Residual Stream Contrast: A Training-Free Counterfactual Listening Test for Whisper Hallucinations

Arnesh Batra [1]

Abstract

The failure we target is simple: Whisper can emit a fluent transcript that remains internally plausible even when the audio is removed. Confidence alone does not test whether the transcript needed the audio. We introduce Residual Stream Contrast (RSC), a training-free white-box score that teacher-forces the same decoded transcript under real and null audio and measures whether the decoder residual trajectory changes. We evaluate it on the 1,663-example activation split, which serves as our benchmark, with every candidate detector scored under identical conditions. RSC is the strongest single detector on this benchmark, reaching .947 AUROC, .943 AUPRC, and .824 TPR at 5% FPR. Relative to AvgLogP, it improves AUROC by +.051 and TPR@5 by +.186, with paired bootstrap intervals excluding zero. A grounding-head atlas localizes the signal to a sparse set of cross-attention heads, led by L8H8 in Whisper-small, and a head-entropy diagnostic built from this atlas provides a secondary, mechanistically grounded check on the same conclusion: residual evidence is load-bearing.

1. Introduction

Modern automatic speech recognition (ASR) has moved from hand-engineered pipelines toward end-to-end sequence models. CTC and RNN-T removed explicit frame alignments (Graves et al., 2006; Graves, 2012); attention-based systems such as Listen, Attend and Spell and Deep Speech 2 showed that large neural recognizers could model recognition end to end (Chan et al., 2015; Amodei et al., 2016); Transformer and Conformer architectures, together with augmentation recipes such as SpecAugment, made sequence modeling more robust and scalable (Vaswani et al., 2017; Gulati et al., 2020; Park et al., 2019). Self-supervised encoders and large weakly supervised recognizers, including wav2vec 2.0, HuBERT, WavLM, and Whisper, pushed this trend into wide deployment (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Radford et al., 2022).

This progress changes the reliability problem. ASR is now used for captions, meeting notes, search, contact centers, accessibility workflows, and audio archives. At the same time, the training recipe often mixes many acoustic domains and languages (Chan et al., 2021; Chen et al., 2021; Ardila et al., 2020), while audio applications increasingly include non-speech and environmental sound (Gemmeke et al., 2017; Salamon et al., 2014; Piczak, 2015). In such settings, a model can output a transcript that is grammatical, repeated, or socially plausible even when the relevant words were not present in the audio.

Standard confidence scores catch many failures, but they ask the wrong question for this particular failure mode. AvgLogP asks whether the emitted tokens were easy for the decoder. A no-speech score asks whether the front end saw speech-like evidence. A voice activity detector asks whether speech was present at all. None of these directly asks whether the specific transcript depended on the audio. A hallucinated phrase can be high confidence because it is fluent under the decoder language prior; a rare but correct phrase can be low confidence while still being grounded.

We make this distinction measurable. Given an emitted transcript, RSC replays the same token sequence under real audio and under a null-audio encoder state. If the decoder residual stream under real audio remains close to the residual stream under null audio, the transcript is suspicious: the decoder appears to be walking through nearly the same internal trajectory without listening. To test this claim fairly, we use the 1,663-example activation split as our benchmark and score every detector under comparison on exactly the same examples, so that no detector benefits from an easier or larger sample.

The paper has one headline claim: RSC is the best single training-free listening detector on the activation benchmark. The grounding-head atlas provides supporting evidence by

[1]Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), New Delhi, India. Correspondence to: Arnesh Batra <arnesh.batra@iiitd.ac.in>.

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

showing where the signal lives. This framing avoids a common confusion: the strongest number in the paper is not the contribution if it comes from a learned or composite witness. The contribution is the counterfactual listening primitive and the matched activation benchmark used to test it.

Contributions.

- We define RSC, a single training-free score for the question: did this transcript change the decoder's residual trajectory relative to no audio?
- We establish the 1,663-example activation split as a matched benchmark on which all detectors are evaluated on identical examples.
- We show that RSC is the best single detector on this benchmark, reaching .947 AUROC, .943 AUPRC, and .824 TPR@5.
- We verify the gain over AvgLogP with paired bootstrap intervals on identical examples (Efron & Tibshirani, 1993).
- We use a grounding-head atlas as a secondary diagnostic that explains and stress-tests the RSC signal.

