BEYOND STATIC BRAIN ALIGNMENT: CROSS-CONDITION NEURAL TRANSPORTABILITY IN SPEECH MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain–model alignment is typically evaluated within the stimulus regime used to fit the neural mapping, leaving open whether the fitted relationship remains predictive when sensory conditions change. We operationalize cross-condition neural transportability using leave-one-condition-out (LOCO) evaluation, fitting a model-to-brain mapping on three speech conditions and testing it on an unseen fourth. Using fMRI from 25 participants hearing 96 sentences across Clean, Noisy, DNN-enhanced, and classical speech-enhanced conditions, we find that conventional within-condition alignment is broad whereas cross-condition evaluation is more selective. Under strict fold-local layer selection, encoder representations show greater transportability than matched decoder representations in Whisper-Tiny, Whisper-Medium, and SpeechT5, with higher transportability concentrated in encoder and intermediate representations. Architecture-matched controls further show that trained Whisper-Tiny exceeds its random-weight counterpart, providing direct evidence that learned structure can improve neural transportability beyond architectural inductive bias in at least one controlled model family. A complementary descriptive analysis shows representation-class AUROC increasing from 0.543 under static evaluation to 0.836 under LOCO, while static alignment and cross-condition transportability remain only moderately associated. Sentence-level LOCO fit does not improve held-out comprehension prediction beyond acoustic condition and Automatic Speech Recognition (ASR) character error rate. Because sentence identity is fixed to condition, LOCO measures joint transfer across acoustic regime and item set rather than the acoustic manipulation alone. Together, these findings show that strong within-condition neural correspondence does not guarantee a transportable model-to-brain mapping, supporting cross-condition neural transportability as a complementary criterion for evaluating brain-aligned speech representations.

1 INTRODUCTION

Deep neural networks have become powerful computational models of biological information processing. Task-optimized and self-supervised representations predict neural responses across vision (Yamins et al., 2014; Conwell et al., 2024), audition and speech (Kell et al., 2018; Tuckute et al., 2023; Millet et al., 2022; Goldstein et al., 2025), and language (Schrimpf et al., 2021; Caucheteux & King, 2022). Model depth can also parallel cortical processing hierarchies (Keshishian et al., 2026), making neural predictivity an important criterion for evaluating learned representations (Feather et al., 2025).

Most brain–model benchmarks, however, evaluate a fitted neural mapping on held-out stimuli drawn from the same broad distribution used for training (Schrimpf et al., 2020). Such tests establish *within-distribution correspondence*, but not whether the same model-to-brain relationship remains predictive when sensory conditions change. This distinction is particularly important for speech, where noise and speech-enhancement processing can substantially alter the acoustic signal while preserving useful linguistic information, and auditory cortex exhibits both noise invariance and rapid adaptation to changing acoustic statistics (Kell & McDermott, 2019; Khalighinejad et al., 2019). Recent work in vision highlights this limitation. Models with similar in-distribution neural predictivity

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.

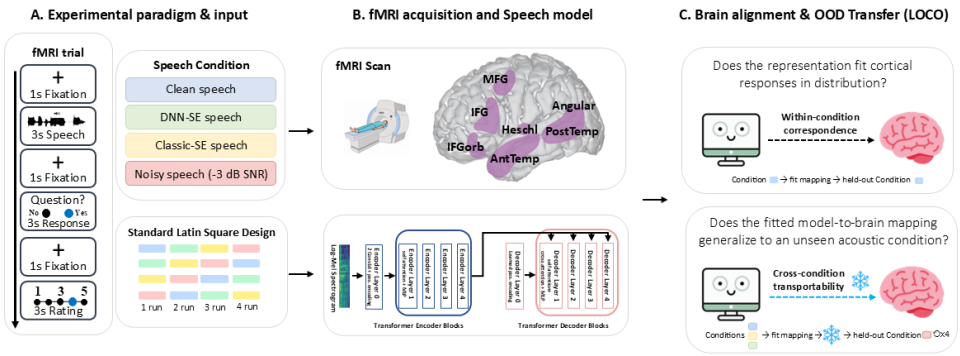


Figure 1: **Cross-condition brain-model alignment benchmark.** (A) Participants heard Mandarin sentences across Clean, DNN-SE, Classic-SE, and Noisy ($-3\,\mathrm{dB}$ SNR) conditions during fMRI, followed by comprehension and subjective intelligibility ratings. (B) Sentence-specific responses from 14 bilateral auditory-language ROIs were compared with hierarchical speech-model representations (e.g., whisper-tiny). (C) Conventional encoding tests within-condition correspondence, whereas LOCO fits the model-to-brain mapping on three conditions, freezes it, and evaluates cross-condition neural transportability on the unseen fourth condition.

can diverge under out-of-distribution (OOD) evaluation (Ren & Bashivan, 2024); neural datasets have been developed explicitly for OOD testing (Gifford et al., 2026); and model-disagreement and closed-loop paradigms seek stimuli that distinguish otherwise compatible computational hypotheses (Golan et al., 2026; Tuckute et al., 2024; Gifford et al., 2025). More broadly, recent perspectives argue that predictive fit should be complemented by stronger tests of generalization and model discrimination (Schmitt et al., 2026; Serre & Pavlick, 2025).

We bring this principle to speech through *cross-condition neural transportability*. Rather than asking only whether a representation predicts cortical responses within a condition, we ask whether the *same fitted model-to-brain mapping remains predictive in an unseen condition*. We operationalize this question using leave-one-condition-out (LOCO) evaluation: the complete neural-encoding pipeline is fit on three conditions, frozen, and evaluated on the held-out fourth. LOCO itself is a standard cross-validation strategy; our contribution is to use condition-held-out transportability as a criterion for brain–model alignment and to test whether it changes what we conclude about speech representations. We evaluate cross-condition neural transportability using fMRI from 25 participants who heard 96 sentences across four acoustic regimes: Clean speech, Noisy speech at $-3\,\mathrm{dB}$ SNR, DNN-enhanced speech (DNN-SE; SEMamba (Chao et al., 2024)), and classically enhanced speech (Classic-SE; Ephraim & Malah, 1984). Because sentence identity was nested within condition, LOCO evaluates joint transfer across acoustic regime and item set rather than an isolated causal acoustic effect. Primary inference is based on strict participant-level matched comparisons with fold-local layer selection; pooled representation-class analyses are used only to visualize how the evaluation criterion changes model discrimination. Our contributions are:

- We operationalize cross-condition neural transportability as a complementary criterion for brain–model alignment, asking whether a fitted mapping remains predictive after an entire speech condition is withheld.
- We show that cross-condition evaluation exposes representation-stage selectivity largely obscured by static alignment, with strict matched comparisons showing greater transportability for encoders than decoders in three sequence-to-sequence systems.
- Using architecture-matched random controls, we show that learned structure can improve transportability beyond architectural inductive bias, with corrected direct evidence for Whisper-Tiny.
- We distinguish transportability from ASR recoverability and trial-wise comprehension, establishing it as a complementary evaluation axis rather than a behavioral proxy.

2 Related Work

Representational similarity analysis and neural encoding provide complementary tools for comparing artificial and biological representations (Kriegeskorte et al., 2008; Nili et al., 2014), while integrative benchmarking has established neural predictivity as an important model criterion (Schrimpf et al., 2020). Such correspondence, however, is typically evaluated within the regime used to fit the neural mapping; strong within-regime prediction does not establish that the same fitted model-to-brain relationship remains predictive under distribution shift. Neural correspondence also need not arise entirely from learned structure: untrained architectures can exhibit non-trivial cortical alignment, while preprocessing and dimensionality choices can produce strong but non-robust correspondence (Kazemian et al., 2025; Hadidi et al., 2026). Modern speech models, including wav2vec 2.0 (Baevski et al., 2020), WavLM (Chen et al., 2022), Whisper (Radford et al., 2023), and SpeechT5 (Ao et al., 2022), provide complementary opportunities to test whether learning contributes beyond architectural inductive bias and whether representation stages differ in the stability of their neural mappings.

Work in vision further shows that models with similar in-distribution neural predictivity can diverge under out-of-distribution evaluation (Ren & Bashivan, 2024), motivating OOD neural datasets, model-disagreement approaches, and closed-loop tests of model predictions (Gifford et al., 2026; Golan et al., 2026; Tuckute et al., 2024; Gifford et al., 2025). These developments support broader calls to evaluate model generalization beyond predictive fit alone (Schmitt et al., 2026; Serre & Pavlick, 2025). Speech provides a natural setting for this question because auditory cortex exhibits both invariance to background noise and adaptation to changing acoustic statistics (Kell & McDermott, 2019; Khalighinejad et al., 2019). We bring these threads together by testing whether a fitted neural mapping remains transportable across clean, noisy, and enhanced speech conditions, while distinguishing cross-condition neural transportability from machine recognition and trial-wise behavioral success.

3 Methods

3.1 Participants, stimuli, and behavioral task

Twenty-five healthy participants with normal hearing completed the fMRI experiment. They heard 96 Mandarin sentences (approximately 10 words each), with 24 assigned to each condition: Clean, Noisy, DNN-SE, and Classic-SE. Noisy speech was mixed with stationary speech-shaped noise at $-3\,\mathrm{dB}$ SNR; DNN-SE used SEMamba (Chao et al., 2024), and Classic-SE used a classical MMSE estimator (Ephraim & Malah, 1984). Stimuli were presented through MRI-compatible headphones. After each sentence, participants completed a two-alternative comprehension question and rated intelligibility on five-point scales. Because sentence identity was nested within condition, LOCO tests joint transfer across acoustic regime and item set rather than an isolated causal acoustic effect. Full task details are provided in Appendix D.

3.2 fMRI acquisition, preprocessing, and cortical regions

Data were acquired on a Siemens Skyra 3T scanner and preprocessed in SPM12 using slice timing correction, motion correction, coregistration, MNI normalization, and 4 mm smoothing. A single trial GLM yielded one beta map per sentence. Multivoxel response vectors were extracted from 14 bilateral auditory and language ROIs: Heschl's gyrus, defined with MarsBaR (Brett et al., 2002), and six language network regions comprising anterior temporal, posterior temporal, angular gyrus, inferior frontal, orbital inferior frontal, and middle frontal cortex following Fedorenko et al. (2010; 2024). Appendix D provides full acquisition and ROI details.

