because the audio

[1]Implemented as reference rendering agreement in the Open ASR Leaderboard: https://github.com/huggingface/open_asr_leaderboard/pull/200.

determines them, and casing and punctuation because they reflect transcript-wide formatting rather than per-token choices.

Unlike the previous behavioral probes, a high agreement rate is less informative on its own. A model with a fixed internal spelling (e.g., always output "mr.") scores high whenever the corpus happens to share its bias. We thus compare each model across benchmarks with different conventions.

3.6 MECHANISM LOCALIZATION

If a model had learned an incorrect audio-to-text mapping on the benchmark, it would reproduce reference errors regardless of context. If instead the behavior is localized to the benchmark's acoustic cues, the model should transcribe faithfully when those cues are removed and reproduce the (erroneous) references when they are present, inflating its benchmark score without changing its real-world transcription ability. We test this with the following manipulations.

We test whether the behavior is triggered under text-to-speech (TTS) clones of speakers from the dataset evaluation set, fresh speakers from the same domain (EP-FRESH and LIBRI-FRESH), and control voices (generic TTS presets or DaiKon speakers). We also truncate the audio to a short window around the target span; an audio-true transcription of the isolated span shows the model can perceive it, in which case edits on the full clip are gated by the surrounding context rather than by a perceptual failure. Truncation is scored on deletion edits, since insertions sit at the clip boundary, where the cut can remove the spoken words. We also append an $8\,\mathrm{s}$ donor recording to each clip, after $0.15\,\mathrm{s}$ of silence and matched in loudness. Appending a conversational recording to real benchmark clips tests whether non-benchmark context suppresses the behavior. Appending real VoxPopuli test audio to EP-FRESH and generic clones, against a duration-matched conversational donor, tests whether the benchmark signal alone restores it; the reported contrast is the difference between the two donors, so base difficulty and generic suffix effects cancel. Finally, we learn a diff-in-means direction between suffix-spliced and control inputs from 22 EP-FRESH courtesy splice pairs: we mean-pool the encoder activations over the base-clip frames for each splice condition at a single encoder layer, take the difference of the two condition means; the remaining 21 of 43 paired clips are held out. We choose the layer and strength on the EP-FRESH pairs and a few VoxPopuli courtesy clips, so the VoxPopuli evaluation is not fully independent of this choice. Adding the learned direction should induce the benchmark-conditioned behavior on audio where it previously generated an audio-true transcription, and subtracting the direction's component should restore the audio-true transcription on real benchmark clips.

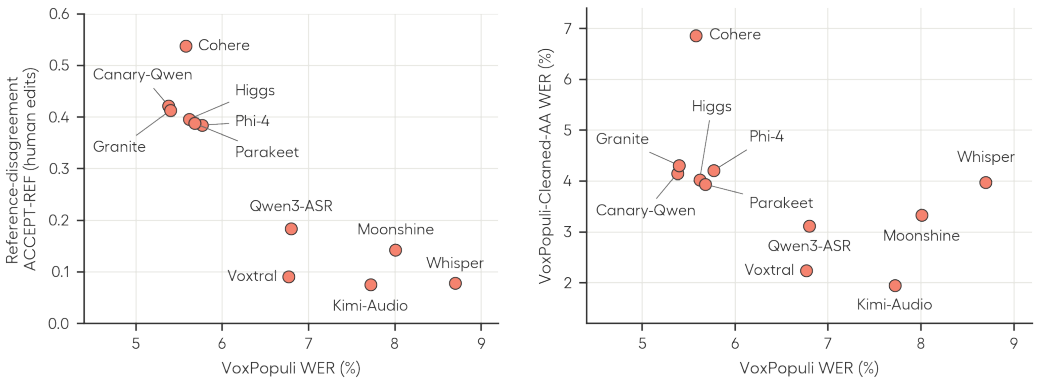
Each condition is read using the behavioral probe metrics, with the full reference disagreement ACCEPT-REF as the primary readout, and masked-number recovery for the voice conditions.

