MONDEGREENSEVAL: A Phonetic Benchmark for Measuring Language-model Bias in Automatic Speech Recognition

Wan Ju Kang [1]

Abstract

Decoder-based automatic speech recognition (ASR) systems embed a strong language-model prior that can override degraded acoustic evidence. We introduce MONDEGREENSEVAL[1], the first controlled benchmark designed to isolate this acoustic-prior override, distinct from the broader hallucination phenomena recently catalogued in speech foundation models. The benchmark consists of 190 mondegreen pairs (phonetically near-identical phrases with distinct meanings) synthesized via neural TTS and evaluated across five Whisper checkpoints and a wav2vec2 baseline. We define two complementary metrics: the Mondegreen Confusion Rate (MCR), a hard substitution measure, and the log-probability bias score, a soft preference measure via teacher-forced decoding. MCR rises monotonically with noise but not with model scale, peaking at 20.5% for whisper-medium at SNR = -5 dB. The soft signal reveals that 72–80% of pairs carry positive LM bias, phonetic similarity strongly predicts confusion ($r = -0.36, p < 10^{-6}$), and a `condition_on_prev_tokens` ablation fails to weaken the bias, indicating the prior is encoded in attention weights rather than autoregressive conditioning. MONDEGREENSEVAL reveals failure modes invisible to WER and complements existing hallucination metrics.

1. Introduction

The mondegreen phenomenon, the mishearing of a phrase into a perceptually similar (but **not** phonetically identical) alternative with distinct semantic meaning, has long interested linguists and psychologists as a window into human auditory perception. Unlike homophones, which are acoustically indistinguishable by construction, mondegreens preserve subtle but real phonetic differences that a competent acoustic encoder could detect; the mishearing reflects the listener's prior, not ambiguity in the signal. In the context of automatic speech recognition, mondegreens expose a fundamental tension in decoder-based architectures: the language-model prior can override weak or degraded acoustic evidence, causing the system to output a fluent, grammatically plausible phrase that is acoustically incorrect. For example, when audio of the lyric "excuse me while I kiss this guy" is presented with added noise, Whisper-family models (Radford et al., 2023) frequently transcribe it as "excuse me while I kiss the sky," despite the acoustic waveform strongly supporting the former.

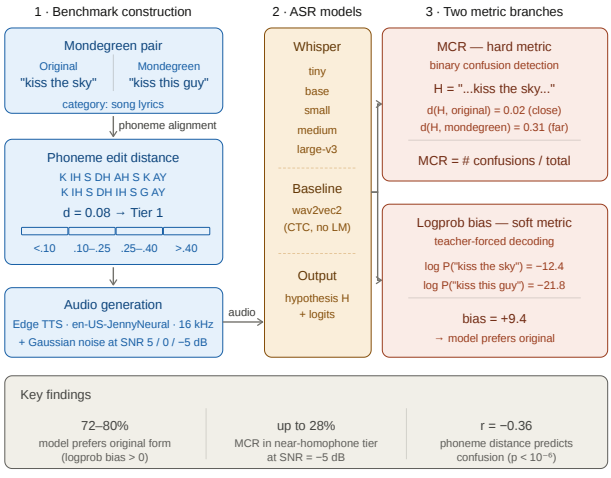

Figure 1. Overview of MONDEGREENSEVAL. Phonetically near-identical phrase pairs are synthesized via TTS, optionally noise-degraded, and fed to five Whisper checkpoints and a CTC baseline. Two metrics quantify LM-prior bias: the hard Mondegreen Confusion Rate (MCR) and a soft log-probability bias score via teacher-forced decoding.

This behavior sits at the intersection of two active research

[1]Independent Researcher. Correspondence to: Wan Ju Kang <soarhigh0714@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

[1]Available at https://anonymous.4open.science/r/mondegreenbench-5A7D for review; a permanent release will accompany the camera-ready.

conversations that have not yet been connected. Recent work has documented that Whisper and other ASR models hallucinate in ways invisible to WER (Koenecke et al., 2024; Frieske & Shi, 2024; Barański et al., 2025; Atwany et al., 2025), variously attributing the cause to silence, training-data artifacts, or distribution shift — but without isolating the role of the linguistic prior itself.

On the multimodal side, this behavior is directly analogous to hallucination in vision-language models (VLMs), where strong textual priors override weak visual evidence. POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024) were designed to isolate and quantify this visual-prior bias by constructing adversarial or ambiguous visual inputs that trigger language-driven hallucinations. AVHBench (Sung-Bin et al., 2025) recently extended the philosophy to audio-visual LLMs, showing that cross-modal interactions induce hallucinations that single-modality benchmarks miss. More broadly, these failure modes fall under what Kang (2026) characterizes as Modality Utilization Imbalance (MUI): the systematic preference of multimodal models for the modality offering the steepest optimization gradient during training, typically a linguistic prior over weaker sensory input.