3.3 Computational model representations

We evaluated log-mel acoustics; random and trained wav2vec 2.0 (Baevski et al., 2020); a random Whisper-Tiny control; WavLM (Chen et al., 2022); Whisper-Tiny, Whisper-Medium, and Whisper-Large-v3 encoder and decoder representations (Radford et al., 2023); SpeechT5 encoder and decoder representations (Ao et al., 2022); and transcript-based Qwen2.5-3B (Yang et al., 2025), Phi-

4-mini (Abouelenin et al., 2025), GPT-2 XL (Radford et al., 2019), and Mistral-7B (Jiang et al., 2023). Audio models received condition-specific speech, whereas text models received ground-truth transcripts and served as condition-invariant linguistic reference spaces rather than matched speech stages. Audio was converted to mono, resampled to $16\,\mathrm{kHz}$, and standardized to $4\,\mathrm{s}$ before layer-wise feature extraction. Whisper encoder representations retained stimulus-aligned states, whereas decoder representations used the final non-special token state; analogous sentence-level representations were extracted from the remaining models. Exact checkpoints, layer definitions, inputs, and extraction procedures are provided in Appendix F.

3.4 Linear encoding, neural predictivity, and ceiling-normalized Brain Score

Following Schrimpf et al. (2021), we fit ridge-regularized linear mappings from model representations to BOLD responses. Features are standardized on training data, sparse random projection is applied above the Johnson–Lindenstrauss target (Achlioptas, 2001), and voxel-wise ridge penalties are selected within training data (Hoerl & Kennard, 1970; Pedregosa et al., 2011). Held-out predictivity is Pearson's r between predicted and measured responses, Fisher-z averaged across static folds, and normalized by a condition- and task-matched cross-subject noise ceiling to yield *Brain Score*. Numerator and denominator use identical model-independent voxel masks and aggregation, and signed Brain Scores are retained for inference. Appendix F provides implementation and ceiling-estimation details.

3.5 Static within-condition encoding and LOCO evaluation

Static encoding uses four within-condition run-wise folds, fitting on 18 sentences and testing on 6 per fold. LOCO instead withholds one complete condition, fitting on 72 sentences from the remaining three conditions and evaluating only on the 24 sentences of the held-out condition; each condition is held out once (Figure 1). All train-derived operations, including layer selection, are frozen before held-out evaluation. Because static and LOCO differ in training-set size, their absolute Brain Score difference is not interpreted as a pure distribution-shift penalty. We instead compare the structure and representation discrimination revealed by the two evaluation criteria.

3.6 Layer selection: descriptive and strict policies

Two layer-selection policies are distinguished. Condition-general summaries select one layer per model stack from internal-validation performance averaged across the four conditions and apply that layer to descriptive static-versus-LOCO analyses. The strict inferential analysis instead reselects the layer independently inside each outer LOCO fold using only the three training conditions; the held-out condition contributes only to final scoring. The chosen layer is shared across all 14 ROIs within a model stack, and encoder and decoder stacks are selected separately. Clean-anchor, condition-specific, final-layer, and early/middle/late policies are retained as sensitivity analyses (Appendix C).

3.7 Participant-level inference, GlobalLOCO, and multiplicity

Participants ($N = 25$) are the independent inferential units. The primary confirmatory comparisons are matched encoder-versus-decoder contrasts within sequence-to-sequence systems and architecture-matched trained-versus-random contrasts; broader family comparisons are complementary. For the strict LOCO analysis, let $l^*(m, c)$ denote the layer selected for model stack m using inner validation on the three training conditions of outer fold c. For participant s, GlobalLOCO is the equal-weight mean of signed ceiling-normalized Brain Scores across the four held-out conditions and 14 ROIs at the corresponding fold-local selected layer:

$$\text{GlobalLOCO}(s, m) = \frac{1}{4R} \sum_{c} \sum_{r} \text{BrainScore}(s, m, l^*(m, c), r, c), \quad R = 14. \tag{1}$$

Participant-level paired differences in this quantity define the effect sizes in the strict matched comparisons. Benjamini-Hochberg false-discovery-rate correction is applied within each prespecified comparison family (Benjamini & Hochberg, 1995). Fold-specific effects are descriptive robustness summaries rather than four independent confirmatory tests.

Representation-class discrimination is also descriptive. The pooled ROC compares six trained speech-encoder stacks with a heterogeneous comparison set comprising decoder stacks and transcript-based linguistic reference representations. Because the same representation identities recur across ROIs, pooled AUROC is not treated as independent inferential evidence. Appendix C.3 reports an ROI-wise sensitivity analysis. The strict participant-level matched tests remain the primary statistical evidence. The RSA appendix uses participant-level two-sided sign-flip permutation inference (Nichols & Holmes, 2002; Winkler et al., 2014); correction families and reporting rules are specified in Appendix E.

3.8 Machine recoverability and behavioral boundary test

Character error rate (CER) provides a sentence-level engineering baseline. The original behavioral pipeline uses Whisper-Tiny, with robustness analyses based on Whisper-Medium, Whisper-Large-v3, and a standardized three-model ensemble. Behavioral prediction is grouped by sentence identity so that all participant responses to one sentence remain in the same outer fold. B1 uses Condition + CER; B2 adds static neural fit; B3 adds LOCO neural fit; and B4 adds both. Held-out log loss is primary, with AUROC and Brier score secondary. Random-weight LOCO is a negative control and the listener's own intelligibility rating is a positive machinery check.

3.9 Complementary matched-layer RSA

For each condition, sentence-level model states and multivoxel brain patterns are converted into representational dissimilarity matrices using Pearson correlation distance, with model–brain RSA computed as the Spearman correlation across the 276 unique upper-triangle entries (Kriegeskorte et al., 2008; Nili et al., 2014). Matched-layer ΔRSA compares each manipulated condition with Clean at the same computational layer, with participant-level effects tested using two-sided sign-flip permutations and BH-FDR correction (Appendix A). LOCO tests whether a fitted predictive mapping transfers across regimes, while RSA tests how model–brain representational geometry changes across conditions.

4 Results

4.1 Cross-condition evaluation changes representation discrimination

Within-condition encoding yields substantial Brain Scores across learned speech encoders, decoder representations, transcript-based references, acoustic controls, and untrained networks, consistent with evidence that diverse representations can explain overlapping neural variance (Schrimpf et al., 2020; Tuckute et al., 2023; Kazemian et al., 2025; Hadidi et al., 2026; Golan et al., 2026). The key question is therefore not simply whether a representation predicts cortical responses, but whether the evaluation criterion distinguishes representations when the sensory regime changes. Figure 2 provides a descriptive view of this distinction. No static stage comparison survives FDR correction, whereas under LOCO speech encoders score higher than the heterogeneous decoder/transcript-reference comparison set across all four held-out conditions. Consistently, pooled representation-class AUROC increases from 0.543 under static encoding to 0.836 under LOCO, indicating substantially stronger descriptive separation between representation classes under cross-condition evaluation. Because this analysis pools heterogeneous comparison representations and repeated model identities across ROIs, the AUROC is deliberately treated as descriptive rather than as independent inferential evidence. The primary inferential evidence therefore comes from the strict fold-local, matched-system comparisons in Section 4.2, with ROI-wise discrimination sensitivity reported in Appendix C.3.

4.2 Strict matched controls identify representation-stage selectivity and architecture-specific learning effects

The strict fold-local comparisons in Figure 3 provide the primary inferential evidence. Matched encoder–decoder contrasts hold model family fixed and therefore provide the cleanest test of representation-stage selectivity. Encoder advantages are significant for Whisper-Tiny ($+0.262$ $[0.059, 0.464]$, $q = 0.018$), Whisper-Medium ($+0.347$ $[0.190, 0.503]$, $q < 0.001$), and SpeechT5

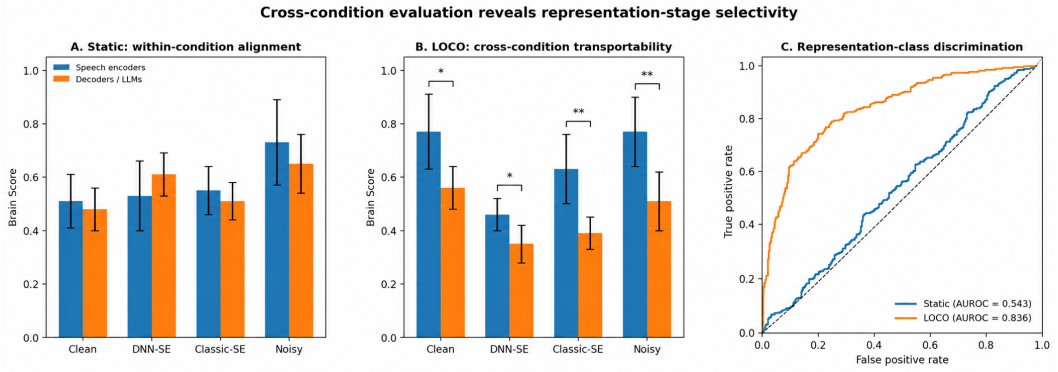


Figure 2: Cross-condition evaluation reveals selectivity that static alignment largely obscures. (A,B) Mean Brain Scores for speech encoders and the decoder/transcript-reference comparison set across the conditions; error bars indicate 95% confidence intervals across participants and statistical markers are paired tests on signed scores. (C) Descriptive pooled ROC from a single Brain Score: AUROC is 0.543 for static alignment and 0.836 for LOCO. ROI-wise sensitivity is reported in Appendix C.3. * $q < 0.05$, ** $q < 0.01$.