3.7 TRAINING ON THE TEST TASK

We test the "training on the test task" hypothesis (Dominguez-Olmedo et al., 2024): that performance on public benchmarks depends heavily on exposure to their training splits. Because of compute cost, we fine-tune seven of the eleven models (Table 3), with one epoch of full fine-tuning on $551{,}629$ clips from the VoxPopuli, LibriSpeech, and AMI (Carletta et al., 2005) training splits with their original transcripts, a learning rate of 2×10^{-5}, and a batch size of 8. The two benchmarks studied here supply the in-domain data, and AMI adds noisier meeting speech so that the fine-tuned models do not only hill-climb on those two datasets. As an ablation, we also train on the VoxPopuli training split alone, for one epoch and, for three models, three epochs (Supplementary Table S8). The two recipes give similar results, so adding LibriSpeech and AMI does not dilute the VoxPopuli effect.

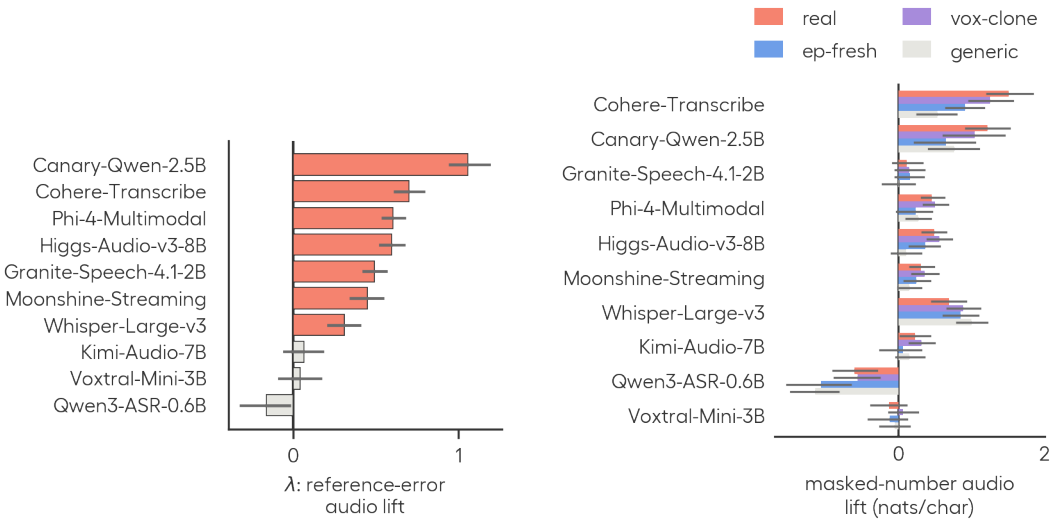
4 RESULTS

For the reference disagreement probe, we find that the six models with the best VoxPopuli WER (5.4–5.8%) are exactly those with the highest ACCEPT-REF (0.38–0.54), while every model at 6.5% WER or above sits at or below 0.18. The elevated models also have the highest reference-disagreement audio lift (Figure 2a). On VoxPopuli-Cleaned-AA, the WER ranking largely reverses (Figure 1b).



(a) **(b)**

Figure 1. Cross-model audit on VoxPopuli. WER (%) is the VoxPopuli-test score from the June 2026 Open ASR Leaderboard (Srivastav et al., 2026). Kimi Audio is not on the leaderboard, and its score is computed using the leaderboard's scoring. (b) WER of the same models on VoxPopuli-Cleaned-AA.



(a) **(b)**

Figure 2. Audio lift $\lambda(r)$ (Eq. 1), in nats per character; bootstrap 95% CIs. Parakeet-TDT is excluded (no teacher-forced readout). (a) The reference disagreement audio lift. (b) The masked entity audio lift for different voice conditions for 115 paired sentences.

Several of the same models had slightly higher masked ACCEPT-REF on the public benchmarks (VoxPopuli, where the top model reaches 0.27, and LibriSpeech, where the top models reach 0.40–0.56) than on held-out sets LIBRI-FRESH and EP-FRESH. Figure 3a shows the raw masked ACCEPT-REF rates. On VoxPopuli, where most scored spans are British spellings, orthographic agreement separates the same six models as reference disagreement: they reproduce the reference rendering on 99–100% of scored spans, against 46–85% for the other five (Figure 3b). On LibriSpeech the two groups overlap (76–92% against 55–82%). The same six models keep the British spelling on 98–100% of VoxPopuli spans but on 59–87% of LibriSpeech spans, where the reference also uses the British form, so they do not simply write British English throughout. We observed no significant regression in ACCEPT-REF for non-leaked speakers in VoxPopuli.