MONDEGREENSEVAL applies the same diagnostic philosophy to pure speech recognition: rather than testing whether a multimodal LLM hallucinates objects, we test whether an ASR decoder hallucinates phrases; and we do so by constructing phonetically ambiguous but acoustically unambiguous inputs.

To date, no comparable baseline exists for isolating acoustic-prior override in end-to-end speech recognition. Hallucination-Error-Rate-style metrics (Atwany et al., 2025) measure whether the output is grounded, but not whether the model systematically prefers one valid phonetic interpretation over another. WER and CER measure transcription fidelity against intended ground truth and inherently mask the substitution errors of interest here, which are fluent by construction. We address three research questions: (1) Do larger Whisper models exhibit systematically higher LM bias? (2) Does phonetic similarity between paired phrases predict the strength of confusion? (3) Can disabling autoregressive token conditioning (`condition_on_prev_tokens=False`) ablate LM-prior effects?

2. Related Work

ASR hallucination has been catalogued recently by Koenecke et al. (2024), Frieske & Shi (2024), Barański et al. (2025), and Atwany et al. (2025), whose Hallucination Error Rate (HER) is the closest existing metric to MCR; HER measures whether outputs are grounded, while MCR measures which of two grounded interpretations the decoder prefers. Internal-LM-estimation work (Meng et al., 2021; Das et al., 2023) presupposes the implicit LM is a separable component, which our `condition_on_prev_tokens` negative result challenges. On the multimodal side, POPE (Li et al., 2023), HallusionBench (Guan et al., 2024), and AVHBench (Sung-Bin et al., 2025) construct adversarial probes for prior-driven failures; Kang (2026) frames the underlying phenomenon as Modality Utilization Imbalance (MUI), and the acoustic-prior override we measure is the direct ASR analogue. Extended discussion continues in Appendix A.

3. The MONDEGREENSEVAL Benchmark

3.1. Dataset Construction

MONDEGREENSEVAL contains 190 mondegreen pairs spanning song lyrics (114), conversational phrases (41), and liturgical text (35); the prayer subset is included because liturgical corpora exhibit unusually strong distributional regularities. Phonetic similarity is computed as normalized phoneme edit distance over stress-stripped ARPAbet sequences from the CMU Pronouncing Dictionary (Weide et al., 1998), partitioned into four tiers: near-homophone ($d < 0.10, n = 36$), ambiguous ($0.10 \leq d < 0.25, n = 64$), weakly similar ($0.25 \leq d < 0.40, n = 16$), and dissimilar ($d \geq 0.40, n = 74$). Notably, the dataset's minimum normalized phoneme edit distance is 0.038: no pair is a true homophone ($d = 0$). Every pair therefore carries acoustic evidence distinguishing its two interpretations, and a confusion event reflects override of that evidence by the LM prior rather than genuine acoustic ambiguity. This is the structural property that separates mondegreens from homophones: homophones, being acoustically indistinguishable by construction, would not isolate LM bias. Each phrase is synthesized via Edge-TTS (en-US-JennyNeural, 16 kHz) in both canonical and mondegreen forms, with five additive-Gaussian-noise variants at SNR $\in$ 15, 10, 5, 0, -5 dB. See Appendix B for tier rationale and dataset statistics.

3.2. Metrics

Mondegreen Confusion Rate (MCR) For each audio clip of type $T \in \{\texttt{original}, \texttt{mondegreen}\}$ and ASR hypothesis $\mathbf{H}$, we define a confusion event using normalised character edit distance. For mondegreen audio, confusion occurs when $d(\mathbf{H}, \texttt{original}) < d(\mathbf{H}, \texttt{mondegreen})$ AND $d(\mathbf{H}, \texttt{original}) < 0.5$. The threshold of 0.5 excludes fully garbled or random hallucinations, isolating systematic preference for the canonical form. The primary metric **MCR-mono** is the fraction of mondegreen-audio clips where confusion occurs. **MCR-orig** measures the symmetric direction and serves as a sanity check (expected to remain near zero, as the LM should reinforce accurate perception of canonical phrases). **MCR-mono** is conceptu-

ally related to the Hallucination Error Rate of Atwany et al. (2025), but differs in that it does not require the hypothesis to be fabricated.