2. Related Work

Confidence, calibration, and ASR uncertainty. Confidence estimation is not the same as grounding. Neural calibration work shows that high predictive probability need not match empirical correctness (Guo et al., 2017). In ASR, token likelihood is shaped by acoustic evidence, language priors, search, and calibration; Whisper-specific confidence studies therefore remain active (Aggarwal et al., 2025; Huo et al., 2025). RSC is complementary to calibration: it does not estimate a probability of correctness. It asks whether the hidden state trajectory changes when the audio evidence is removed.

Whisper hallucinations and mitigation. Recent work documents Whisper hallucination harms and failure conditions, including pauses, non-vocal audio, long-form drift, and noisy speech (Koenecke et al., 2024; Frieske & Shi, 2024; Bain et al., 2023). Local confidence contrasts offer output-side hallucination detection (Corpataux et al., 2026), while mitigation methods alter decoding or attention behavior (Ahn et al., 2026; Wang et al., 2025). RSC instead stays in monitoring mode: it can be run after ordinary decoding, does not retrain the model, and ranks a finished transcript by its audio dependence.

Mechanistic audio interpretability. Mechanistic analyses of ASR and audio-language models use activations, logit lenses, and specialist heads to ask whether models rely on audio evidence (Glazer et al., 2026b;a). Our grounding-head atlas follows this direction, but with a deployment-facing target: a low-FPR hallucination monitor whose signal can be traced to residual counterfactuals and cross-attention heads.

3. Method

Problem setting. Let x be an audio segment and let $y = (y_1, \ldots, y_T)$ be the transcript already decoded by Whisper. A monitor receives (x, y) and returns a scalar hallucination-risk score. Higher scores should rank transcripts that look less supported by the audio. The monitor is post-hoc: it does not search for a new transcript and does not change the decoded text.

Why likelihood is not enough. Average log probability mixes at least two effects: how much the audio supports the transcript and how much the decoder language prior likes the transcript. A useful decomposition is

$$\ell(y; x) = \ell(y; \varnothing) + \frac{1}{T} \sum_{t=1}^{T} \log \frac{p(y_t \mid y_{<t}, x)}{p(y_t \mid y_{<t}, \varnothing)}, \quad (1)$$

where $\varnothing$ denotes the run's null-audio encoder state. The second term is an audio-PMI style contrast (a-PMI). On our activation benchmark, a-PMI is useful but remains weaker than RSC, especially at low false-positive rates. This motivates a residual-state diagnostic rather than a probability-only diagnostic.

Residual Stream Contrast. Let $r_t(x)$ be the decoder residual vector saved at a fixed diagnostic site while teacher-forcing y under the real audio encoder state, and let $r_t(\varnothing)$ be the corresponding vector under the null-audio state. Let $\hat{r}$ denote robust normalization with calibration-set medians and median absolute deviations. RSC is

$$\mathrm{RSC}(x, y) = \frac{1}{T} \sum_{t=1}^{T} \frac{\langle \hat{r}_t(x), \hat{r}_t(\varnothing) \rangle}{\|\hat{r}_t(x)\|_2 \, \|\hat{r}_t(\varnothing)\|_2}. \quad (2)$$

High RSC means real-audio and no-audio residual trajectories are similar. The score is therefore a direct measurement of the non-listening hypothesis.

What the contrast isolates. The null run is not a new decoding baseline; it is a counterfactual replay of the same transcript. Teacher forcing fixes the token path, robust scaling prevents large-norm coordinates from dominating the cosine, and token averaging makes the detector sensitive to sustained non-listening rather than one odd token. Read with confidence, low confidence plus low RSC suggests hard but audio-dependent speech, while high confidence plus high RSC is the suspicious fluent-prior regime.

Algorithm 1 Residual Stream Contrast

Input: audio x, emitted tokens $y_{1:T}$, null state $\varnothing$, residual site s, robust center/scale (m, q)
Encode x to get audio state e_x.
Teacher-force $y_{1:T}$ with e_x and with $\varnothing$.
for $t = 1$ to T **do**
  Read residuals $r_t(x)$ and $r_t(\varnothing)$ at site s.
  Normalize $\hat{r}_t = (r_t - m)/(q + \epsilon)$.
  $c_t \leftarrow \cos(\hat{r}_t(x), \hat{r}_t(\varnothing))$.
end for
return $\frac{1}{T}\sum_t c_t$.