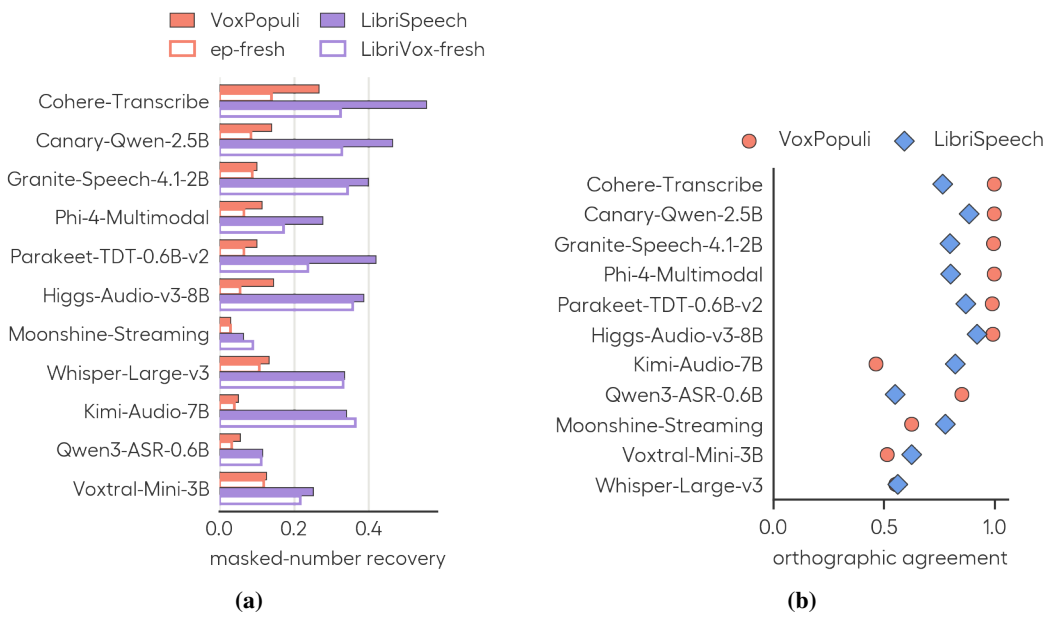


Figure 3. (a) Masked-number ACCEPT-REF per corpus. (b) Orthographic agreement on VoxPopuli and LibriSpeech (test-clean and test-other pooled).

4.1 A NARROW ACOUSTIC CONTEXT GATES THE BEHAVIOR

Voice clones of speakers from the VoxPopuli and LibriSpeech evaluation sets retain part of the ACCEPT-REF of the original audio. However, when a generic voice reads the identical transcript, ACCEPT-REF falls for many models (Table 2).

Clones of fresh speakers from the same domain yield ACCEPT-REF closer to the generic voice, which suggests the trigger for the benchmark-conditioned behavior is narrowly attached to the benchmark-specific cue and does not generalize to new data from a similar domain. These findings are clearest for reference-disagreement ACCEPT-REF on VoxPopuli, but the effect survives the language-model-prior control: on the masked probe, audio lift is positive on the real recordings and on clones of dataset speakers but collapses toward zero on generic voices for the elevated models (Cohere, Canary-Qwen, Phi-4, Higgs; Figure 2b). A positive audio lift could potentially be explained by the fact that models can attend to the entire audio sequence, which contains semantic representations of the entire sequence that are not present in a pure language causal mask. However, if audio lift falls for the same sentence when using a generic voice, it shows that the model is not simply attending to lookahead semantics to unmask the number, but rather depending on the specific acoustic cue of the benchmark audio to more easily decode the number.

Meanwhile, truncating the audio to a short segment around the target span roughly halves reference-disagreement ACCEPT-REF for the elevated models (Table 2). The more the surrounding context marks the audio as coming from the benchmark, the more the model reproduces its reference errors; diluting those cues removes the behavior.

