A CROSS-DESCRIPTION TEST OF SPARSE FEATURE CORRESPONDENCE BETWEEN LANGUAGE MODELS AND BRAIN RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model features can predict brain responses. Do the features that matter most for this prediction also matter for predicting the model's own hidden state? We compare sparse-autoencoder (SAE) features using fMRI from 59 people listening to two narrations. We select features and fit separate predictors of fMRI responses and the model's final hidden state on one narration, evaluate on the other, and repeat in reverse. We subtract each feature's contribution from the fixed predictions and compare the signed decreases in held-out variance explained. Across six model–SAE pairs and both directions, mean rank correlation is $-.056$, and no model passes both timing and feature-identity tests in both directions. Simulations show why this null is inconclusive. At 256 features, known positive correspondence is detected in only 1.6%–10.4% of replicates across 64 conditions from two model families, including planted correlation $.80$. The false-positive fraction is set to 5%. An exploratory sweep recovers median correspondence of $.56$ at 24 features and $.02$ at 256, despite the same planted $.80$. Separate controls recover the fMRI predictor's planted ranking much better with Gaussian predictors of the same shape. These simulations test reliance estimation with fixed predictor matrices, leaving the full transfer procedure's sensitivity unmeasured. The empirical null cannot rule out shared brain–model organization: prediction and repeatability need a separate check that the analysis can recover known feature agreement.

1 INTRODUCTION

Encoding models use the internal states of language models (LMs) to predict fMRI responses recorded while people listen to stories (Huth et al., 2016; Wehbe et al., 2014; Jain & Huth, 2018; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Kumar et al., 2024). Sparse autoencoders (SAEs) decompose these dense residual-stream vectors into sparsely active features (Cunningham et al., 2023; Lieberum et al., 2024; He et al., 2024). Recent studies use these features to predict brain responses and map predictive accuracy across cortex (Lepori et al., 2026; Guo et al., 2026). Related work aligns vision-model SAE units with voxels (Mao et al., 2025). These results raise a further question: when the same SAE features predict both brain responses and an LM's hidden state, do the two fitted predictors rely on the same features?

We fit two regressions to the same feature bank: one predicts cortical blood-oxygen-level-dependent (BOLD) responses, and the other predicts the LM's hidden state. A feature can help one predictor and contribute little to the other. For each feature, we subtract its fitted contribution from the fixed prediction and recompute held-out R^2_{oos}, which measures prediction relative to the training mean. The decrease in this score is the feature's *reliance value*. It is negative if subtracting the contribution improves prediction. The values form a *reliance profile*. We compare their rankings between predictors. This calculation leaves brain activity and the LM unchanged.

These rankings may be difficult to estimate even when the features predict the targets. Good predictions can combine features whose individual contributions are hard to distinguish. A ranking that repeats across participant groups may also reflect a stable property of the predictors or fitting procedure. We therefore use simulations with known contributions to ask how well the fitted profiles recover each ranking and the agreement between them.

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

We select features from one Narratives narration (Nastase et al., 2021) and fit regressions to predict each listener's BOLD responses and the LM's hidden state. We evaluate these regressions on the other narration with the feature bank and regression parameters fixed. Reversing direction selects a new bank and fits new regressions (Figure 1). The pretrained LM and SAE are unchanged.

Neither prediction target is used to select the bank. We standardize features to remove arbitrary differences in their units and evaluate on another narration to test transfer beyond the selection text. Temporal and feature-identity nulls test whether shared timing or layer composition can explain agreement between the profiles.

The empirical comparison finds no consistent correspondence. Simulations introduced after review show why that result is inconclusive: with 256 features, the estimator detects planted correspondence in only 1.6%–10.4% of replicates across the tested positive conditions. Smaller banks recover more cross-system correspondence, while separate Gaussian-predictor controls recover individual rankings more accurately. We present the simulations first because poor recovery limits what the empirical null can tell us about shared organization between brain and model.

2 Related work

Encoding models and their evaluation. Encoding studies predict brain responses from stimulus features (Huth et al., 2016; Wehbe et al., 2014; Jain & Huth, 2018). Studies using LM activations often call the predictive correlation neural predictivity or brain alignment (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Kumar et al., 2024). These scores can reflect acoustics, story position, or temporal autocorrelation as well as linguistic content. Evaluation methods address these confounds by holding out contiguous blocks, comparing predictions with stimulus-level baselines, and assigning separate penalties to feature groups (Nunez-Elizalde et al., 2019; Dupré la Tour et al., 2022; Hadidi et al., 2026). We adopt these methods when evaluating prediction and correspondence.