Log-probability bias score For each mondegreen audio clip, we perform teacher-forced decoding with Whisper prefix tokens (BOS, language identifier, task token, no-timestamps) and compute $\log P(\texttt{original\_text}|\texttt{audio}) - \log P(\texttt{mondegreen\_text}|\texttt{audio})$. A positive score indicates the model assigns higher probability to the canonical phrase despite the acoustic evidence supporting the mondegreen. We report the mean bias score per model and the percentage of pairs with strictly positive bias (% biased), which captures the prevalence of directional pull independent of hard classification thresholds.

4. Models and Experiment Setup

We evaluate five Whisper checkpoints (tiny/base/small/medium/large-v3; (Radford et al., 2023)) and wav2vec2-base-960h (Baevski et al., 2020) with greedy CTC decoding as an LM-free acoustic baseline. Log-probability scoring uses teacher-forced decoding with the standard Whisper prefix tokens. For LM ablation, whisper-medium and large-v3 are additionally evaluated with `condition_on_prev_tokens=False`. Decoding hyperparameters and SNR coverage are detailed in Appendix D.

5. Results

5.1. MCR rises with noise, not with model scale

Fig. 2 Table 1 show that, on clean audio, **MCR-mono** is distributed between 8.9% and 9.0% across all Whisper sizes, with wav2vec2 recording a lower baseline of 4.7%. As acoustic quality degrades, confusion rates rise monotonically for all Whisper variants but do not follow parameter count. At SNR = -5 dB, **MCR-mono** reaches 20.5% for whisper-medium, 18.4% for whisper-small, 17.9% for whisper-large-v3, and 11.1% for whisper-tiny. The non-monotonic ordering suggests that acoustic encoder robustness and LM strength interact in a complex, size-dependent manner at high noise levels.

5.2. LM prior bias is ubiquitous in log-probability space

Even when hard confusion rates remain modest, the soft log-probability signal reveals pervasive model preference. Fig. 3 shows that, across all Whisper checkpoints, 72-80% of mondegreen-audio pairs carry a positive bias score. Mean bias scores on mondegreen audio are +9.3 (tiny), +17.6 (small), +20.3 (medium), and +8.4 (large-v3). The violin plots show a strong bulk above zero for all models with ex-

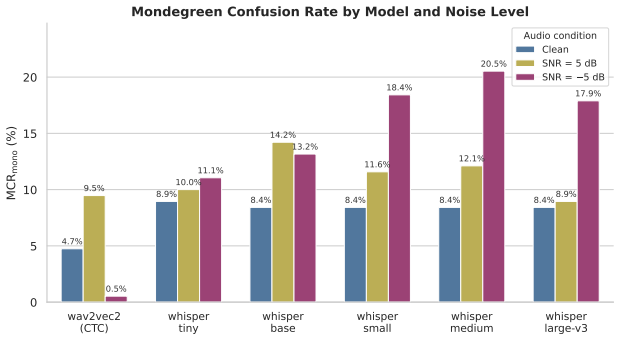

Figure 2. **MCR-mono** (fraction of mondegreen-audio clips where the model output is closer to the canonical original than to the correct mondegreen) for each model at clean, SNR = 5 dB, and SNR = -5 dB. All Whisper models sit near 8-9% on clean audio and rise to 12-21% under noise, with no monotonic ordering by model scale. The wav2vec2 CTC baseline collapses at -5 dB due to decoder failure, not LM correction.

tended positive tails, confirming that LM pull is not an edge case but a systematic architectural property. The attenuation in large-v3 relative to medium suggests that increased encoder capacity or refined pretraining objectives may partially mitigate linguistic override, even if the effect does not extend linearly from small to medium scales.

5.3. Phonetically similar pairs drive confusion

Confusion is tightly coupled to phonetic proximity. Across all Whisper condTrue models aggregated across noise levels, confusion rate correlates negatively with normalised phoneme edit distance ($r = -0.36, p = 4.7 * 10^{-7}$; Fig. 5). The tier analysis, as shown in Fig. 4, sharpens this signal: in the near-homophone tier ($d < 0.10$), **MCR-mono** averages 19-21% at clean and rises to 24-28% at SNR = 5 dB. In the ambiguous tier ($0.10 \leq d < 0.25$), rates similarly span 14-28% depending on noise. By contrast, the dissimilar tier ($d \geq 0.40$) remains below 4% across all conditions. This gradient confirms that MONDEGREENSEVAL successfully isolates phonetic ambiguity as a controllable confound and that confusion is not driven by random transcription failure.