Table 1. Results on the 1,663-example activation benchmark. TPR5 is TPR at 5% FPR. Every detector is evaluated on the same examples.

Detector	n	AUROC	AUPRC	TPR5
AvgLogP	1663	.896	.884	.638
a-PMI	1663	.900	.912	.797
Top-8 heads	1663	.905	.849	.814
RSC	1663	**.947**	**.943**	**.824**

Benchmark and protocol. We use the activation split as our benchmark. It contains 1,663 examples with decoder residuals, cross-attention summaries, and output probabilities. Every detector – AvgLogP, a-PMI, top-8 head entropy, and RSC – is evaluated on exactly these examples. This matched design prevents any detector from benefiting from an easier or larger sample. We report AUROC, AUPRC, and TPR at 5% FPR. TPR@5 is emphasized because a useful hallucination filter cannot reject many correct speech segments. Paired bootstrap differences are computed on identical benchmark examples.

4. Results

Headline: RSC is the strongest single listening score. Figure 1 and Table 1 give the headline result on the 1,663-example activation benchmark with 779 positives. RSC reaches .947 AUROC, .943 AUPRC, and .824 TPR@5. It outperforms AvgLogP (.896/.884/.638), a-PMI (.900/.912/.797), and top-8 head entropy (.905/.849/.814). This is the claim we emphasize because every detector is evaluated on identical examples.

Identical-example bootstrap confirms the gain. On the activation benchmark, RSC improves over AvgLogP from .896 to .947 AUROC and from .638 to .824 TPR@5. Paired bootstrap resampling gives a $+.051$ AUROC difference with interval $[.035, .071]$ and a $+.186$ TPR@5 difference with interval $[.147, .229]$. The identical-example design matters: the gain is not an artifact of RSC being computed on a different or easier sample.

Grounding heads are a useful ceiling, not the contribution. Top-8 head entropy, built directly from the grounding-head atlas described below, reaches .905 AUROC and .814 TPR@5 on the activation benchmark, competitive at low FPR but requiring head-level attention diagnostics rather than a single residual contrast. This is informative rather than redundant: it shows that the non-listening signal is not unique to the residual stream and is also visible in cross-attention behavior, which strengthens the read that RSC measures a real mechanistic property rather than a numerical artifact.

Grounding heads make the signal auditable. The grounding-circuit atlas ranks decoder heads by attention and temporal-grounding diagnostics. The top heads are L8H8, L7H9, L2H5, L5H3, L8H0, L10H5, L6H6, and L8H4. L8H8 has head AUROC .860, head temporal-grounding AUROC .891, grounding gap .228, and circuit score .980 over 1,285 atlas examples. Top-8 head entropy alone reaches .905 AUROC and .814 TPR@5 on the activation benchmark, as reported above. These heads do not replace RSC; they show that the non-listening signal is not a shapeless scalar but is concentrated in interpretable cross-attention structure.

Audits and slices. Label-shuffle audits stay near chance: pooled shuffled AUROC means are .506 for RSC, .506 for AvgLogP, and .498 for top-8 head entropy, with all reported pooled detectors passing the audit. Slice robustness is less uniform. AvgLogP is best on the Earnings22 benchmark slice, while RSC is best on the UrbanSound8K slice and the environmental condition family. Some environmental slices have extreme class imbalance, which makes fixed thresholds brittle. This unevenness is part of the result: RSC is a strong listening diagnostic, not a universal deployment policy.

5. Discussion and Limitations

RSC is most useful as a second-stage monitor. A production system can first use cheap confidence, VAD, or segmentation filters, then run RSC on high-stakes or suspicious segments. The cost is white-box access and one null-audio teacher-forced pass. The benefit is a different kind of evidence: instead of asking whether the transcript was likely, RSC asks whether the transcript's internal trajectory needed the audio.