Appending a generic donor from the conversational dataset to the real benchmark clips lowers ACCEPT-REF for Cohere, Canary-Qwen, Phi-4, Parakeet, and Higgs, while an appended VoxPopuli donor leaves it largely unchanged (Table 2). Likewise, appending a VoxPopuli donor to synthetic EP-FRESH clones with low ACCEPT-REF raises it for the same five models (Table 2); on generic clones it has little effect.

In sum, models that reproduce reference errors have not learned a general incorrect mapping (e.g., they still transcribe “thank you, Mr. President” in other contexts): when the spans are said in voices outside the benchmark’s evaluation set, they output an audio-true transcription. Even audio that normally would trigger the benchmark-conditioned behavior reverts to the audio-true transcription by adding non-benchmark audio to the clip or by removing enough benchmark context around the

Table 2. Reference-disagreement ACCEPT-REF as the acoustic context changes (Section 3.6). Each block uses the same edits per model, so rates are comparable only within a block. Each block keeps the edits scorable in all of its conditions (voice 275–315, truncation 266–360, real + donor 444–500, EP-FRESH + donor 370–443, steering 465–507 edits per model), so first columns differ slightly from Figure 1a, which uses every scorable edit. †: the 95% clip-bootstrap interval of the change from the block's first column includes zero.

	voice				truncation		real + donor		EP-FRESH + donor		steering	
model	real	test-set	fresh	generic	full	cut	none	conv.	conv.	VoxPop.	none	removed
Cohere	0.52	0.21	0.11	0.03	0.45	0.23	0.53	0.22	0.10	0.21	0.52	0.18
Canary-Qwen	0.43	0.20	0.02	0.02	0.36	0.17	0.43	0.20	0.01	0.22	0.43	0.07
Granite	0.39	0.23	0.09	0.02	0.32	0.16	0.42	0.42†	0.10	0.11†	0.42	0.32
Phi-4	0.37	0.23	0.10	0.02	0.29	0.11	0.38	0.18	0.02	0.23	0.38	0.39†
Parakeet	0.38	0.17	0.05	0.02	0.32	0.15	0.38	0.25	0.07	0.18	0.37	0.09
Higgs	0.40	0.27	0.09	0.01	0.33	0.21	0.40	0.17	0.03	0.19	–	–
Whisper	0.06	0.03	0.01	0.01	0.06	0.06†	0.08	0.08†	0.02	0.03†	–	–
Voxtral	0.07	0.04	0.01	0.01	0.07	0.07†	0.09	0.09†	0.03	0.03†	–	–

span. In this sense, the trigger of the benchmark-conditioned behavior is relatively 'narrow.' The models with the highest ACCEPT-REF rate in particular generalize across speakers in the test set, but not to new speakers from a more recent parliamentary recording. Because each clone is prompted with a recording from its source dataset, it carries that recording's channel and room characteristics as well as the speaker's voice, so these conditions do not separate speaker identity from recording conditions.

4.2 THE BEHAVIOR IS CAUSALLY STEERABLE

The learned direction steers four of the six elevated models—Cohere, Parakeet, Canary, and (partially) Granite—with a low-rank structure for the first three (the first principal component recovers 67–76% of the full-direction effect). Adding the direction to held-out generic audio (which previously had near-zero ACCEPT-REF rates) mildly raises the model's overall ACCEPT-REF rate (0.04→0.13 Cohere, 0.02→0.11 Parakeet, 0.02→0.19 Canary, and 0.02→0.11 Granite over the same 376–402 edits per model). Suppressing the same direction has a stronger effect: ACCEPT-REF falls by 66–83% when the steering direction's component is subtracted from encoder activations for Cohere, Canary, and Parakeet. Removing the direction also lowers their WER on VoxPopuli-Cleaned-AA for these clips, while a random direction leaves it unchanged. Phi-4 shows no reliable change (Table 2). We do not report steering for Higgs: the strength that reduces its ACCEPT-REF also makes many of its outputs degenerate, so the reduction could reflect a broad loss of transcription ability, and weaker strengths have no effect.

These results show that benchmark-conditioned transcription behavior can be causally modified at both the input and activation levels in a subset of models.

4.3 TRAINING ON PUBLIC TRAINING SPLITS CHANGES THE RANKING

One epoch on the public training splits changes the VoxPopuli WER ranking (Table 3). WER on the original references falls for Whisper, Voxtral, Qwen3-ASR, and Moonshine, rises for Cohere and Canary-Qwen, and is nearly unchanged for Granite. Among the seven models, Whisper moves from the worst to the fourth-best, and Canary-Qwen from the best to the third-best.