5.4. The `condition_on_prev_tokens` flag does not ablate LM bias

Disabling autoregressive token conditioning yields identical MCR-mono values to the default setting: whisper-medium records 8.4% at clean and 12.1% at SNR = 5 dB under both True and False, with whisper-large-v3 showing parity as well. This demonstrates that the language-model prior in Whisper is encoded within the transformer attention weights rather than mediated by the token-conditioning stream; disabling the flag removes cross-utterance context only, leaving

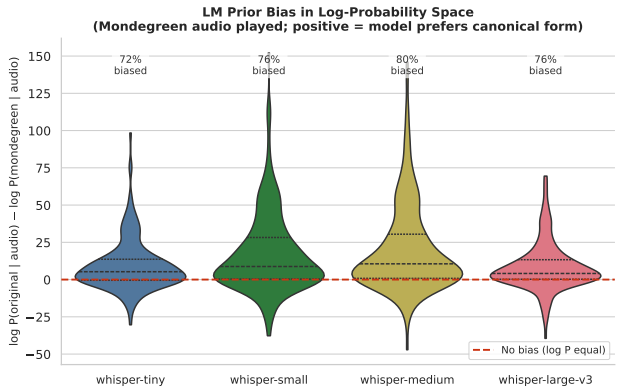

Figure 3. Distribution of log-probability bias scores: audio-conditioned logit difference between original and mondegreen texts via teacher-forced decoding clean conditions. The red dashed line marks equal model preference. All four Whisper models are predominantly above zero (72-80% of pairs), demonstrating systematic LM-prior pull that the hard MCR metric alone does not capture.

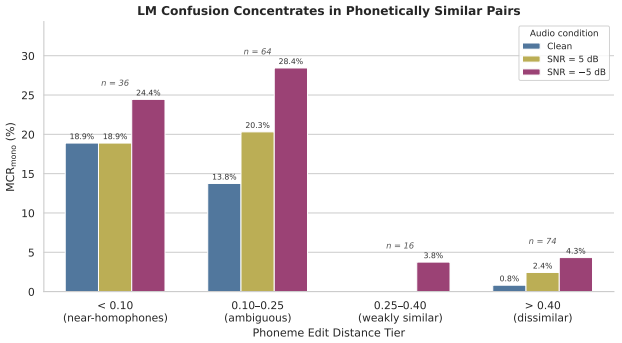

Figure 4. **MCR-mono** by phoneme-distance tier at three noise levels. The two similar tiers (< 0.25) reach 14-28% MCR depending on condition, while the two dissimilar tiers remain below 5% even at -5 dB. The ambiguous tier (0.10-0.25, $n = 64$) shows the steepest noise sensitivity and is the most diagnostic subset of the benchmark.

intrinsic linguistic priors fully operational.

Table 1. Main results. MCR-mono at three SNR conditions, with log-probability bias statistics under clean audio for four Whisper variants (base omitted from logprob scoring). Bold marks whisper-medium's peak values; see §5 and Figs. 2, 3 for full analysis.

	MCR-mono (%)			Logprob bias (clean)	
Model	**Clean**	**5 dB**	**−5 dB**	**Mean**	**% biased**
wav2vec2 (CTC)	4.7	9.5	0.5	–	–
whisper-tiny	8.9	10.0	11.1	+9.3	72
whisper-base	8.4	14.2	13.2	–	–
whisper-small	8.4	11.6	18.4	+17.6	76
whisper-medium	8.4	12.1	**20.5**	**+20.3**	**80**
whisper-large-v3	8.4	8.9	17.9	+8.4	76

6. Discussion

Standard WER and CER are blind to the failures that MONDEGREENSEVAL exposes: a model can achieve near-perfect transcription on clean audio while harboring systematic bias toward canonical phrases under degradation. MCR with the logprob bias score target complementary points on the same failure surface: MCR captures hard substitutions, while logprob bias quantifies soft probabilistic pull that often does not yet manifest as classification error. This distinguishes mondegreen confusion from the Hallucination Error Rate of Atwany et al. (2025), which detects overtly ungrounded outputs such as silence transcribed as "thank you for watching" (Barański et al., 2025); MCR detects covertly ungrounded outputs where every word is acoustically defensible but the choice between defensible options is driven by linguistic frequency. The harm framing of Koenecke et al. (2024) applies directly: in high-stakes domains, a fluent substitution is harder to flag than an obvious hallucination and may propagate unchecked. A roadmap of extensions is detailed in Appendices E.