The current evidence has clear boundaries. The activation benchmark is smaller than what a purely output-level benchmark could reach because internal activations are more expensive to collect. The grounding-head diagnostic is competitive with RSC at low FPR, but it requires head-level attention extraction rather than a single residual contrast, so it sacrifices some of the simplicity of the single score. Non-Whisper checks use output-level CTC confidence only,

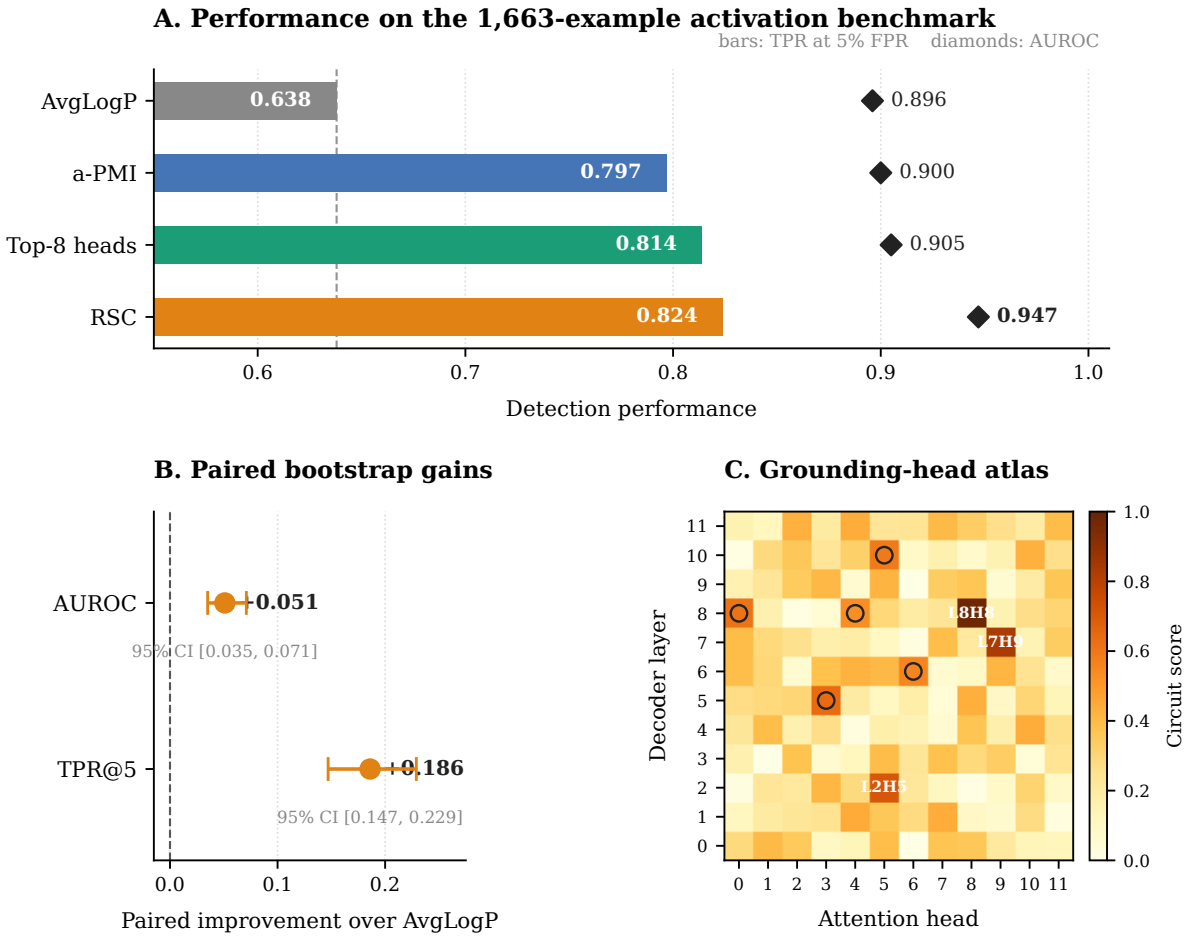

Figure 1. Results on the 1,663-example activation benchmark. Panel A compares all detectors on identical examples; bars show TPR at 5% FPR and diamonds show AUROC. Panel B reports paired-bootstrap improvements of RSC over AvgLogP with 95% confidence intervals. Panel C presents the grounding-head atlas, showing that the non-listening signal is concentrated in a sparse set of decoder cross-attention heads.

so they should be read as smoke tests rather than evidence that RSC transfers to encoder-only or CTC systems. Labels are also imperfect: WER can punish harmless variants, and non-speech examples assume lexical output is undesirable. Bootstrap intervals, label-shuffle audits, and slice tables support a conservative reading: the residual contrast is real and useful at low FPR, while deployment policy must still handle domain shift and class imbalance.