($+0.314$ $[0.135, 0.492]$, $q = 0.003$), whereas Whisper-Large-v3 shows the same direction but does not reach significance ($+0.111$, $q = 0.215$). The most consistent cross-condition distinction in the dataset is therefore stage-specific: encoder representations support more transportable fitted neural mappings than matched decoder representations in three of four sequence-to-sequence systems. Architecture-matched random controls address a complementary question: whether learning contributes beyond network structure. Trained Whisper-Tiny exceeds its random-weight counterpart by $+0.244$ signed GlobalLOCO Brain Score (95% CI $[0.081, 0.406]$, $q = 0.010$), whereas the positive wav2vec 2.0 contrast does not survive correction ($+0.120$ $[-0.015, 0.254]$, $q = 0.079$). The learning result is therefore narrower than the stage effect: training can improve neural transportability beyond architectural inductive bias, with corrected direct evidence here for Whisper-Tiny. For context, trained speech representations also exceed transcript-based text LLM references ($+0.306$ $[0.160, 0.452]$, $q < 0.001$), whereas contrasts with log-mel and pooled untrained controls do not survive correction.

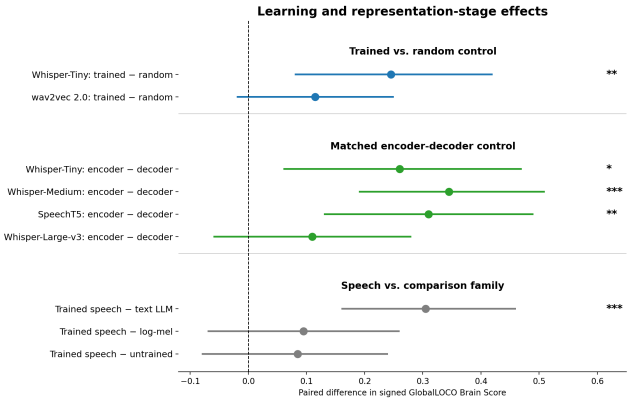


Figure 3: Strict fold-local evidence for representation-stage and learning effects. Participant-level paired differences in signed GlobalLOCO Brain Score. Matched encoder-minus-decoder effects are significant for Whisper-Tiny, Whisper-Medium, and SpeechT5; trained-minus-random effects are significant for Whisper-Tiny but not wav2vec 2.0 after correction. Broader family-level contrasts are shown for context. Points are mean paired differences; horizontal bars indicate 95% confidence intervals.

4.3 LAYER-DEPTH PROFILES LOCALIZE REPRESENTATION-STAGE SELECTIVITY

The matched contrasts establish a representation-stage difference; Figure 4 asks where that difference emerges across model depth. Across Whisper-Tiny, Whisper-Medium, and SpeechT5, signed LOCO Brain Scores are highest in encoder or intermediate representations and decline across decoder states. Within-encoder trajectories vary across architectures, so the result is not a universal rise–then–fall profile. The shared pattern is instead a transition from more transportable encoder/intermediate representations to less transportable decoder states. This localizes the stage effect without identifying its underlying mechanism.

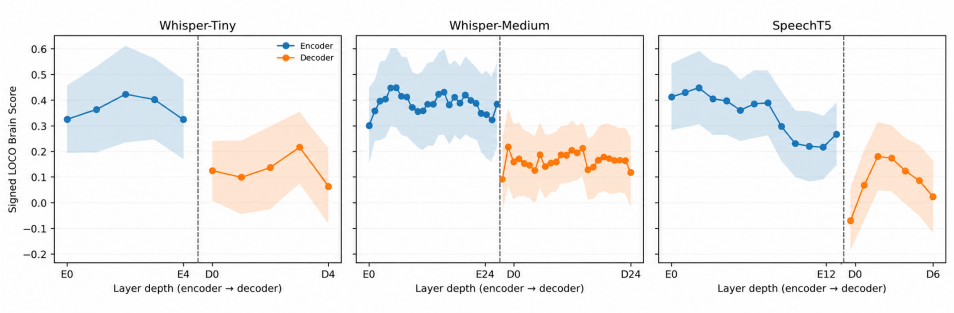


Figure 4: Layer-depth localization of cross-condition neural transportability. Signed LOCO Brain Scores are plotted across encoder and decoder depth for Whisper-Tiny, Whisper-Medium, and SpeechT5. Curves show participant means and shaded bands indicate 95% confidence intervals.

4.4 STATIC CORRESPONDENCE ONLY PARTLY PREDICTS CROSS-CONDITION TRANSPORTABILITY

Static and LOCO Brain Scores are positively but only moderately related across 238 model-stack × ROI summaries (Spearman $\rho = 0.392$; Figure 5). The association remains similar under strict fold-local layer selection that excludes the held-out condition ($\rho = 0.381$), indicating that the distinction between static correspondence and cross-condition transportability is not explained by layer-selection policy alone. Speech encoders tend to occupy higher-LOCO regions of the scatter, whereas decoder and transcript-based text LLM references overlap more under static evaluation and remain lower under cross-condition transfer. Static correspondence therefore contains information about transportability but does not determine it. Because static and LOCO differ in training-set size and ceiling estimation, the relationship is interpreted structurally rather than as a quantitative distribution-shift penalty.

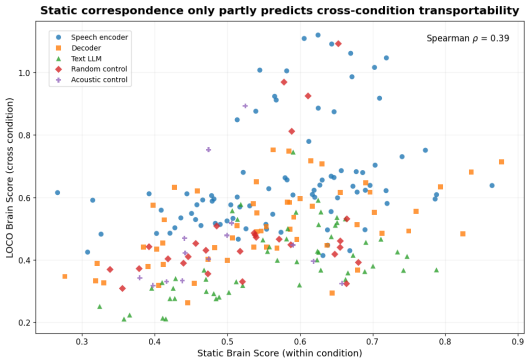


Figure 5: Static correspondence only partly predicts cross-condition transportability. Condition-general static and LOCO Brain Scores are shown across 238 model-stack × ROI summaries and colored by representation class. The association is moderate (Spearman ρ approximately 0.39). The observed scatter is descriptive because the protocols differ in training-set size and ceiling estimates.

4.5 Neural transportability varies across regimes and is distinct from trial-wise comprehension

Cross-condition transportability does not follow a simple Clean-to-degraded severity axis. Effects vary across held-out Clean, DNN-SE, Classic-SE, and Noisy conditions, although the Whisper-Tiny learning effect is not confined to a single target regime (Appendix C). Machine recoverability also depends strongly on the recognizer (Figure 6), while complementary RSA reveals qualitatively different patterns of representational reorganization across conditions rather than monotonic degradation (Appendix A), consistent with prior evidence for both noise invariance and adaptation in auditory cortex (Kell & McDermott, 2019; Khalighinejad et al., 2019). Because sentence identity is nested within condition, the supported interpretation is that different held-out regimes and item sets pose distinct challenges to the fitted neural mapping rather than defining a universal severity ordering. We next ask whether neural transportability reduces to machine recoverability or trial-wise comprehension. ASR performance tracks meaningful variation in behavioral difficulty: Whisper-Tiny CER is 0.083, 0.688, 0.875, and 0.812 across Clean, DNN-SE, Classic-SE, and Noisy, compared with 0.021, 0.483, 0.517, and 0.275 for Whisper-Large-v3. Sentence-level CER correlates with comprehension at $r = -0.543$, -0.670, and -0.632 for Whisper-Tiny, Whisper-Medium, and Whisper-Large-v3, respectively, and at $r = -0.673$ for the three-model ensemble. Despite this relationship between machine recognition and human comprehension, adding sentence-level LOCO neural fit does not improve held-out comprehension prediction beyond Condition + CER. With Whisper-Tiny CER, log loss changes from 0.4298 to 0.4302 ($\Delta = +0.00037$, 95% CI $[-0.00154, +0.00234]$); static fit, random-weight LOCO, and combined static-plus-LOCO models likewise provide no improvement. By contrast, the listener's intelligibility rating reduces log loss by -0.0402 $[-0.0591, -0.0203]$, showing that the evaluation can detect an informative trial-level predictor. Neural transportability therefore characterizes the robustness of a fitted model-to-brain relationship across conditions, but is not a direct proxy for trial-wise comprehension in this dataset.

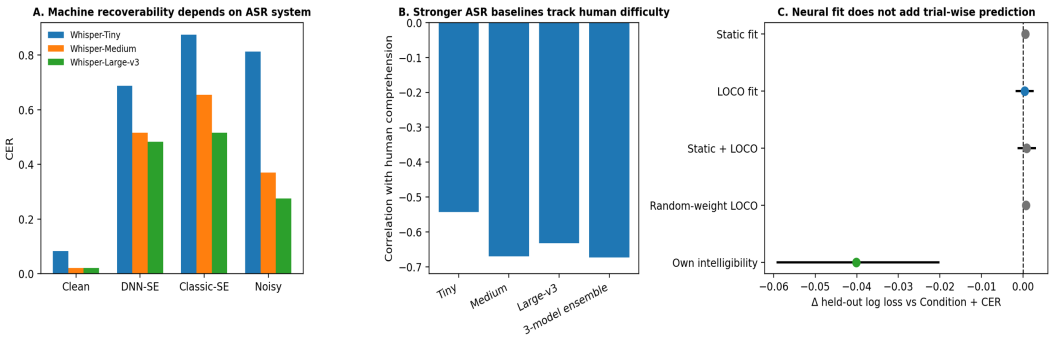


Figure 6: Machine recoverability and behavioral relevance define a boundary for neural transportability. (A) CER across Clean, DNN-SE, Classic-SE, and Noisy for Whisper-Tiny, Whisper-Medium, and Whisper-Large-v3, showing recognizer-dependent apparent difficulty. (B) Sentence-level correlations between CER and human comprehension. (C) Change in held-out comprehension log loss relative to Condition + CER. LOCO fit does not improve prediction; static fit and random-weight LOCO show small degradations, whereas the listener's own intelligibility rating produces a clear improvement. Negative Δ log loss indicates improvement.

5 Discussion and Conclusion

5.1 Cross-condition transportability complements static brain–model alignment

Within-condition neural correspondence and cross-condition transportability are related but non-equivalent. Static encoding asks whether a representation predicts cortical responses within the fitted regime; LOCO asks whether the same fitted relationship survives an unseen condition. Their moderate association, stronger discrimination under LOCO, and strict matched comparisons show that strong static alignment does not guarantee a transportable model-to-brain mapping.