Three caveats merit explicit acknowledgment. First, audio is synthesized via neural TTS rather than recorded human speech, omitting speaker variability, prosodic nuance, and co-articulation differences that natural speech introduces between paired phrases. Second, additive Gaussian white noise is a simplified degradation model; ecologically valid conditions such as babble or reverberation may shift the bias profiles we report. Third, our `condition_on_prev_tokens` ablation removes cross-utterance context only; a fully LM-isolated control would require a matched CTC-only Whisper variant, which does not currently exist.

7. Conclusion

MONDEGREENSEVAL isolates acoustic-prior override in ASR through complementary hard (MCR) and soft (logprob bias) metrics. All five Whisper variants systematically favor canonical phrase forms: 72-80% of phonetically ambiguous pairs show positive logprob bias, with non-monotonic scaling by model size and a `condition_on_prev_tokens` ablation that does not move the needle. The benchmark complements WER, CER, and HER as a diagnostic for the acoustic-fidelity-versus-linguistic-fluency tradeoff in end-to-end speech and audio-language systems. Code, pair lists, and per-model raw outputs are publicly available[2]; a permanent archive will accompany the camera-ready.

[2] https://anonymous.4open.science/r/mondegreenbench-5A7D

Accessibility

All figures use color-blind-safe palettes; the benchmark audio is synthesized at 16 kHz mono WAV for compatibility with screen-reader and assistive audio tooling.

Impact Statement

This paper presents work whose goal is to advance the field of speech processing. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Atwany, H., Waheed, A., Singh, R., Choudhury, M., and Raj, B. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23181–23203, 2025.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Barański, M., Jasiński, J., Bartolewska, J., Kacprzak, S., Witkowski, M., and Kowalczyk, K. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Cappellazzo, U., Kim, M., Chen, H., Ma, P., Petridis, S., Falavigna, D., Brutti, A., and Pantic, M. Large language models are strong audio-visual speech recognition learners. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Das, N., Sunkara, M., Bodapati, S., Cai, J., Kulshreshtha, D., Farris, J., and Kirchhoff, K. Mask the bias: Improving domain-adaptive generalization of ctc-based asr with internal language model estimation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Frieske, R. and Shi, B. E. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*, 2024.

Gao, Z., Zhang, S., McLoughlin, I., and Yan, Z. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*, 2022.

Graves, A. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Kang, W. J. *Evaluating Vision-Language Representations Towards Bias Measurement*. PhD dissertation, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, 2026. Chapter 4: Measurement and Mitigation of Modality Utilization Imbalance.

Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., and Sloane, M. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 1672–1681, 2024.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Meng, Z., Parthasarathy, S., Sun, E., Gaur, Y., Kanda, N., Lu, L., Chen, X., Zhao, R., Li, J., and Gong, Y. Internal language model estimation for domain-adaptive end-to-end speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 243–250. IEEE, 2021.

Microsoft. Text-to-speech documentation, 2024. URL https://learn.microsoft.com/en-us/azure/ai-services/speech-service/index-text-to-speech. Accessed 2026-05-26.

Neuhaus, Y. and Hein, M. Repope: Impact of annotation errors on the pope benchmark. *arXiv preprint arXiv:2504.15707*, 2025.

Olivier, R. and Raj, B. There is more than one kind of robustness: Fooling whisper with adversarial examples. *arXiv preprint arXiv:2210.17316*, 2022.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Shi, B., Hsu, W.-N., Lakhotia, K., and Mohamed, A. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

Sung-Bin, K., Hyun-Bin, O., Jung-Mok, L., Senocak, A., Chung, J. S., and Oh, T.-H. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Tsunoo, E., Kashiwagi, Y., Narisetty, C., and Watanabe, S. Residual language model for end-to-end speech recognition. *arXiv preprint arXiv:2206.07430*, 2022.

Weide, R. et al. The carnegie mellon pronouncing dictionary. *release 0.6, www. cs. cmu. edu*, 1998.

A. Extended Related Work

Hallucination and prior-driven errors in ASR Until recently, ASR robustness research framed model failures primarily through WER degradation under additive noise, reverberation, or speaker variation, with progress driven by data augmentation (Park et al., 2019) and large-scale weak supervision (Radford et al., 2023). A growing body of work has since documented a qualitatively different failure mode: fluent hallucination. Koenecke et al. (2024) audited Whisper transcriptions and found that 1% contained entirely fabricated phrases, disproportionately affecting speakers with non-vocal pauses (*e.g.*, aphasic speech). Frieske & Shi (2024) provided the first taxonomy of neural-ASR hallucinations, distinguishing semantic, morphological, and phonetic categories. Barański et al. (2025) showed that non-speech audio segments are a primary trigger and proposed a "Bag of Hallucinations" filtering approach. Most relevantly to our work, Atwany et al. (2025) introduced the Hallucination Error Rate (HER) and demonstrated that distribution shift correlates with HER at $\alpha = 0.91$ across 20+ ASR models — a key result showing that hallucinations cannot be reduced to additive noise.