Sparse features. SAEs represent activations through sparse features and a learned decoder (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024; Lieberum et al., 2024; He et al., 2024). The dictionaries used here all have JumpReLU-form encoders (Rajamanoharan et al., 2024). Llama Scope checkpoints were trained as Top-K SAEs and released in a fixed-threshold JumpReLU form (He et al., 2024). The learned features depend on the model, layer, dictionary width, and seed (Leask et al., 2025; Paulo & Belrose, 2025). Feature splitting and absorption further complicate their interpretation (Chanin et al., 2025). Brain–SAE studies use sparse features for interpretable encoding and cortical mapping (Lepori et al., 2026; Guo et al., 2026; Mao et al., 2025). The closest study, Lepori et al. (2026), asks whether alignment reflects primary rather than idiosyncratic variation in LM representations. Our comparison instead uses the feature-by-feature losses of two separately fitted predictors, followed by tests of whether their rank agreement can be recovered from known planted profiles.

Earlier work questions whether successful encoding identifies the underlying computation (Antonello & Huth, 2024; Tuckute et al., 2024). The model-reliance literature likewise distinguishes prediction from reliance on a variable (Fisher et al., 2019). We apply that distinction to BOLD and LM predictors fitted with the same SAE features, then test whether the analysis can recover known agreement between their reliance profiles. The simulations specify ablation-loss rankings directly, because agreement between signed coefficients need not produce agreement between ablation losses.

3 Study design

3.1 Narrations and brain data

We use the Physical and Social audio runs from Narratives (OpenNeuro `ds002345`; Nastase et al. 2021; OpenNeuro 2021). Both describe the same animation of geometric shapes: Physical describes their movements, while Social presents them as characters with intentions (Nastase et al., 2021). This contrast lets us test transfer between descriptions of a common stimulus with different vocabulary. Both recordings last about 458 seconds (7 minutes 38 seconds), including the opening