Operational reading. RSC is a triage signal for one counterfactual listening question, not a replacement for calibration, segmentation, or human review. Confident low-RSC transcripts can pass ordinary checks; low-confidence low-RSC transcripts look like hard audio; high-RSC transcripts deserve escalation because the decoder follows a similar residual path even without audio. The activation benchmark is therefore the headline: it compares all detectors on identical examples instead of hiding the claim inside a larger composite.

Ethically, RSC should separate difficult grounded speech from non-listening: a low-resource phrase can be uncertain but grounded, while a common phrase can be confident and survive audio removal.

Conclusion. RSC turns hallucination detection into a counterfactual listening test: RSC is the best single training-free detector, its gain over AvgLogP is paired and bootstrapped, and fluent output should change residual traces when audio is present.

Impact Statement

This paper presents work whose goal is to advance the reliability of automatic speech recognition systems. RSC is a diagnostic tool intended to help practitioners detect and audit hallucinated transcripts, which is a positive application for accessibility, captioning, and archival use cases where unfaithful transcripts can cause real harm. We are not aware of a plausible misuse of a hallucination *detector*

itself, though we note that, like any monitoring signal, RSC should be treated as a triage aid rather than a certification of correctness, and should not be used as the sole basis for high-stakes decisions without human review. Beyond these considerations, there are no potential societal consequences of our work that we feel must be specifically highlighted here.

Acknowledgements

The author thanks the ICML reviewers for their time and feedback.

References

Aggarwal, V., Nair, S. S., Verma, Y., and Jogi, Y. Adopting whisper for confidence estimation, 2025.

Ahn, H., Chae, J., Park, Y., and Shim, K. Whisper-cd: Accurate long-form speech recognition using multi-negative contrastive decoding, 2026.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 173–182. PMLR, 2016.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, 2020. European Language Resources Association.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Bain, M., Huh, J., Han, T., and Zisserman, A. Whisperx: Time-accurate speech transcription of long-form audio, 2023.

Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. Listen, attend and spell, 2015.

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., and Norouzi, M. Speechstew: Simply mix all available speech recognition data to train one large neural network, 2021.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., Wang, Y., You, Z., and Yan, Z. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio, 2021.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. Fleurs: Few-shot learning evaluation of universal representations of speech, 2022.

Corpataux, S., Scius-Bertrand, A., and Wolf, B. Detecting whisper hallucinations with local confidence contrasts. In Kucharavy, A., Delgado, P., Schurch Todeschini, V., and Rumley, S. (eds.), *Proceedings of the Fourth Swiss AI Days*, volume 309 of *Proceedings of Machine Learning Research*, pp. 38–45. PMLR, 2026.

Del Rio, M., Ha, P., McNamara, Q., Miller, C., and Chandra, S. Earnings-22: A practical benchmark for accents in the wild, 2022.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

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

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 776–780, 2017.

Glazer, N., Aharon, L., and Fetaya, E. Are audio-language models listening? audio-specialist heads for adaptive audio steering, 2026a.

Glazer, N., Segal-Feldman, Y., Segev, H., Shamsian, A., Buchnick, A., Hetz, G., Fetaya, E., Keshet, J., and Navon, A. Beyond transcription: Mechanistic interpretability in ASR. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37407–37416, 2026b.

Graves, A. Sequence transduction with recurrent neural networks, 2012.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376, 2006.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., and Wu, Y. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pp. 5036–5040, 2020.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

Huo, M., Zhang, Y., and Tang, Y. Identifying and calibrating overconfidence in noisy speech recognition, 2025.

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 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1672–1681, 2024.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

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. In *Interspeech 2019*, pp. 2613–2617, 2019.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2022.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, 2014.

Snyder, D., Chen, G., and Povey, D. Musan: A music, speech, and noise corpus, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 993–1003, Online, 2021. Association for Computational Linguistics.

Wang, Y., Alhmoud, A., Alsahly, S., Alqurishi, M., and Ravanelli, M. Calm-whisper: Reduce whisper hallucination on non-speech by calming crazy heads down, 2025. Accepted to Interspeech 2025.