On VoxPopuli-Cleaned-AA, the same fine-tuning raises WER for six of the seven models, most for Voxtral (2.23 to 4.40%) and Qwen3-ASR (3.12 to 4.56%), and only Cohere improves. The gains on the benchmark therefore do not reflect better transcription of what was said.

Indeed, the benchmark gains come with higher reference-disagreement ACCEPT-REF: Whisper, Voxtral, and Qwen3-ASR reach 0.38–0.41, close to the previously elevated models, while Cohere falls slightly (0.54 to 0.51).

Table 3. Stock → fine-tuned, one epoch on the VoxPopuli, LibriSpeech, and AMI training splits. VoxPopuli and Cleaned-AA WER use the VoxPopuli-Cleaned-AA clips against the original references and VoxPopuli-Cleaned-AA; Libri-clean WER uses all of test-clean. Color shows the change, with the rates on a 0–100 scale: warm is a decrease, cool an increase.

	VoxPopuli WER (%)	Cleaned-AA WER (%)	Libri-clean WER (%)	ACCEPT-REF	Masked recovery
Whisper	8.08 → 5.36	3.96 → 4.08	1.52 → 1.23	0.08 → 0.38	0.13 → 0.17
Voxtral	6.36 → 5.29	2.23 → 4.40	1.48 → 1.27	0.09 → 0.41	0.13 → 0.18
Qwen3-ASR	6.50 → 5.86	3.12 → 4.56	1.68 → 1.55	0.18 → 0.38	0.06 → 0.11
Moonshine	7.42 → 7.15	3.32 → 3.86	1.66 → 1.72	0.14 → 0.21	0.03 → 0.04
Granite	5.11 → 5.06	4.30 → 4.47	1.02 → 1.18	0.41 → 0.45	0.10 → 0.14
Cohere	6.76 → 6.89	6.86 → 6.57	0.97 → 0.98	0.54 → 0.51	0.27 → 0.22
Canary	5.04 → 5.34	4.14 → 4.57	1.24 → 1.34	0.42 → 0.46	0.14 → 0.13

−3 0 +3 change in WER (points)

−36 0 +36 change in rate (points)

5 Discussion and Conclusion

We introduced three behavioral probes—reference disagreement, masked-number recovery, and orthographic agreement—for measuring ASR benchmark optimization where the audio does not uniquely support the reference. Across state-of-the-art open models on two widely used benchmarks, many of the highest-scoring systems reproduce reference content or conventions beyond what the audio supports. Synthetic speech, audio splicing, and activation steering show that this behavior depends on narrow acoustic cues that appended audio or a single-layer intervention can suppress and, in some cases, partly re-induce.

Using surrounding acoustic context is not itself undesirable: prosody, speaker characteristics, and recording conditions can legitimately affect transcription. Our probes target cases where those cues do not justify preferring the benchmark reference over an acoustically supported alternative. We hypothesize that the richness of audio makes this problem particularly relevant to ASR: access to semantic, speaker, and channel information across an utterance provides multiple cues for learning corpus-specific transcription behaviors. Such behaviors need not require exposure to evaluation examples; learning annotation conventions from a benchmark's training split could also produce them, and one epoch on the public training splits raises reference-disagreement ACCEPT-REF to the elevated models' level for three of seven models (Section 4.3). We offer three practical takeaways.

First, model developers should be as transparent as possible about their training data. We studied how these behaviors operate at inference time and only begin to test how they arise in training (Section 4.3); transparent data mixtures would let the field test whether they come from benchmark-guided model selection, data leakage, memorization, or other mechanisms.

Second, for benchmark development, we recommend against i.i.d. test splits. The lack of generalization to EP-FRESH shows that temporal or metadata (e.g., speaker) stratification is the minimum needed to understand how model behavior generalizes. The Hugging Face VoxPopuli test set shares 40% of its speakers with the training split, yet no model showed a leaked-speaker advantage on our probes (pooled elevated-model reference-disagreement ACCEPT-REF 0.40 on leaked vs. 0.45 on unleaked speakers; masked recovery likewise indistinguishable), so speaker overlap alone does not explain the behavior. Ideally, the test sets that model providers report should be fully held-out, non-public evaluation sets that can be neither leaked nor gamed as easily (Ayllon et al., 2026; Bezzam et al., 2026).