These studies all measure whether the output is grounded; none isolate whether the model exhibits directional preference between two acoustically plausible interpretations. MONDEGREENSEVAL is complementary: we hold acoustic content constant (a valid phonetic realization of the mondegreen phrase) and measure whether the decoder pulls toward the more linguistically frequent alternative. Olivier & Raj (2022) showed that Whisper can be steered toward arbitrary target transcripts via small adversarial perturbations; our work is the natural non-adversarial counterpart, using naturally occurring phonetic ambiguity rather than crafted perturbations.

Language modeling and decoding in end-to-end speech Modern ASR systems span attention-based encoder-decoder models (Whisper; Radford et al. (2023)), CTC-based acoustic models (wav2vec 2.0; Baevski et al. (2020)), RNN-T transducers (Graves, 2012), Conformer-based architectures (Gulati et al., 2020), and non-autoregressive variants such as Paraformer (Gao et al., 2022)). All autoregressive variants inherently combine an acoustic encoder with a language-modeling component. The internal-LM estimation (ILME) literature (Meng et al., 2021; Tsunoo et al., 2022; Das et al., 2023) has quantified this implicit LM and proposed subtraction-based correction at inference. ILME, however, presupposes that the implicit LM is a separable, estimable component. Our negative result on `condition_on_prev_tokens`, that the flag does not weaken the prior, suggests that for transformer encoder-decoder models like Whisper, the LM influence is not localizable to the autoregressive token stream but is distributed across attention weights. This finding is consistent with the encoder-decoder asymmetry hypothesis of Atwany et al. (2025) and the mechanistic-interpretability observation that Whisper's decoder behaves as a weak standalone LM even without acoustic input.

Hallucination benchmarks in multimodal models The VLM community pioneered the construction of adversarial probe benchmarks for prior-driven failures. POPE (Li et al., 2023) uses polling-based binary queries about object presence; HallusionBench (Guan et al., 2024) extends this to image-context reasoning under intentional ambiguity; RePOPE (Neuhaus & Hein, 2025) corrects annotation artifacts in the original POPE and shows that they shift model rankings significantly. AVHBench (Sung-Bin et al., 2025) is the closest sibling to our work: it constructs a 5,816-pair benchmark for cross-modal hallucination in audio-visual LLMs such as SALMONN (Tang et al., 2024) and Qwen2-Audio (Chu et al., 2024), explicitly probing whether visual context can override acoustic grounding (and vice versa). MONDEGREENSEVAL differs in two ways: (1) it targets unimodal speech-to-text systems, not multimodal LLMs, isolating the LM prior internal to a single decoder; and (2) it uses phonetic rather than semantic ambiguity as the controlled confound. The two benchmarks are complementary — AVHBench measures whether one modality drowns out another; MONDEGREENSEVAL measures whether one decoder modality drowns out its own sensory input.

A complementary line of work frames these phenomena not as failures of grounding but as a structural property of how multimodal architectures allocate representational capacity. Kang (2026) introduces Modality Utilization Imbalance (MUI) as a unifying lens, demonstrating across the InternVL, Gemma, and Qwen3-VL families that VLMs systematically over-rely on text priors when visual evidence is weak — a phenomenon traced to the "Greedy Learner" dynamic during pretraining and mitigated at inference time via adaptive routing and soft gating of visual features. The acoustic-prior override we measure in Whisper is the direct ASR analogue of MUI: rather than dropping or attenuating an irrelevant image, an MUI-style intervention in speech would need to attenuate the decoder's own linguistic prior when acoustic evidence becomes ambiguous. MONDEGREENSEVAL is, in this sense, the audio-side measurement instrument that such an intervention would target — and the non-monotonic scaling we report in §5 (medium ¿ large-v3 in bias) closely echoes the inverse-scaling-by-policy pattern Kang (2026) reports for VLMs, where smaller models benefit more from hard switching and larger models from soft gating.