Our contribution is therefore an evaluation principle rather than a new encoding architecture. LOCO itself is standard cross-validation; here, it tests the stability of a fitted brain–model relationship under condition shift. This extends OOD neural evaluation in vision (Ren & Bashivan, 2024; Gifford et al., 2026), model-disagreement approaches (Golan et al., 2026), and broader calls for stronger computational tests (Schmitt et al., 2026; Serre & Pavlick, 2025) to speech. Successful transfer indicates mapping stability under the evaluated shift, not equivalence between artificial and biological computation.

5.2 Representation stage shows the most consistent matched pattern

Representation stage provides the most consistent matched result. Encoders are more transportable than matched decoders in Whisper-Tiny, Whisper-Medium, and SpeechT5, while Whisper-Large-v3 shows the same direction without reaching significance. Layer-depth profiles localize this distinction to encoder and intermediate representations before transportability declines across decoder states. Although within-encoder trajectories vary across architectures, the shared pattern is a reproducible stage-level difference. This is consistent with intermediate speech representations retaining information useful for stable cortical mapping while later autoregressive states specialize for generation (Goldstein et al., 2025), although the mechanism remains unresolved.

Learning effects are narrower. Trained Whisper-Tiny exceeds its architecture-matched random control, whereas the wav2vec 2.0 contrast does not survive correction. Learned structure can therefore improve transportability beyond architectural inductive bias, but current evidence is architecture-specific. Representation stage is thus the more reproducible source of transportability selectivity in this dataset.

5.3 Neural transportability defines a complementary evaluation axis

Transportability also varies across held-out Clean, Noisy, DNN-SE, and Classic-SE conditions without following a simple severity ordering. Machine recoverability depends on the recognizer, while complementary RSA shows qualitatively different representational reorganization across conditions (Appendix A), consistent with cortical noise invariance and adaptation (Kell & McDermott, 2019; Khalighinejad et al., 2019). Because sentence identity is nested within condition, these differences reflect joint transfer across acoustic regime and item set rather than isolated acoustic effects.

Behavior further constrains the interpretation of LOCO. Stronger ASR systems track comprehension more closely, yet sentence-level LOCO fit does not improve held-out comprehension prediction beyond Condition + CER, whereas the listener's intelligibility rating does. Neural transportability is therefore not a direct proxy for comprehension, just as recognition does not establish neural correspondence. More broadly, machine recognition, within-condition neural correspondence, cross-condition transportability, and behavioral relevance probe related but non-equivalent properties of representations (Schrimpf et al., 2020; Feather et al., 2025).

Taken together, cross-condition neural transportability complements conventional alignment by testing whether a fitted brain–model relationship remains predictive when the sensory regime changes. It reveals representation-stage selectivity and architecture-specific learning effects not reducible to static fit, recognition, or trial-wise comprehension.

5.4 Limitations

Several limitations constrain these conclusions. Sentence identity is nested within condition, so LOCO jointly changes acoustic regime and item set rather than isolating a pure acoustic OOD effect; with only 24 held-out sentences per fold, item-level uncertainty remains despite participant-level inference over $N = 25$. Static and LOCO also use different training-set sizes, so absolute Brain Score differences are not interpreted as pure distribution-shift penalties. Architecture-matched random controls are available only for wav2vec 2.0 and Whisper-Tiny, limiting general claims about learning. Encoder and decoder representations use different extraction conventions, which may contribute to stage differences. CER robustness relies on Whisper-family recognizers, and the behavioral analysis contains only 96 sentence identities with a binary outcome. The absence of incremental behavioral prediction from LOCO should therefore be treated as a boundary of this dataset rather than evidence that transportability is generally unrelated to behavior.

AI USE STATEMENT

We used generative AI tools for implementing analysis code, dataset cleaning, and translation, as well as for language editing, manuscript organization, drafting parts of the paper, and retrieving related work. No other required-disclosure tasks under the ICLR 2027 AI policy involved AI assistance. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

The study used de-identified human fMRI data collected with written informed consent under institutional ethics approval. No identifying participant information is reported.

REPRODUCIBILITY STATEMENT

The main paper specifies the LOCO folds, inferential units, matched controls, discrimination analysis, and layer-selection policies. Appendix B provides the full model inventory and detailed static/LOCO maps; Appendix C reports layer-policy, fold-specific, and ROI-wise discrimination sensitivities; Appendix G reports ASR and behavioral robustness. These materials make the inferential path from descriptive benchmark to strict matched claim explicit.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 03 2025.

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001. doi: 10.1145/375551.375608.

Junyi Ao, Rui Wang, Long Zhou, et al. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.393. URL https://aclanthology.org/2022.acl-long.393/.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 01 1995. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Matthew Brett, Jean-Luc Anton, Romain Valabregue, et al. Region of interest analysis using an spm toolbox. In *8th international conference on functional mapping of the human brain*, volume 16, pp. 497. Sendai, Japan, 2002.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5:134, 2022. doi: 10.1038/s42003-022-03036-1.

Rong Chao, Wen-Huang Cheng, Moreno La Quatra, et al. An investigation of incorporating mamba for speech enhancement. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 302–308, 2024. doi: 10.1109/SLT61566.2024.10832332.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. 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. doi: 10.1109/JSTSP.2022.3188113.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, et al. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15:9383, 2024. doi: 10.1038/s41467-024-53147-y.

Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32 (6):1109–1121, 1984. doi: 10.1109/TASSP.1984.1164453.

Jenelle Feather, Meenakshi Khosla, N. Apurva Ratan Murty, et al. Brain-model evaluations need the neuroai turing test, 2025. URL https://arxiv.org/abs/2502.16238.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, et al. New method for fMRI investigations of language: Defining rois functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194, 2010. doi: 10.1152/jn.00032.2010. PMID: 20410363.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25:289–312, 2024. doi: 10.1038/s41583-024-00802-4.

Alessandro T. Gifford, Maya A. Jastrzębowska, Johannes J. D. Singer, et al. In silico discovery of representational relationships across visual cortex. *Nature Human Behaviour*, 9:2079–2098, 2025. doi: 10.1038/s41562-025-02252-z.

Alessandro T. Gifford, Radoslaw M. Cichy, Thomas Naselaris, et al. A 7T fmri dataset of synthetic images for out-of-distribution modeling of vision. *Nature Communications*, 17:1589, 2026. doi: 10.1038/s41467-026-69345-9.

Tal Golan, Heiko H. Schütt, and Nikolaus Kriegeskorte. Making models disagree to learn how brains compute. *Nature Reviews Neuroscience*, 2026. doi: 10.1038/s41583-026-01070-0.

Ariel Goldstein, Haocheng Wang, Leonard Niekerken, et al. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, 9:1041–1055, 2025. doi: 10.1038/s41562-025-02105-9.

Nima Hadidi, Ebrahim Feghhi, Bryan H. Song, et al. Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. *Nature Communications*, 17:5769, 2026. doi: 10.1038/s41467-026-72253-7.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Atlas Kazemian, Eric Elmoznino, and Michael F. Bonner. Convolutional architectures are cortex-aligned de novo. *Nature Machine Intelligence*, 7:1834–1844, 2025. doi: 10.1038/s42256-025-01142-3.

Alexander J. E. Kell and Josh H. McDermott. Invariance to background noise as a signature of non-primary auditory cortex. *Nature Communications*, 10:3958, 2019. doi: 10.1038/s41467-019-11710-y.

Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, et al. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16, 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.03.044. URL https://www.sciencedirect.com/science/article/pii/S0896627318302502.

Menoua Keshishian, Gavin Mischler, Samuel Thomas, et al. Parallel hierarchical encoding of linguistic representations in the human auditory cortex and recurrent automatic speech recognition systems. *Nature Machine Intelligence*, 8:257–269, 2026. doi: 10.1038/s42256-026-01185-0.

Bahar Khalighinejad, Jose L. Herrero, Ashesh D. Mehta, et al. Adaptation of the human auditory cortex to changing background noise. *Nature Communications*, 10:2509, 2019. doi: 10.1038/s41467-019-10611-4.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, et al. Toward a realistic model of speech processing in the brain with self-supervised learning. In *NeurIPS*, volume 35, pp. 33428–33443, 2022. doi: 10.52202/068431-2422.

Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002. doi: 10.1002/hbm.1058.

Hamed Nili, Cai Wingfield, Alexander Walther, et al. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10(4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November 2011. ISSN 1532-4435.

Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Yifei Ren and Pouya Bashivan. How well do models of visual cortex generalize to out of distribution samples? *PLOS Computational Biology*, 20(5):e1011145, 2024. doi: 10.1371/journal.pcbi.1011145.

Lea-Maria Schmitt, Dota Tianai Dong, Emin Çelik, et al. A decade of comparing brains to dnns: Progress and perspectives. *Trends in Cognitive Sciences*, 2026. ISSN 1364-6613. doi: 10.1016/j.tics.2026.06.011. URL https://www.sciencedirect.com/science/article/pii/S1364661326001385.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, et al. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.07.040.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, et al. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

Thomas Serre and Ellie Pavlick. From prediction to understanding: Will ai foundation models transform brain science? *Neuron*, 113(21):3504–3508, 2025. ISSN 0896-6273. doi: 10.1016/j.neuron.2025.09.039.

Greta Tuckute, Jenelle Feather, Dana Boebinger, et al. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):e3002366, 2023. doi: 10.1371/journal.pbio.3002366.

Greta Tuckute, Aalok Sathe, Shashank Srikant, et al. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8:544–561, 2024. doi: 10.1038/s41562-023-01783-7.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, et al. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2014.01.060.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

A MATCHED-LAYER REPRESENTATIONAL REORGANIZATION

The main LOCO analysis evaluates whether a fitted model-to-brain mapping remains predictive across a held-out condition. Representational similarity analysis (RSA) addresses a complementary question: whether model-brain representational geometry itself changes across conditions. These properties are related but not equivalent. A representation can reorganize geometrically while still supporting a transportable predictive mapping, or remain geometrically similar while its fitted mapping transfers poorly. We therefore use matched-layer ΔRSA to characterize condition-dependent representational reorganization separately from LOCO neural transportability.