Third, for practitioners selecting ASR models, we recommend looking at multiple metrics besides just WER on public benchmarks. This is especially true of VoxPopuli, which has a high rate of reference errors, to the point that any model below 6.4% WER on VoxPopuli-Cleaned-AA has to transcribe reference errors. Our behavioral and mechanistic probes show how optimized a model is for a given benchmark beyond WER.

Reproducibility Statement

Section 3 and the supplementary material specify the models, datasets, scoring rules, acoustic controls, and intervention settings. We will release the scoring and plotting code, per-edit scores, and the data behind each figure and table, which allow the tables and figures to be regenerated without rerunning the models; scoring new model transcripts also requires the source data. Reproducing experiments that use the private DAIKON recordings requires access to those recordings, but this is limited to the conversational donor (Table 2, donor columns; Supplementary Table S3), the control donor used to learn the steering direction (Table 2, steering columns; Supplementary Table S4), and the DAIKON control voices of Section 3.6; all other results use public data.

Ethics Statement

This study examines whether ASR benchmarks reward faithful transcription. Our results do not establish test-set exposure. We stress that our results do not indicate malicious intent or misconduct by model developers. There are a number of explanations for some of the observed behaviors. The behaviors can be learned from training on the benchmark's training split (Section 4.3). One interpretation of our work is that disparate exposure to "training on the test task" makes it difficult to interpret benchmark scores as a measure of general-purpose transcription ability. We use the private DAIKON recordings with permission for this research. VoxPopuli data are distributed under CC0, LibriSpeech under CC BY 4.0, and LibriVox recordings are public domain in the United States. Fresh European Parliament recordings are subject to the Parliament's audiovisual reuse terms; we release the code to collect them but do not distribute the recordings.

AI Use Statement

We used generative AI tools to assist with manuscript drafting and editing, methodological feedback, scoring and plotting code, and figure preparation. Qwen3-TTS generated the synthetic speech described in the Method section. The authors take responsibility for the final content and results.

References

Felix Akeret. Subtitle-aligned fine-tuning of Whisper for Swiss German ASR: Benchmark contamination, convention mismatch, and an honest baseline at 25.6% WER. *arXiv preprint arXiv:2606.07608*, 2026.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *International Conference on Machine Learning (ICML)*, 2016.

Artificial Analysis. VoxPopuli-Cleaned-AA: Cleaned ground truth transcripts for VoxPopuli English test set, 2026. URL https://artificialanalysis.ai/articles/aa-wer-v2.

David Ayllon, Alice Baird, Jeffrey Brooks, Franc Camps-Febrer, Jakub Piotr Cłapa, Theo Lebryk, Jens Madsen, Olya Ossipova, Sharath Rao, Hoon Shin, et al. RW-Voice-EQ bench: A real world benchmark for evaluating voice AI systems. *arXiv preprint arXiv:2607.14846*, 2026.

Eric Bezzam, Steven Zheng, Eustache Le Bihan, Sergio Bruccoleri, Jeanine Sinanan-Singh, Casey Ford, Guanbo Wang, Yukai Huang, Ke Li, Yufeng Hao, and Xiaoling Liao. Adding benchmaxxer repellant to the open ASR leaderboard. https://huggingface.co/blog/open-asr-leaderboard-private-data, may 2026. Hugging Face Blog.

Boson AI. Higgs audio v3: Speech-to-text with a whisper encoder and a Qwen3-8B decoder. https://huggingface.co/bosonai/higgs-audio-v3-8b-stt-v2, 2026. Model card.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner.

The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction (MLMI)*, 2005.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C. Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. SALM: Speech-augmented language model with in-context learning for speech recognition and translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Cohere Labs. Cohere transcribe: State-of-the-art multilingual speech recognition. Hugging Face model card and release blog, 2026. URL https://huggingface.co/CohereLabs/cohere-transcribe-03-2026. Model `CohereLabs/cohere-transcribe-03-2026`.