B. Tier Rationale and Dataset Statistics

Phonetic similarity between each pair is computed using stress-stripped ARPAbet sequences derived from the CMU Pronouncing Dictionary (Weide et al., 1998), with `g2p_en` used as a fallback for out-of-vocabulary items. Normalized phoneme edit distance spans 0.038 to 1.000 ($\mu = 0.332$). We partition pairs into four tiers.: near- homophones ($d < 0.10, n = 36$), plausibly ambiguous ($0.10 \leq d < 0.25, n = 64$), weakly similar ($0.25 \leq d < 0.40, n = 16$), and dissimilar ($d \geq 0.40, n = 74$). The first two tiers constitute the primary diagnostic regime where acoustic evidence is most vulnerable to override. Audio is synthesized using Edge-TTS (Microsoft, 2024) with the en-US-JennyNeural voice at 16 kHz mono WAV for- mat. Each phrase is rendered in both its canonical and mondegreen forms. Noisy variants are generated via additive Gaussian white noise at SNR levels of 15, 10, 5, 0, and -5 dB. The benchmark includes clean audio and the five noisy conditions for all pairs across both original and mondegreen text types.

C. Additional Figures

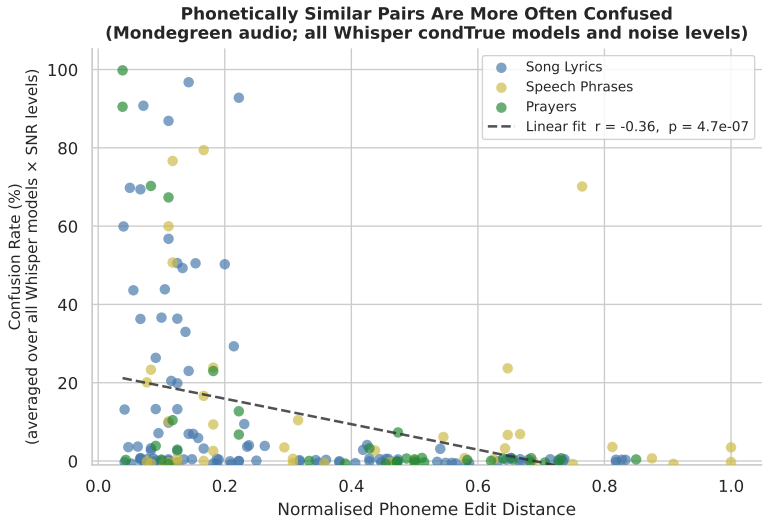

Figure 5. Per-pair confusion rate (averaged over all Whisper condTrue models and noise levels) against normalized phoneme edit distance, colored by semantic category. The negative correlation confirms that phonetic proximity between original and mondegreen phrases is the primary driver of LM confusion, not transcription error.

Fig. 5 plots per-pair confusion rate against normalized phoneme edit distance, aggregated over all Whisper condTrue models and SNR levels and colored by semantic category. The continuous view reproduces the tier finding of body Fig. 4: confusion concentrates in the $d < 0.25$ region with consistent slope across the three semantic domains.

Fig. 6 shows MCR-mono across all six audio conditions (clean and SNR $\in$ 15, 10, 5, 0, -5 dB) as a per-model line plot, complementing the three-bar summary of body Fig. 2. The widening spread among Whisper variants under noise (particularly the divergence of medium and large-v3 below 5 dB SNR) supports the claim that acoustic encoder robustness, not just LM strength, becomes the dominant factor at low SNR.

D. Decoding Hyperparameters and SNR Coverage

For LM ablation, whisper-medium and whisper-large-v3 are additionally evaluated with condition on prev tokens=False. Log-probability scoring is conducted on all four Whisper models at clean audio and SNR = 5 dB. All experiments use identical decoding constraints (no beam width expansion beyond defaults) to ensure fair comparison. The decoding hyperparameters are shown in Table 2.

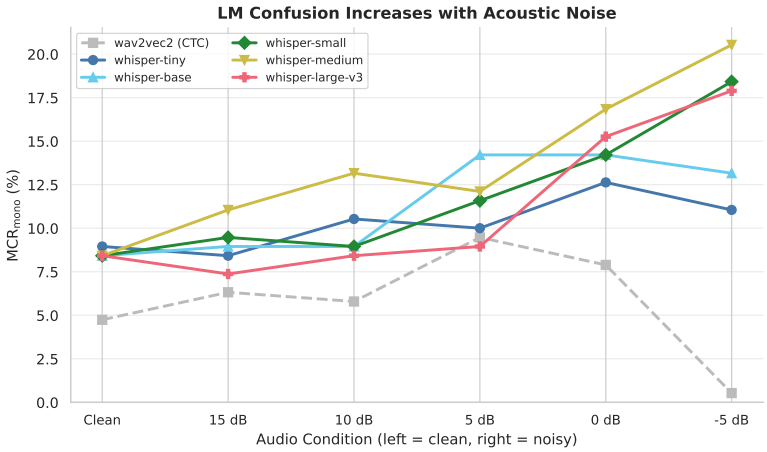

Figure 6. **MCR-mono** versus audio condition from clean (left) to SNR = -5 dB (right), one line per model. Whisper models rise monotonically with noise; wav2vec2 (dashed) rises then collapses at -5 dB as the CTC decoder fails. The spread among Whisper models widens under noise, suggesting acoustic encoder robustness becomes the dominant factor at low SNR.