A. Benchmark Coverage

Table 2. Per-dataset coverage of the activation benchmark. "Act." counts benchmark examples with cached internal activations; "Act. pos." counts positive (hallucinated) examples among them.

Dataset	Act.	Act. pos.
BoH-Aug	194	30
Earnings22	121	40
FLEURS	187	11
LibriSpeech	344	28
UrbanSound8K	288	274
VoxPopuli	129	7
ESC-50	200	195
MUSAN	200	194

The public corpora that make up our benchmark include LibriSpeech (Panayotov et al., 2015), UrbanSound8K (Salamon et al., 2014), Earnings-22 (Del Rio et al., 2022), FLEURS (Conneau et al., 2022), VoxPopuli (Wang et al., 2021), ESC-50 (Piczak, 2015), and MUSAN (Snyder et al., 2015). Coverage counts in Table 2 are computed from our benchmark construction; empirical claims in the paper are limited to these examples.

B. Additional Interpretability Diagnostics

The activation benchmark contains 1,663 examples with 779 positives. RSC has AUROC .947, AUPRC .943, TPR@1 .317, TPR@5 .824, and TPR@10 .878. AvgLogP reaches AUROC .896, AUPRC .884, and TPR@5 .638; a-PMI reaches AUROC .900, AUPRC .912, and TPR@5 .797; top-8 head entropy reaches AUROC .905, AUPRC .849, and TPR@5 .814. Label-shuffle audits pass for all reported pooled detectors, with shuffled AUROC intervals centered near chance.

C. Implementation Notes

Null audio is implemented by replacing the encoder state with the run's null-audio state during teacher forcing. The same emitted token sequence is replayed, so the score measures support for the transcript rather than search variation. Residuals, cross-attention summaries, and log probabilities are cached for reuse across the benchmark construction pipeline. Robust score scaling uses median and median absolute deviation over the matched calibration pool. All reported thresholds are evaluated at fixed FPR operating points on held-out benchmark examples, and pairwise differences are bootstrapped on identical examples.

Non-Whisper checks use output-level CTC average log probability for wav2vec 2.0 and HuBERT models (Baevski et al., 2020; Hsu et al., 2021). Those checks are smoke tests only; internal RSC requires decoder residual access in the Whisper-family model.

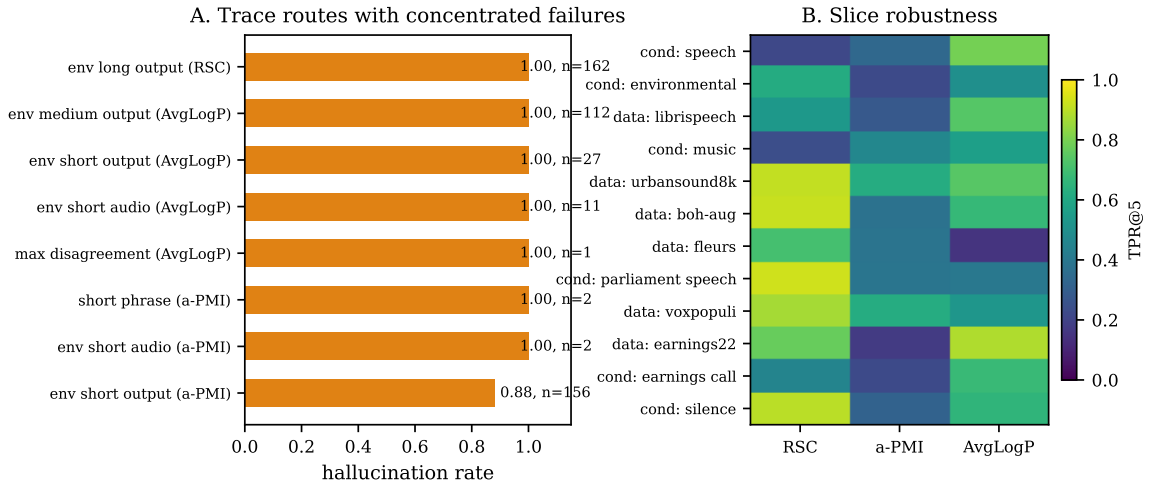

Figure 2. Appendix diagnostics from our benchmark. Panel A: trace routes show concentrated failure regimes. Panel B: slice robustness is uneven across datasets and condition families.