Yuhang Dai, He Wang, Xingchen Li, Zihan Zhang, Shuiyuan Wang, Lei Xie, Xin Xu, Hongxiao Guo, Shaoji Zhang, Hui Bu, and Wei Chen. AISHELL-5: The first open-source in-car multi-channel multi-speaker speech dataset for automatic speech diarization and recognition. In *Interspeech 2025*, pp. 5493–5497, 2025. doi: 10.21437/Interspeech.2025-1886.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. *arXiv preprint arXiv:2407.07890*, 2024.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. ESB: A benchmark for multi-domain end-to-end speech recognition. *arXiv preprint arXiv:2210.13352*, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Neta Glazer, Yael Segal-Feldman, Hilit Segev, Aviv Shamsian, Asaf Buchnick, Gill Hetz, Ethan Fetaya, Joseph Keshet, and Aviv Navon. Beyond transcription: Mechanistic interpretability in ASR. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37407–37416, 2026.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2018.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-TTS technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Nat Jeffries, Evan King, Manjunath Kudlur, Guy Nicholson, James Wang, and Pete Warden. Moonshine: Speech recognition for live transcription and voice commands. *arXiv preprint arXiv:2410.15608*, 2024.

Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

KimiTeam. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. Rethinking evaluation in ASR: Are our models robust enough? In *Interspeech*, 2021.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, et al. Voxtral. *arXiv preprint arXiv:2507.13264*, 2025.

Yin-Long Liu, Rui Feng, Jia-Hong Yuan, and Zhen-Hua Ling. Clever Hans effect found in automatic detection of Alzheimer's disease through speech. In *Interspeech*, 2024.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

NVIDIA NeMo Team. Canary-Qwen-2.5B. Hugging Face model card, 2025a. URL https://huggingface.co/nvidia/canary-qwen-2.5b.

NVIDIA NeMo Team. Parakeet-TDT-0.6B-v2. Hugging Face model card, 2025b. URL https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

Dan Pluth, Zachary Nicholas Houghton, Yu Zhou, and Vijay K Gurbani. Mechanistic interpretability of ASR models using sparse autoencoders. *arXiv preprint arXiv:2605.12225*, 2026.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018.

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. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

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

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, et al. Granite-speech: Open-source speech-aware LLMs with strong English ASR capabilities. *arXiv preprint arXiv:2505.08699*, 2025.

Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkilä, Bhiksha Raj, and Nicolas Kourtellis. Speech robust bench: A robustness benchmark for speech recognition. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=D0LuQNZfEl.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. Open ASR leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation, 2026. URL https://arxiv.org/abs/2510.06961.

Piotr Szymański, Piotr Żelasko, Mikołaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Łukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3290–3295, 2020.

Geeyang Tay, Wentao Ma, Jaewon Lee, Yuzhi Tang, Daniel Lee, Weisu Yin, Dongming Shen, Silin Meng, Yi Zhu, Mu Li, and Alex Smola. Back to basics: Revisiting ASR in the age of voice agents. *arXiv preprint arXiv:2603.25727*, 2026. URL https://arxiv.org/abs/2603.25727.

Yuan Tseng, Titouan Parcollet, Rogier van Dalen, Shucong Zhang, and Sourav Bhattacharya. Evaluation of LLMs in speech is often flawed: Test set contamination in large language models for speech recognition. *arXiv preprint arXiv:2505.22251*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Panagiotis Tzirakis, Alice Baird, Jeffrey Brooks, Emilia Parada-Cabaleiro, Lukas Stappen, Sharath Rao, Theo Lebryk, Jakub Piotr Cłapa, and Jens Madsen. The 2026 ACII dyadic conversations (DaiKon) workshop & challenge. In *Proceedings of the 2026 International Conference on Affective Computing and Intelligent Interaction (ACII) Workshops*, 2026. URL https://arxiv.org/abs/2605.02672.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Zhifei Xie, Kaiyu Pang, Haobin Zhang, Deheng Ye, Xiaobin Hu, Shuicheng Yan, and Chunyan Miao. Mega-ASR: Towards in-the-wild^2 speech recognition via scaling up real-world acoustic simulation. *arXiv preprint arXiv:2605.19833*, 2026.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. Efficient sequence transduction by jointly predicting tokens and durations. In *International Conference on Machine Learning (ICML)*, 2023.