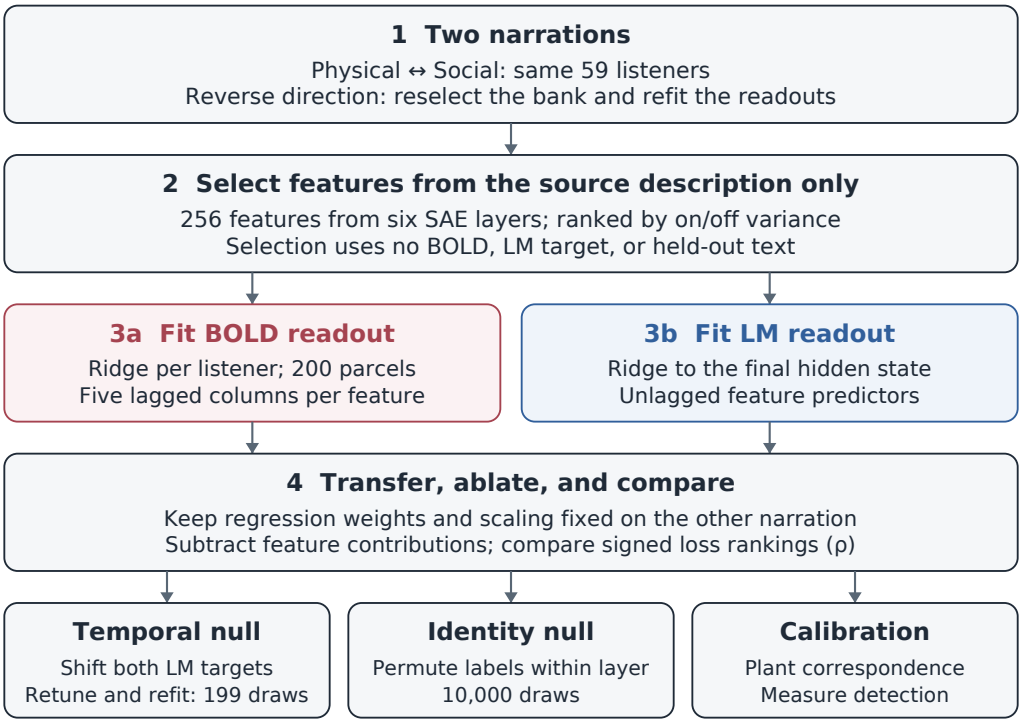

Figure 1: Feature selection and evaluation across narrations. We select features and fit the BOLD and LM readouts on one narration, then evaluate them on the other with regression weights and scaling fixed. For each feature, we subtract its contribution from the fixed predictions and recompute held-out R^2_{oos}. Spearman ρ compares the signed decreases between readouts. Reverse transfer selects a new bank and refits the readouts. Both null tests must reject to support correspondence; simulations assess recovery of planted correspondence with fixed predictor matrices.

jingle and silence. All 59 participants heard both in counterbalanced order. After removing the shared jingle from the transcripts, their content-vocabulary Jaccard overlap is .10 (Appendix A).

We extracted BOLD time series from fMRIPrep derivatives (Esteban et al., 2019) in MNI152NLin6Asym space. We used the Schaefer-2018 atlas to group voxels into 200 cortical parcels spanning seven networks (Schaefer et al., 2018). We removed 24 motion terms and six anatomical CompCor components by regression (Behzadi et al., 2007), detrended the signals, and applied a 0.008-Hz high-pass filter. Each parcel was then standardized within run using uncensored time points. Time points with excessive motion or signal change, or non-steady-state signals, were excluded from fitting and evaluation (details in Appendix B). The retained time grids contain 303 and 305 TRs, respectively, at 1.5 seconds per TR.

We test Gemma 2 2B and 9B with Gemma Scope, Llama 3.1 8B with Llama Scope, and Gemma 3 1B, 4B, and 12B with Gemma Scope 2 (Gemma Team, 2024; Grattafiori et al., 2024; Gemma Team, 2025), each at six residual-stream layers. Each model–direction pair is a *cell*: P→S means fitting on PHYSICAL and evaluating on SOCIAL.

3.2 FEATURE SELECTION FROM THE SOURCE DESCRIPTION

We tokenize the transcript and process it in consecutive 1,024-token causal windows. At each selected layer, the SAE encodes every token. We average each feature's activations within each TR, placing the full dictionary on the fMRI time grid. The LM target is the final hidden state averaged over the same tokens. Predicting this target assesses an internal representation, not effects on next-token predictions.

Feature selection uses only source-description activations. A feature is eligible if it is active in at least five source TRs and no more than 95% of them. Within each layer, we rank eligible features by their on/off variance, $p(1-p)$, which is largest at $p = .5$. Here p is the fraction of source TRs with nonzero activation. We interleave the six rankings to balance layer contributions. The main bank contains $K = 256$ features. Nested banks of 64 to 1,024 test sensitivity to bank size. We select one bank per transfer direction and repeat selection inside each training fold for within-description analyses. Selection uses no BOLD values, LM targets, held-out text, or fitted results.

3.3 Fitting the BOLD and LM predictors

To predict BOLD responses up to six seconds after the words, we include five columns per feature, at lags of 0 to 4 TRs. We fit one ridge regression per participant to all 200 parcels. At $K = 256$, this gives 1,280 predictor columns for about 300 time points and requires strong regularization. The LM regression uses unlagged features because its target comes from the same tokens.

Both regression models are evaluated on five contiguous outer test blocks. A four-TR gap on each side of a test block prevents overlap between training and test lagged predictors. BOLD cleaning and filtering use the full run before splitting, so preprocessing is shared across within-description folds (Appendix B). Inner folds select the ridge penalty from a fixed twelve-value grid. We standardize predictors and targets using training rows only and select separate penalties to maximize Pearson r and out-of-sample R^2_{oos}. The latter compares predictions with the training mean. A negative value means they perform worse than that constant. As specified before examining results, the prediction benchmark uses r-tuned models and the reliance analysis uses R^2_{oos}-tuned models.