A.1 REPRESENTATIONAL SIMILARITY ANALYSIS AND MATCHED-LAYER ΔRSA

A.1.1 RDM CONSTRUCTION

For each condition, sentence-level model representations and multivoxel fMRI response patterns were converted into representational dissimilarity matrices (RDMs). For two sentences i and j, dissimilarity was defined as Pearson correlation distance:

$$d(i, j) = 1 - \mathrm{corr}_{\mathrm{P}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\,. \tag{A1}$$

With 24 sentences, each condition yields a symmetric 24×24 RDM containing $24 \times 23/2 = 276$ unique off-diagonal sentence pairs. Brain RDMs were constructed separately for each participant, ROI, and condition; model RDMs were constructed separately for each model layer and condition (Kriegeskorte et al., 2008; Nili et al., 2014).

A.1.2 LAYER-WISE MODEL-BRAIN RSA

For participant s, model m, layer l, ROI r, and condition c, model-brain correspondence was quantified as the Spearman correlation between the unique upper-triangle entries of the corresponding model and brain RDMs:

$$\mathrm{RSA}(s, m, l, r, c) = \rho_S\left[\mathrm{vec}_u(\mathrm{RDM}_{\text{model}}(m, l, c)), \mathrm{vec}_u(\mathrm{RDM}_{\text{brain}}(s, r, c))\right] \tag{A2}$$

Here vec_u extracts the 276 unique off-diagonal entries. RSA therefore measures whether a computational representation and a cortical ROI impose a similar geometry over the sentence set without fitting a model-to-brain encoding map. In the Clean-Noisy analysis, text LLMs received identical ground-truth transcripts across conditions and therefore supplied condition-invariant linguistic reference geometries.

A.1.3 MATCHED-LAYER ΔRSA

Condition effects were evaluated at the same computational layer, avoiding the confound introduced by selecting different peak layers in different conditions. For target condition c:

$$\Delta\mathrm{RSA}(s, m, l, r, c) = \mathrm{RSA}(s, m, l, r, c) - \mathrm{RSA}(s, m, l, r, \mathrm{Clean}) \tag{A3}$$

Positive ΔRSA indicates stronger model-brain correspondence in the target condition than in Clean; negative ΔRSA indicates stronger correspondence in Clean. DNN-SE and Classic-SE are retained as sensitivity analyses; because sentence identity is fixed to condition in the four-condition experiment, these contrasts include both condition and item-set variation.

A.1.4 STATISTICAL INFERENCE

The 25 participant-level ΔRSA values, not the 276 RDM entries, are the independent inferential observations. For each fixed hypothesis, a two-sided participant-level sign-flip permutation test recomputed the one-sample t statistic after independently multiplying each participant effect by $+1$ or -1 (Nichols & Holmes, 2002; Winkler et al., 2014). The permutation p value is the proportion satisfying $|t_{\text{perm}}| \geq |t_{\text{obs}}|$. BH-FDR is then applied within the prespecified family (Benjamini & Hochberg, 1995). The validated Clean-Noisy analysis uses layer-wise inference with both panel-wise and whole-search-space correction. DNN-SE and Classic-SE use the available mean-over-layers and peak-group summary families and are reported separately.

A.2 CLEAN → NOISY: REPRESENTATIONAL REORGANIZATION

The validated Clean-Noisy analysis identified 89 layer-specific effects after panel-wise BH-FDR correction, comprising 67 positive and 22 negative ΔRSA effects. Positive effects were concentrated in learned speech representations and predominantly in the left hemisphere, whereas most negative effects involved text LLM reference spaces and right-hemisphere regions. A single BH-FDR correction across the complete model × ROI × layer search space retained three effects: increased Whisper-Tiny correspondence in left anterior and posterior temporal cortex and decreased Qwen2.5-3B correspondence in right anterior temporal cortex. These globally corrected effects show that acoustic degradation does not simply reduce model-brain similarity; it can increase correspondence with selected learned speech representations while decreasing correspondence with a fixed linguistic reference geometry.

Table A1: Clean-Noisy effects surviving whole-search-space BH-FDR correction of participant-level sign-flip permutation p values.

ROI	Model-layer	Mean ΔRSA	t_{obs}	p_{perm}	q_{global}
L_{AntTemp}	Whisper-Tiny Enc2	$+0.0860$	5.42	1.43×10^{-5}	0.0228
L_{PostTemp}	Whisper-Tiny Enc4	$+0.1112$	5.92	3.95×10^{-6}	0.0126
R_{AntTemp}	Qwen2.5-3B L12	-0.0489	-5.24	2.29×10^{-5}	0.0244

A.3 CLEAN → DNN-SE: SELECTIVE TEMPORAL GAINS AND HIGHER-LEVEL LOSSES

The Clean-DNN-SE sensitivity analysis shows a mixed pattern. Four peak-group effects are positive and localized to temporal cortex, involving log-mel and wav2vec 2.0 representations. Two mean-over-layers effects are negative, involving Phi-4-mini in left posterior temporal cortex and Mistral-7B in right angular gyrus. The supported conclusion is selective rather than global: DNN-SE is associated with localized increases in acoustic/intermediate-speech correspondence in temporal cortex together with decreases in selected higher-level linguistic representations.

Table A2: Clean-DNN-SE effects surviving global BH-FDR within the available summary families. Peak-group and mean-over-layers effects belong to different inferential families and should not be treated as equivalent layer-wise tests.

Inference family	ROI	Model-layer	Mean ΔRSA	t_{obs}	p_{perm}	q_{global}
Peak-group	L_{AntTemp}	log-mel	$+0.0711$	3.84	1.12×10^{-3}	0.0390
Peak-group	L_{PostTemp}	log-mel	$+0.0944$	4.35	8.00×10^{-5}	0.0049
Peak-group	L_{PostTemp}	wav2vec 2.0 Enc9	$+0.0836$	4.59	1.05×10^{-4}	0.0049
Peak-group	R_{PostTemp}	log-mel	$+0.1063$	4.56	1.05×10^{-4}	0.0049
Mean layers	L_{PostTemp}	Phi-4-mini	-0.0291	-3.78	7.00×10^{-4}	0.0490
Mean layers	R_{Angular}	Mistral-7B	-0.0407	-3.84	6.35×10^{-4}	0.0490

A.4 CLEAN → CLASSIC-SE: PREDOMINANTLY REDUCED MODEL-BRAIN CORRESPONDENCE

Clean-Classic-SE shows a different pattern. Thirteen mean-over-layers effects survive global BH-FDR correction, 12 of which are negative. These reductions span acoustic, self-supervised speech, ASR, and text representations and are concentrated in auditory and temporal regions. The only positive effect is Phi-4-mini in right orbital IFG. No peak-group effect survives the corresponding global correction. Unlike the bidirectional Clean-Noisy result and the mixed Clean-DNN-SE pattern, Classic-SE is characterized primarily by broad reductions in model-brain representational correspondence.

Table A3: Clean-Classic-SE mean-over-layers effects surviving global BH-FDR correction.

ROI	**Model-layer**	**Mean ΔRSA**	t_{obs}	p_{perm}	q_{global}
L_{Heschl}	log-mel	-0.0596	-3.49	1.28×10^{-3}	0.0329
L_{Heschl}	wav2vec 2.0	-0.0531	-3.31	2.26×10^{-3}	0.0396
L_{Heschl}	WavLM	-0.0720	-3.68	7.25×10^{-4}	0.0329
L_{Heschl}	Whisper-Tiny	-0.0389	-3.21	3.61×10^{-3}	0.0447
L_{Heschl}	Whisper-Medium	-0.0480	-3.54	1.41×10^{-3}	0.0329
L_{AntTemp}	Qwen2.5-3B	-0.0431	-3.72	7.75×10^{-4}	0.0329
L_{PostTemp}	log-mel	-0.0627	-3.61	1.24×10^{-3}	0.0329
L_{PostTemp}	Qwen2.5-3B	-0.0381	-3.43	2.55×10^{-3}	0.0396
R_{Heschl}	log-mel	-0.0401	-3.05	3.68×10^{-3}	0.0447
R_{PostTemp}	log-mel	-0.0454	-3.18	3.83×10^{-3}	0.0447
R_{PostTemp}	wav2vec 2.0	-0.0597	-3.61	1.79×10^{-3}	0.0357
R_{IFGorb}	Phi-4-mini	$+0.0308$	3.77	1.24×10^{-3}	0.0329
R_{IFG}	log-mel	-0.0533	-3.15	4.40×10^{-3}	0.0474

A.5 RELATION TO CROSS-CONDITION TRANSPORTABILITY

The three RSA contrasts do not support a single monotonic degradation axis. Clean-Noisy shows a structured bidirectional redistribution, DNN-SE combines localized temporal gains with selected higher-level losses, and Classic-SE is dominated by broader reductions in correspondence. These findings are complementary to, but distinct from, LOCO. RSA asks how model-brain representational geometry changes across conditions; LOCO asks whether a predictive mapping fitted in other conditions remains valid in the held-out condition. Positive ΔRSA does not imply stronger transportability, and negative ΔRSA does not imply mapping failure. Appendix A therefore provides representational context for regime-dependent reorganization, while the principal contribution remains cross-condition neural transportability.

B EXPERIMENTAL INVENTORY AND DETAILED MAPS

This appendix provides the complete experimental inventory underlying the static and LOCO analyses. It specifies the four train-three/test-one folds and the computational representation panel, then reports the full within-condition and cross-condition maps used as descriptive context for the stricter participant-level contrasts in the main text.

B.1 LEAVE-ONE-CONDITION-OUT PROTOCOL AND MODEL PANEL

LOCO uses four outer folds. In each fold, all 72 sentences from three conditions are used to fit the model-to-brain mapping, while all 24 sentences from the fourth condition are held out for final evaluation. Train-derived standardization, dimensionality control, ridge tuning, and strict fold-local layer selection are completed entirely on the training side before the target condition is scored.

Table B1: Leave-one-condition-out (LOCO) folds.