Table 2. Model architectures and decoding hyperparameters. All models are loaded from Hugging Face via `transformers`. Inference hyperparameters are identical across models unless noted. The log-probability scorer uses the same prefix tokens as standard decoding: `<|startoftranscript|> <|en|> <|transcribe|> <|notimestamps|>`.

	tiny	**base**	**small**	**medium**	**large-v3**
Architecture					
Parameters	39M	74M	244M	764M	1,543M
Encoder layers	4	6	12	24	32
Decoder layers	4	6	12	24	32
d_{model}	384	512	768	1024	1280
Attention heads	6	8	12	16	20
Mel filterbanks	80	80	80	80	128
Weights dtype	fp32	fp32	fp32	fp32	fp16
Inference (all models)					
Input sampling rate: 16,000 Hz Hop length: 160 samples Audio chunk: 30 s					
Language: `en` Task: `transcribe` Timestamps: disabled					
Beam search: disabled (greedy) Batch size: 1					
`condition_on_prev_tokens`: `True` (standard) / `False` (ablation, medium & large-v3 only)					

E. Roadmap for Extensions

Along multimodal LLMs A natural extension of MONDEGREENSEVAL is evaluating audio-visual large language models in the spirit of AVHBench (Sung-Bin et al., 2025). Audio-visual systems such as SALMONN (Tang et al., 2024), and Qwen2-Audio (Chu et al., 2024) integrate visual lip-movement cues with acoustic input, and Whisper-Flamingo and Cappellazzo et al. (2025) extend this to LLM-fused decoders. In such settings, does visual context suppress or amplify the acoustic-textual LM prior? A model that simultaneously receives degraded audio of "kiss this guy" and clear lip-read cues should theoretically shift logprob bias toward the acoustic modality. Testing mondegreen pairs under joint audio-visual conditions will clarify whether cross-modal fusion acts as a regularization mechanism that grounds predictions in sensory evidence, or whether, as AVHBench observes, strong text priors persist across modality additions.

Along model families: Conformer, Paraformer, and MMS The current benchmark focuses on the OpenAI Whisper family and wav2vec2. Expanding evaluation to Conformer-based encoder-decoder architectures (Gulati et al., 2020), non-autoregressive Paraformer-v2 (Shi et al., 2022), and Meta's Massively Multilingual Speech models (Pratap et al., 2024) is critical. Paraformer-v2 in particular is reported as noise-robust by design and uses CTC for token extraction, predicting that MCR should be substantially lower than autoregressive Whisper at matched SNR — a prediction that would, if confirmed, isolate the autoregressive decoder as the locus of mondegreen confusion. The recent Phi-4-Multimodal-Instruct (Abouelenin

et al., 2025), a fully decoder-based speech model, is hypothesized to exhibit stronger mondegreen bias for the same reason, consistent with the architectural observation in Atwany et al. (2025) that decoder-only speech models prioritize linguistic fluency over exact acoustic matching.

Deeper inspections of the acoustic-textual imbalance Whisper-family models possess highly asymmetrical encoder-decoder capacities; the acoustic encoder processes raw waveforms with deep hierarchical representations, while the language-modeling head maps these features into a discrete token space via a relatively lightweight projection. Early mechanistic-interpretability work on Whisper showed that the decoder alone behaves as a weak standalone LM and that encoder attention is highly localized. We hypothesize that this architectural imbalance favors text modality utilization: when acoustic features are noisy, the bottleneck in mapping them to tokens becomes a stochastic process, allowing the language prior to dominate the output distribution. Future work should employ probing classifiers and representational similarity analysis (RSA), as well as attention-map-based hallucination detection (recently shown effective for SpeechLLMs), to track how mondegreen confusion rates correlate with encoder feature degradation at each transformer layer.

Open-sourcing the benchmarking toolkit To facilitate community adoption, we will open-source a comprehensive benchmarking pipeline that automates mondegreen audio synthesis, noise injection, multi-model evaluation, and dual-metric scoring (MCR and logprob bias). By standardizing the testing procedure, this toolkit will enable rapid stress-testing of newly released speech and audio-LLM models, ensuring that acoustic-linguistic tradeoffs are routinely audited alongside conventional WER, CER, and HER benchmarks.