3.4 Measuring reliance after transfer

For transfer, we refit each regression on the source description with its most frequently selected penalty and apply its regression weights and scaling unchanged to the other description. We repeat in reverse. We then subtract each feature's contribution from the fixed prediction, including all five BOLD lags. This calculation, which we call *ablation*, is equivalent to zeroing the feature's standardized inputs at their training mean. Reliance is the original minus adjusted R^2_{oos}, averaged over target dimensions and, for BOLD, participants:

$$I_k = R^2_{\text{oos}}(Y, \widehat{Y}) - R^2_{\text{oos}}(Y, \widehat{Y}^{(-k)}). \tag{1}$$

Here $\widehat{Y}^{(-k)}$ is the prediction after subtracting feature k's contribution. A negative I_k means this subtraction improves prediction. We compare the two reliance profiles using Spearman correlation. The result describes the fitted predictors (Fisher et al., 2019); it does not establish causal feature use by the brain or LM. Because correlated features can substitute for one another, we also subtract the joint contributions of source-defined clusters.

For one target, let $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ be the prediction error and $\boldsymbol{g}_k$ the contribution being subtracted. Expanding the squared error gives the target's ablation loss:

$$\Delta_k = \frac{\|\boldsymbol{g}_k\|^2 + 2\boldsymbol{e}^\top \boldsymbol{g}_k}{\sum_{t \in \text{test}} (y_t - \bar{y}_{\text{train}})^2}. \tag{2}$$

The reported I_k averages these losses across target dimensions and, for BOLD, participants. The squared contribution is nonnegative, but its cross-term with prediction error can make the loss negative. In a noiseless planted predictor, the error is zero and only the squared contribution remains. This distinction explains why calibration must control the loss rankings directly rather than coefficient signs.

Coefficient magnitudes depend on predictor units, which can create apparent agreement when the two regression models share activation scales. Standardization based on training data removes these arbitrary units. Held-out ablation losses still depend on how often features activate, their correlations, distribution shift, and estimation error (Section 4.4).

3.5 Testing against temporal and identity nulls

Shared narrative timing can produce positive correspondence without shared feature reliance. To test this possibility, we circularly shift the LM targets 199 times. Source and held-out targets receive

independent offsets of at least ten TRs, drawn once per direction and bank size and shared across the six models. For each draw, we reselect the penalty, refit and transfer the LM predictor, and compare its new reliance profile with the fixed BOLD profile. This temporal null tests whether correct timing gives more agreement than shifted timing. For each draw, we average the twelve cell correlations using the same shift index to form a pooled null. We compare the observed mean with these 199 null means using $p = (1 + \#\{\rho_{\text{null}} \geq \rho_{\text{observed}}\})/200$. Sharing offsets across models preserves their common retiming within each direction. We report raw per-cell p values and assess multiple testing across the twelve primary temporal tests with Benjamini–Hochberg correction.

The identity null tests whether exact feature identities matter by shuffling LM feature labels within each layer for 10,000 draws, preserving layer composition. Both null tests use the upper tail. Lower-tail results are explicitly labelled. Evidence of correspondence requires passing both tests in both transfer directions. The identity null need not be centered at zero because the permutations preserve between-layer structure. Passing this test shows that exact feature pairing matters relative to those permutations. The decision rule also requires positive correspondence and the temporal test.

3.6 Calibration with known feature reliance

The simulations combine Gemma 2 2B and Llama 3.1 8B, variance-selected and correlation-pruned banks, and both transfer directions, giving eight designs. Each uses the held-out narration's predictor matrix for the source-selected bank at $K = 256$. BOLD predictors retain the five lags, participant-by-time rows, and censoring pattern. LM predictors use TR-level rows without lags. The four-TR purge and feature groups are also retained.

We plant agreement between the noiseless ablation-loss rankings that the analysis is intended to recover. For a planted linear predictor, group k has loss proportional to $\beta_k^2 c_k$, where $c_k = \|\sum_{j \in G_k} \tilde{\boldsymbol{x}}_j\|^2$ and G_k contains the group's standardized predictor columns. We generate two positive loss profiles p, q with a chosen Spearman correlation. For non-silent groups, setting

$$\beta_k = s_k \sqrt{p_k / c_k}, \qquad s_k \in \{-1, +1\}, \tag{3}$$

gives the BOLD predictor the desired loss ordering; we use q analogously for the LM predictor. Signs are drawn independently for the two systems. Silent groups receive zero coefficients. Finite rank resolution and ties can shift the realized correlation from its nominal value, so we record it in every replicate. This construction replaces the first calibration's signed coefficients, whose correlation did not specify agreement between ablation losses (Appendix K).