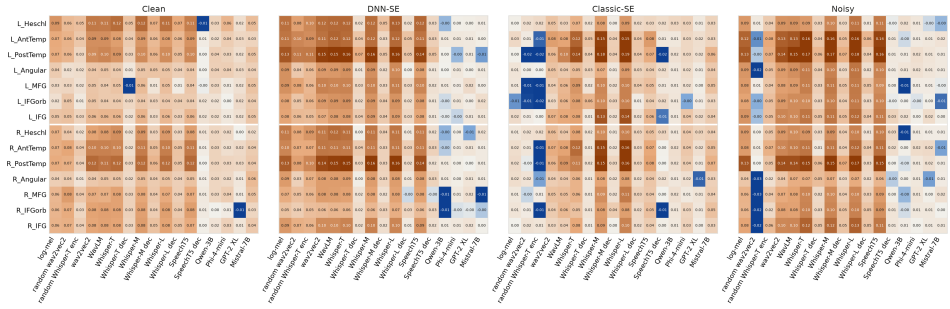
Fold	Training conditions (72 sentences)	Held-out condition (24 sentences)
1	Clean + DNN-SE + Classic-SE	Noisy
2	Clean + DNN-SE + Noisy	Classic-SE
3	Clean + Classic-SE + Noisy	DNN-SE
4	DNN-SE + Classic-SE + Noisy	Clean

Table B2: Computational representation panel used for static and LOCO analyses.

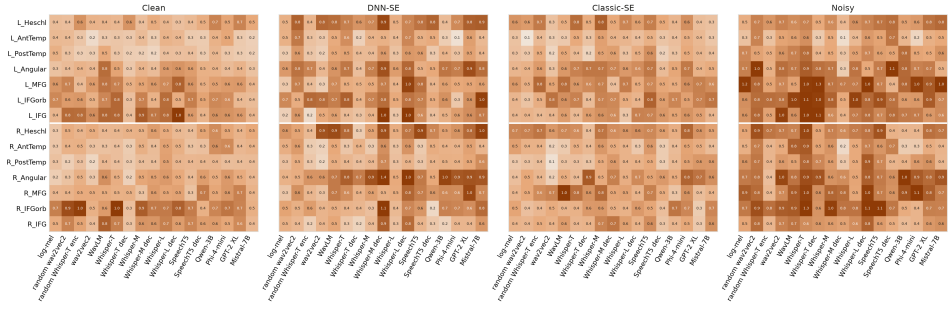
Family	Representations	Input	Primary role
Acoustic	log-mel	Waveform	Low-level acoustic control
Untrained	Random wav2vec 2.0 random Whisper-Tiny encoder	Waveform	Architecture-matched controls
SSL speech	wav2vec 2.0 WavLM	Waveform	Learned speech encoders
ASR/speech	Whisper-Tiny Whisper-Medium Whisper-Large-v3 SpeechT5 encoders + decoders	Waveform	Matched stage comparisons
Text LLM	Qwen2.5-3B Phi-4-mini GPT-2 XL Mistral-7B	Ground-truth transcript	Condition-invariant linguistic reference spaces

B.2 WITHIN-CONDITION BRAIN SCORE AND RSA COMPARISON

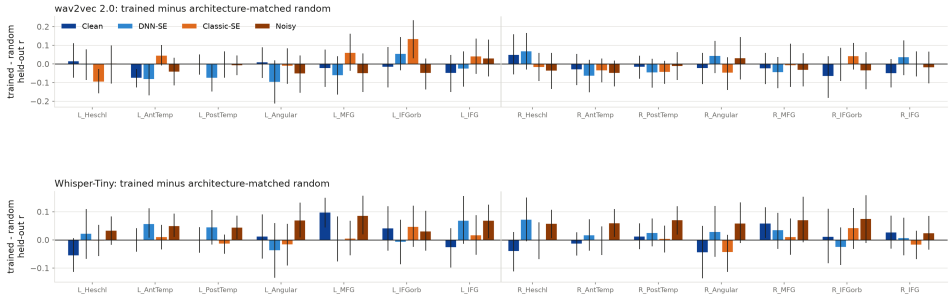
Static encoding evaluates each condition separately using four run-wise folds (18 training sentences and 6 held-out sentences per fold). The detailed map juxtaposes layer-selected RSA with ceiling-normalized static Brain Score across all 14 ROIs and representation families. Architecture-matched trained-minus-random contrasts are shown for wav2vec 2.0 and Whisper-Tiny. These maps are descriptive: negative normalized participant values may be clipped only for visualization, whereas all inferential contrasts use signed participant-level scores.



(A) RSA, best layer per parcel and condition



(B) Static Brain Score

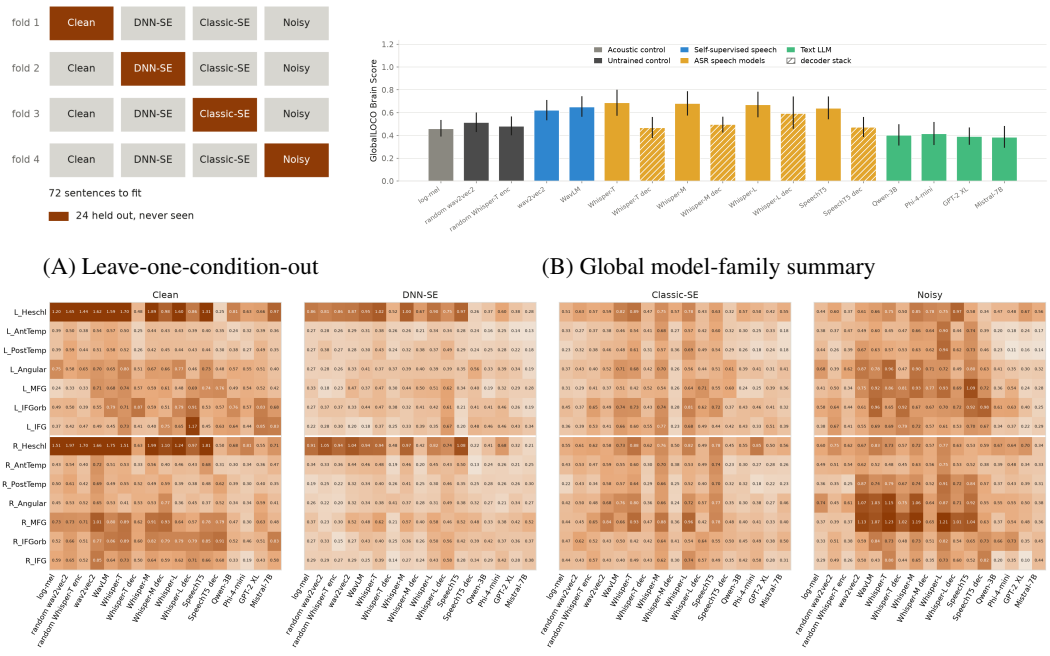


(C) Trained vs. Random contrasts

Figure B1: Detailed within-condition correspondence across Clean, DNN-SE, Classic-SE, and Noisy speech. Top: model-brain RSA. Middle: static ceiling-normalized Brain Score. Bottom: architecture-matched trained-minus-random contrasts.

B.3 CROSS-CONDITION LOCO MAP

The LOCO map summarizes cross-condition neural predictivity across the same 14 ROIs and representation panel. The global panel averages the four held-out conditions after fitting on the remaining three conditions; the condition-specific heatmaps show the held-out-fold pattern. The static-versus-LOCO relationship is descriptive because the protocols differ in training-set size and ceiling estimation. Consequently, absolute displacement is not interpreted as a pure OOD penalty. Held-out-condition summaries likewise do not imply a causal severity ordering because sentence identity is fixed to condition.



(A) Leave-one-condition-out

(B) Global model-family summary

(C) LOCO Brain Score maps

Figure B2: (A) Neural mappings are fit on three conditions (72 sentences) and tested on the held-out condition (24 sentences). (B) Global cross-condition Brain Scores across model families, with colors indicating representation families and hatching indicating decoder representations. (C) Held-out-condition-specific cross-condition Brain Score maps across 14 ROIs and model representations. Together, the panels summarize global and region-specific cross-condition neural generalization.

C LAYER-POLICY AND FOLD-SPECIFIC ROBUSTNESS

Layer choice can materially affect brain-model comparisons, particularly when model stacks differ in depth. We therefore separate descriptive layer summaries from the strict fold-local policy that supports the main inferential claims and test whether the broad cross-condition pattern depends on a single layer convention.

C.1 LAYER-POLICY SENSITIVITY

Seven descriptive policies are reported: (i) condition-general, selecting one layer from internal-validation performance averaged across the four conditions; (ii) Clean-anchor, selecting the layer preferred under Clean and applying it unchanged elsewhere; (iii) condition-specific, selecting a separate best layer for each condition and therefore remaining descriptive rather than transport-safe; (iv) final layer; and (v–vii) early, middle, and late summaries averaging layers within coarse depth thirds. Encoder and decoder stacks are treated separately whenever both are available.

The strict policy is different. For each outer LOCO fold and model stack, the selected layer maximizes mean inner-validation raw neural predictivity using only the three training conditions. The held-out condition contributes no information to layer selection. The layer is shared across all 14 ROIs within the stack, while encoder and decoder stacks are selected independently. This fold-local policy prevents target-condition leakage and defines the narrowest matched claims in the main text.

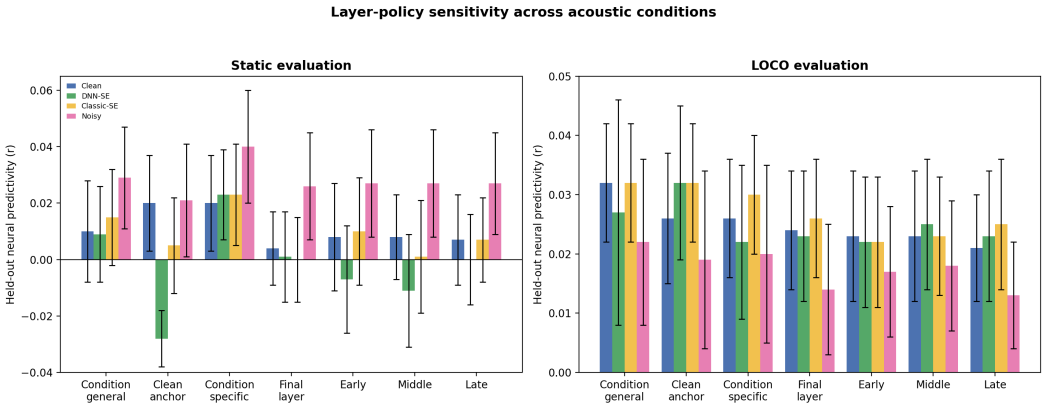


Figure C1: Layer-policy sensitivity. Alternative layer summaries preserve the broad cross-condition pattern. These summaries are descriptive; strict fold-local layer selection is used for the conservative inferential claims.