Random loadings mix each signal into the target dimensions. We scale it to one or two times the mean per-target standard deviation of the empirical residuals, then add recentered residual rows sampled independently with replacement. Each system has one planted profile across target dimensions and, for BOLD, participants. Within blocked folds of the resulting matrix, we retune and fit the regressions, subtract feature contributions, and compare the recovered profiles.

For every design and signal scale, 500 replicates at nominal $\rho = 0$ set the reference distribution. We run 250 replicates at each positive level, $\rho \in \{.2, .4, .6, .8\}$, and count how often recovered correspondence exceeds the reference distribution's 95th percentile. The grid therefore contains 64 positive conditions and 16 reference conditions. The empirical temporal and identity nulls are not used to make these detection decisions.

The calibration fixes the banks and fits within the held-out narration, without repeating feature selection or transfer. Independent residual-row sampling removes temporal autocorrelation. Shared planted rankings and large signals are intended to favor recovery, but these choices do not establish an upper bound on the full transfer test's sensitivity.

3.7 Decision criteria and sensitivity checks

Sensitivity analyses vary bank size, correlation pruning, grouped ablation, token pooling, and participant averaging. We also compare reliance profiles across participants and folds to assess reliability.

The prespecified rule requires the global correspondence statistic to pass both nulls, with positive correspondence in both directions. The result must persist at three adjacent bank sizes including

256, occur in at least two model families after correction, and agree between the variance-selected bank and a second construction. Interpreting a null also requires at least 80% detection at planted $\rho = .40$ (Appendix H). After examining the first results, we added a requirement that the bank improve held-out prediction beyond the stimulus baseline (Section 4.3). This requirement was not prespecified.

4 RESULTS

4.1 DETECTION STAYS BELOW 11% AT 256 FEATURES

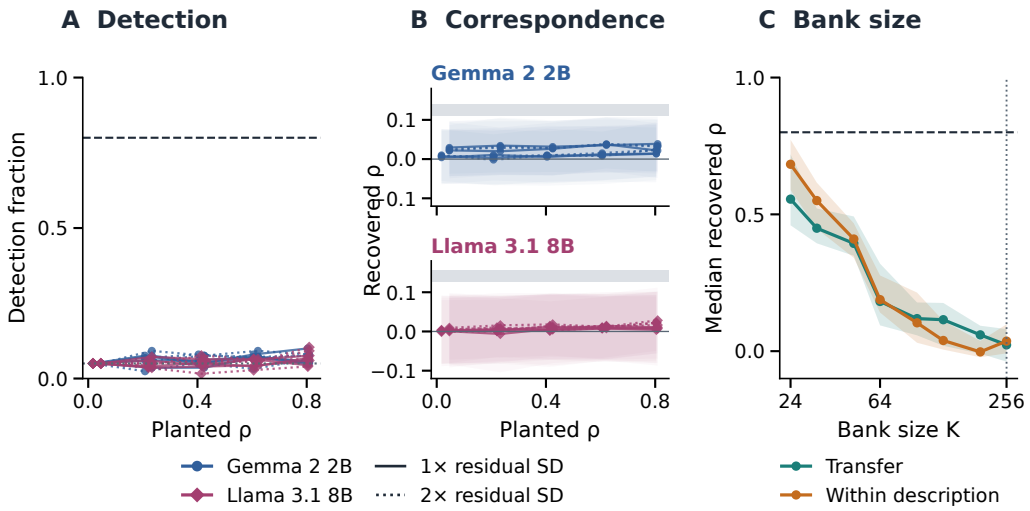

Figure 2: Recovery of planted correspondence. **A:** detection versus realized planted ρ for each design and signal scale; dashed line: 80% criterion; dotted line: 5% false-positive fraction. **B:** recovered ρ by model family. Lines show condition means, colored bands mean $\pm$ replicate SD, and gray bands the range of null 95th percentiles. Solid/dotted curves in A–B use one/two residual SDs. **C:** exploratory bank-size sweep at planted $\rho = .80$: median recovered ρ and interquartile range over designs and seeds. Transfer and within-description labels identify the empirical predictor designs (Appendix K.1).

The corrected simulations test whether the fitted reliance profiles recover known correspondence (Figure 2A–B). For each design and signal scale, detection requires a recovered correlation above the 95th percentile of replicates with planted $\rho = 0$. This sets the false-positive fraction to 5%. This is not an independent estimate of the empirical test's error rate.