Table C1: Strict fold-local layers selected from training-side inner validation. Each column names the held-out condition; the layer shown was selected using only the other three conditions. Encoder and decoder stacks are selected separately and the chosen layer is shared across ROIs.

Model stack	Clean	DNN-SE	Classic-SE	Noisy
(encoder) log-mel	Enc0	Enc0	Enc0	Enc0
(encoder) random wav2vec2	Enc10	Enc2	Enc9	Enc7
(encoder) random Whisper-T enc	Enc1	Enc1	Enc0	Enc0
(encoder) wav2vec2	Enc2	Enc9	Enc1	Enc4
(encoder) WavLM	Enc7	Enc7	Enc8	Enc5
(encoder) Whisper-T	Enc1	Enc1	Enc2	Enc2
(decoder) Whisper-T	Dec0	Dec2	Dec3	Dec2
(encoder) Whisper-M	Enc2	Enc4	Enc4	Enc4
(decoder) Whisper-M	Dec0	Dec15	Dec13	Dec20
(encoder) Whisper-L	Enc24	Enc8	Enc31	Enc24
(decoder) Whisper-L	Dec30	Dec32	Dec30	Dec21
(encoder) SpeechT5	Enc3	Enc3	Enc4	Enc6
(decoder) SpeechT5	Dec5	Dec3	Dec2	Dec5
(LLM) Qwen-3B	L24	L1	L1	L35
(LLM) Phi-4-mini	L31	L3	L3	L25
(LLM) GPT-2 XL	L48	L0	L48	L48
(LLM) Mistral-7B	L19	L0	L0	L19

C.2 Fold-specific trained-minus-random effects

To assess whether the learning effect is driven by a single target condition, participant-level trained-minus-random differences are displayed separately for each held-out fold. Fold-specific panels are descriptive because the confirmatory learning test averages the four folds within participant before group-level inference.

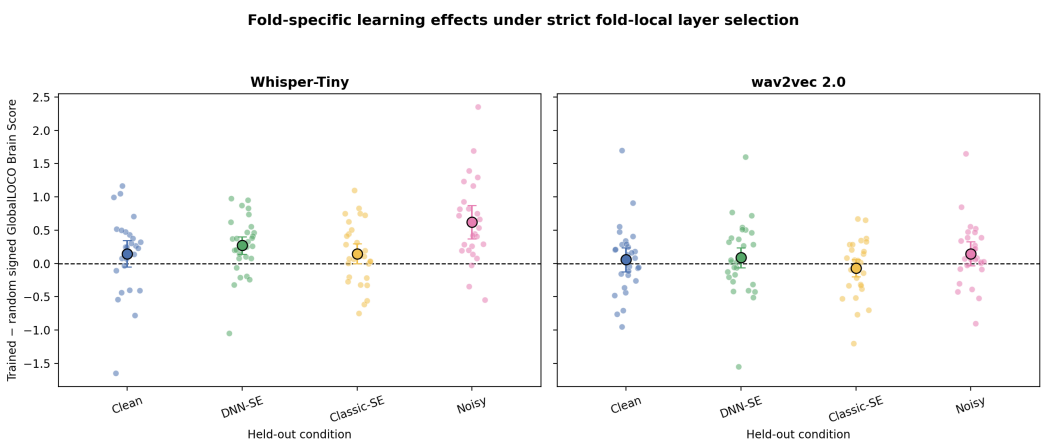


Figure C2: Fold-specific paired trained-minus-random effects under strict fold-local layer selection. Small points are participant-level differences; large points and bars show means and 95% confidence intervals. Fold-specific panels are descriptive; confirmatory inference averages across folds.

C.3 ROI-WISE REPRESENTATION-CLASS DISCRIMINATION SENSITIVITY

The main-text representation-class ROC pools model-stack × ROI summaries and is descriptive because the same representation appears repeatedly across ROIs. As a sensitivity analysis, encoder-versus-decoder/transcript-reference AUROC is computed separately within each of the 14 ROIs, and static and LOCO discrimination are compared within the same region. Mean ROI-wise AUROC increases from approximately 0.59 under static alignment to 0.91 under LOCO, with most ROIs showing stronger discrimination under cross-condition evaluation. This supports the same qualitative conclusion without treating ROI observations as independent model samples.

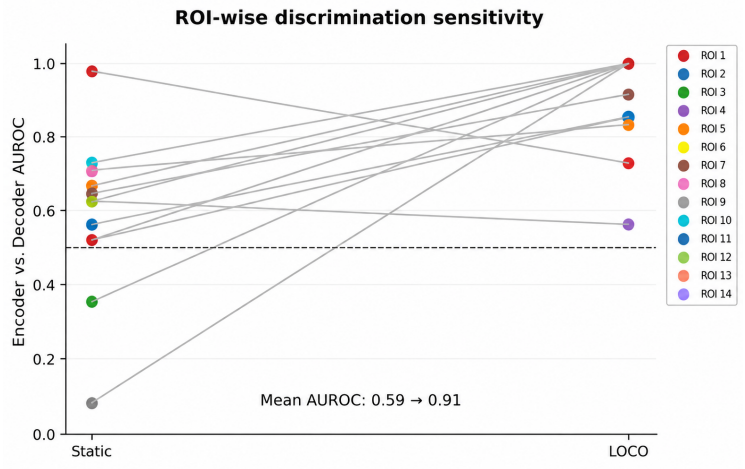


Figure C3: ROI-wise representation-class discrimination sensitivity. AUROC is computed separately within each of the 14 bilateral ROIs; lines connect the same ROI under static and LOCO evaluation. The dashed line marks chance performance (AUROC = 0.5).

C.4 SUMMARY

The broad conclusions are preserved across alternative layer-selection policies. Strict fold-local selection excludes the held-out condition from layer choice, fold-specific analyses indicate that the learning result is not driven by one target regime, and ROI-wise sensitivity shows stronger representation-class discrimination under LOCO than static alignment. These analyses support the

interpretation that cross-condition transportability is not an artifact of one layer-selection convention, one ROI aggregation, or one held-out fold.

D EXPERIMENTAL TASK AND FMRI IMPLEMENTATION

D.1 PARTICIPANTS, STIMULI, AND EXPERIMENT

Twenty-five healthy native Mandarin speakers with self-reported normal hearing completed the fMRI experiment. The full experiment comprised 96 approximately 10-word sentences. Twenty-four were presented as Clean speech; the remaining 72 were distributed across 24 Noisy trials, 24 DNN-SE trials, and 24 Classic-SE trials. Noisy stimuli were generated by mixing each sentence with stationary speech-shaped noise at $-3\,\mathrm{dB}$ SNR. DNN-SE used SEMamba, whereas Classic-SE used the minimum mean-square error estimator of Ephraim & Malah (1984). Stimuli were delivered through MRI-compatible headphones.

Trials were organized into four 24-trial runs using Latin-square counterbalancing across runs. Each trial comprised a $1\,\mathrm{s}$ fixation, a $3\,\mathrm{s}$ sentence, a brief fixation interval, and a $3\,\mathrm{s}$ response period. After each sentence, participants answered a two-alternative comprehension question scored as correct or incorrect and rated intelligibility on five-point mean-opinion-score scales (1 = bad, 5 = excellent). Sentence identity was fixed to condition rather than counterbalanced; this design feature is therefore carried forward explicitly in the interpretation of condition and LOCO effects.

D.2 MRI ACQUISITION AND PREPROCESSING

MRI data were acquired on a Siemens Magnetom Skyra 3T scanner. T1-weighted MP-RAGE anatomy used 1-mm isotropic voxels. Functional gradient-echo EPI used TR $= 2000\,\mathrm{ms}$, TE $=$ $24\,\mathrm{ms}$, FOV $= 220 \times 220\,\mathrm{mm}$, 38 slices, $3.4 \times 3.4 \times 4.0\,\mathrm{mm}$ voxels, and a 90-degree flip angle. SPM12 preprocessing comprised slice-timing correction, motion correction, functional-anatomical coregistration, normalization to MNI space, and 4-mm FWHM spatial smoothing.

D.3 ROI DEFINITION AND SENTENCE-SPECIFIC RESPONSE PATTERNS

Language-network ROIs comprised bilateral anterior temporal cortex, posterior temporal cortex, angular gyrus, inferior frontal cortex, orbital inferior frontal cortex, and middle frontal gyrus, following the functional-anatomical language-network framework (Fedorenko et al., 2010; 2024). Bilateral Heschl's gyri were defined anatomically and extracted using MarsBaR (Brett et al., 2002), yielding 14 ROIs across seven bilateral auditory-language regions. ROI masks were fixed before model comparison.

The first-level GLM contained one regressor for each of the 96 sentence trials, convolved with the canonical hemodynamic response function; six head-motion parameters were included as nuisance covariates. Each trial beta map supplied one sentence-specific activation pattern. Within each ROI, voxel-wise beta estimates formed one multivoxel response vector per sentence and participant. These vectors served as neural targets for encoding and as brain representations for RSA.

E STATISTICAL INFERENCE AND MULTIPLICITY

E.1 PARTICIPANT-LEVEL NEURAL INFERENCE

Participants, not voxels, layers, sentence pairs, or ROIs, are the independent inferential units for group-level neural comparisons. For each prespecified contrast, the relevant score is first aggregated within participant according to the analysis definition and the paired participant-level difference is then tested against zero. Reported confidence intervals are 95% intervals over participant-level paired effects. Brain Score visualizations may floor negative normalized participant scores at zero, but all inferential contrasts retain signed values.

E.2 PRESPECIFIED FDR FAMILIES FOR ENCODING ANALYSES

Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995) is applied separately within prespecified hypothesis families rather than across every descriptive panel. The principal families are: (i) trained versus architecture-matched random controls, (ii) encoder versus decoder within matched sequence-to-sequence systems, and (iii) trained-speech family contrasts against comparison representation families. The four held-out-regime stage comparisons in Figure 2 form a secondary family. Fold-specific trained-minus-random panels are descriptive and are not treated as four independent confirmatory replications. Stimulus-feature permutation controls preserve the fold and ridge pipeline while breaking stimulus-feature correspondence.

E.3 RSA SIGN-FLIP INFERENCE AND CORRECTION

For RSA, let $\delta(s, h)$ denote participant s's matched-layer ΔRSA for fixed hypothesis $h = (\text{model}, \text{layer}, \text{ROI}, \text{contrast})$. The observed statistic is the one-sample t statistic across the 25 participant-level effects. Under the null hypothesis of symmetry around zero, each participant effect is independently multiplied by $+1$ or -1 and the statistic is recomputed; the two-sided permutation p value is the proportion satisfying $|t_{\text{perm}}| \geq |t_{\text{obs}}|$ (Nichols & Holmes, 2002; Winkler et al., 2014). Thus, the 25 participants — not the 276 RDM pair entries — are exchangeable units. The Clean-Noisy analysis applies BH-FDR across layers within each prespecified model $\times$ ROI panel and additionally reports correction across the complete model $\times$ ROI $\times$ layer search space. DNN-SE and Classic-SE summaries use the available mean-over-layers and peak-group families and are reported separately rather than compared by raw counts of significant effects.

F COMPUTATIONAL MODELS AND FEATURE EXTRACTION

F.1 ACOUSTIC PREPROCESSING AND LAYER-WISE REPRESENTATIONS

All auditory stimuli were converted to mono, resampled to 16 kHz, and truncated or zero-padded to 4 s before feature extraction. Identical preprocessing was applied across Clean, Noisy, DNN-SE, and Classic-SE waveforms. Hidden representations were extracted at every analyzed layer. For Whisper, encoder states corresponding to the stimulus interval were retained, whereas decoder representations were taken from the final non-special token state during autoregressive decoding. Other speech systems were converted to one sentence-level vector per analyzed layer using the corresponding hidden states. Text LLMs received the same ground-truth Mandarin transcript across acoustic conditions; the final-token hidden state at each layer provided a condition-invariant linguistic representation.

F.2 MODEL INVENTORY

Table F1: Computational model inventory.

Model	Parameters	Depth	Input
log-mel	–	–	waveform
Random wav2vec 2.0	95M	12 encoders	waveform
Random Whisper-Tiny (encoder)	8M	4 encoders	waveform
wav2vec2-Base (Baevski et al., 2020)	95M	12 encoders	waveform
WavLM-Base (Chen et al., 2022)	95M	12 encoders	waveform
Whisper-Tiny (Radford et al., 2023)	39M	4 enc + 4 dec	waveform
Whisper-Medium (Radford et al., 2023)	769M	24 enc + 24 dec	waveform
Whisper-Large-v3 (Radford et al., 2023)	1550M	32 enc + 32 dec	waveform
SpeechT5 (Ao et al., 2022)	144M	12 enc + 6 dec	waveform
Qwen2.5-3B (Yang et al., 2025)	3.1B	36 layers	transcript
Phi-4-mini (Abouelenin et al., 2025)	3.8B	32 layers	transcript
GPT-2 XL (Radford et al., 2019)	1.5B	48 layers	transcript
Mistral-7B (Jiang et al., 2023)	7.3B	32 layers	transcript

F.3 LINEAR ENCODING IMPLEMENTATION

Ridge regression (Hoerl & Kennard, 1970) maps model features X (sentences $\times$ features) to voxel responses Y (sentences $\times$ voxels), following the model-to-brain encoding logic of Schrimpf et al. (2021). Feature standardization is fit on the training fold and applied unchanged to held-out stimuli. When original feature dimensionality exceeds the Johnson-Lindenstrauss target, sparse random projection (Achlioptas, 2001) is applied with $\epsilon = 0.1$; the study-specific targets are 2724 dimensions for static encoding and 3912 for LOCO. In the current roster this projection affects only the highest-dimensional representations. The implementation uses scikit-learn-compatible ridge machinery (Pedregosa et al., 2011).

Each voxel receives its own ridge penalty selected by inner cross-validation. Static encoding uses leave-one-run-out internal validation inside its training data; the held-out test run is never used to tune the penalty. LOCO performs all tuning on the three-condition training side of the outer fold. Held-out voxel predictivity is Pearson r across test sentences. For static encoding, voxel r values are computed within each run and combined by Fisher averaging; LOCO has one held-out condition and therefore one outer test block. ROI raw predictivity is the arithmetic mean across the model-independent set of ceiling-retained voxels.

F.4 CROSS-SUBJECT NOISE CEILING AND NORMALIZED BRAIN SCORE

Biological measurements contain irreducible inter-participant noise, so raw model predictivity is normalized by an empirical ceiling following Schrimpf et al. (2021). For each task and condition, one participant is treated as the target and the remaining participants as a source population. Source participants are subsampled at increasing pool sizes, their neural responses are used to predict the held-out participant, and prediction accuracy is tracked as a function of source-pool size. An asymptotic ceiling is extrapolated to the large-population limit, with bootstrap resampling used to stabilize the estimate.

Ceilings are estimated independently of every computational model and separately for static/run-wise and LOCO evaluation. Voxels with effectively zero ceiling ($< 5 \times 10^{-5}$) are excluded because they contain no measurable shared signal; voxels whose estimate reaches the numerical upper boundary (≥ 0.99) are excluded because the asymptote is not data-constrained. The same retained voxel mask and arithmetic aggregation are used for numerator and denominator:

$$\text{BrainScore}(s, r, c, t) = \frac{r_{\text{raw}}(s, r, c, t)}{\text{Ceiling}(r, c, t)}. \tag{F1}$$

Finite-sample normalized scores can exceed 1. Negative values are preserved for statistical inference and clipped to zero only in figures explicitly labeled as such. Ceiling normalization serves a measurement role rather than an inferential one: significance is assessed on participant-level score contrasts, whereas the ceiling rescales predictivity by the amount of response variance reliably shared across participants.

G ASR AND BEHAVIORAL ROBUSTNESS

G.1 RECOGNITION ROBUSTNESS ACROSS ASR SYSTEMS

CER was recomputed with Whisper-Tiny, Whisper-Medium, and Whisper-Large-v3 to test whether the behavioral boundary depends on a weak recognizer. Stronger ASR systems substantially reduce CER in the degraded conditions and track human sentence-level comprehension more closely. The standardized three-model ensemble provides an additional robustness summary.

Table G1: Mean Character Error Rate across acoustic conditions and ASR systems.

Condition	Whisper-Tiny CER	Whisper-Medium CER	Whisper-Large-v3 CER
Clean	0.083	0.021	0.021
DNN-SE	0.688	0.517	0.483
Classic-SE	0.875	0.654	0.517
Noisy	0.812	0.371	0.275

Across 96 sentences, CER-comprehension correlations are $r = -0.543$ for Whisper-Tiny, $r = -0.670$ for Whisper-Medium, $r = -0.632$ for Whisper-Large-v3, and $r = -0.673$ for the standardized three-model ensemble. Thus, the behavioral baseline becomes stronger as recognition improves, but this does not rescue the incremental contribution of sentence-level neural fit.

G.2 Held-out comprehension prediction

Behavioral prediction is grouped by sentence identity so that all 25 participant responses to the same sentence remain in the same outer fold. B1 = Condition + CER is the reference model; B2 adds static neural fit, B3 adds LOCO neural fit, and B4 adds both. Held-out log loss is the primary metric, with AUROC and Brier score as secondary metrics. Random-weight LOCO is a negative control and the listener's own intelligibility rating is a positive machinery check. Lower log loss and Brier score are better; higher AUROC is better.

Table G2: Held-out comprehension prediction across behavioral evaluation metrics. B1 (Condition + CER) is the reference for incremental log-loss comparisons.

Behavioral model	CV log loss ↓	AUROC ↑	Brier ↓	Δ log loss vs B1
B0: Condition	0.4439	0.6722	0.1403	+0.01409 [−0.00297, +0.03149]
B1: Condition + CER	0.4298	0.7132	0.1366	reference
B2: Static fit	0.4304	0.7123	0.1368	+0.00056 [+0.00022, +0.00096]
B3: LOCO fit	0.4302	0.7137	0.1368	+0.00037 [−0.00154, +0.00234]
B4: + Static + LOCO	0.4307	0.7129	0.1369	+0.00089 [−0.00110, +0.00290]
Neg. ctrl: random LOCO	0.4305	0.7120	0.1369	+0.00066 [+0.00026, +0.00112]
Pos. ctrl: intelligibility	0.3896	0.7892	0.1229	−0.04016 [−0.05905, −0.02032]

The neural extensions do not improve held-out comprehension prediction. B2 (static fit) and the random-weight negative control produce small but reliable increases in log loss relative to B1, while B3 (LOCO fit) and B4 (static + LOCO) do not differ detectably from the reference. In contrast, the participant's own intelligibility rating produces a clear improvement, demonstrating that the evaluation pipeline can detect an informative trial-level predictor. No standalone calibration statistic is reported because calibration slope/intercept or expected calibration error was not prespecified in the available analysis outputs.

G.3 Interpretation

These behavioral analyses define a boundary rather than a failure of the neural benchmark. Cross-condition neural transportability discriminates computational representations, but sentence-level LOCO fit is not a direct trial-wise proxy for whether a listener understands a sentence. Conversely,

CER can track behavioral difficulty without establishing neural correspondence. Machine recognition, within-condition neural correspondence, cross-condition neural transportability, and behavioral success should therefore be treated as related but non-equivalent evaluation axes.