Detection ranges from 1.6% to 10.4% across 64 positive conditions: eight designs, two signal scales, and four planted correlations. Mean detection is 6.0%. Even at planted $\rho = .80$, detection is only 4.0%–10.4%, well below the 80% criterion. The recovered correlations barely change between planted $\rho = 0$ and .80: condition means range from $-.001$ to .029 and from .004 to .039, respectively. These differences are small relative to replicate standard deviations of .06–.10 and null thresholds of .11–.16 (Appendix K).

Poor detection persists despite large injected signals, with standard deviations equal to or twice those of the empirical residuals. Doubling the signal does not consistently improve detection. These simulations assess recovery with fixed empirical predictors and independently resampled residual rows, using fits within the held-out narration. They leave the sensitivity of feature selection, transfer, and the joint null-testing rule unmeasured.

Table 1: Recovery in existing exploratory simulations. All entries are medians at one residual SD. In A, the two columns share 24 replicates (eight source-selected designs, three seeds), with planted cross-system $\rho = .80$. B uses a separate experiment with 32 replicates per condition (eight designs, four seeds). It measures each system's self-recovery; cross-system correspondence was not measured in B.

A. Paired bank-size experiment		
K	BOLD self-recovery	Recovered correspondence
24	0.920	0.556
48	0.632	0.394
256	0.040	0.023
B. Predictor and noise controls at $K = 256$		
Predictors / noise	BOLD self-recovery	LM self-recovery
SAE / empirical	0.060	0.140
SAE / white	0.062	0.122
Gaussian / empirical	0.998	0.302
Gaussian / white	0.999	0.322

4.2 Smaller banks improve recovery in exploratory simulations

We distinguish recovery of an individual ranking from recovery of agreement between two rankings. *Self-recovery* is the correlation between a fitted reliance profile and its own planted ranking. *Recovered correspondence* is the correlation between the two fitted profiles. A predictor can recover part of its own ranking while agreement with the other predictor remains poorly estimated.

Table 1 compares these quantities in the same replicates of a bank-size sweep. With planted cross-system $\rho = .80$, median BOLD self-recovery is .920 at $K = 24$ and .040 at $K = 256$. Recovered correspondence is .556 and .023, respectively. At $K = 48$, self-recovery is .632 and recovered correspondence is .394. Each bank is a random subset of the selected 256 features. These exploratory results use only three seeds per design and do not establish a bank size at which the full transfer test has adequate power. Recovery also declines in the separate within-description sweep (Figure 2C).

A separate control changes the predictor matrix or residual noise at $K = 256$ (Table 1B). With empirical residual noise, replacing SAE predictors by independent Gaussian entries of the same shape raises median BOLD self-recovery from .060 to .998. LM self-recovery remains lower, rising from .140 to .302. Replacing residual noise with white noise while keeping the SAE predictors gives BOLD self-recovery of .062. The fitting procedure and matrix dimensions are retained, but the Gaussian comparison also changes temporal structure, repeated rows, and injected realizations. It cannot isolate feature correlation or establish a universal sample-size requirement (Appendix K.1).

Recovery therefore varies with the bank size, predictor matrix, and target. The difference between BOLD and LM self-recovery also shows why recovering one ranking is insufficient for a comparison that depends on both.

4.3 Prediction improves on stimulus controls only within description

In the empirical transfer analysis, the selected features predict both targets. Across twelve model–direction pairs, regressions tuned for r reach mean parcelwise BOLD r of .07–.09 and LM r of .30–.47. The BOLD regressions used for reliance, tuned for R^2_{oos}, explain less than 1% of held-out variance (.003–.008; Table 8). Their median absolute single-feature ablation losses are $1\text{–}3 \times 10^{-5}$ in R^2_{oos} units across cells. In four cells, most ablations improve prediction. The correspondence statistic therefore ranks small changes of both signs, even when the overall prediction correlation is positive (Table 9).

We test added predictive value with a banded-ridge baseline. One block contains word and phoneme counts, the audio envelope and its first difference, and story-position bases. A second contains mean and summed token surprisal from the same LM. The full model adds the SAE bank as a third block. Each block has its own penalty, with block weights chosen inside the same nested folds. Paired differences between the full model and baseline are averaged across six models within each

participant, then across participants. We bootstrap the 59 participants for confidence intervals (details in Appendix G).

Adding SAE features improves within-description R^2_{oos} for both narrations, with both intervals excluding zero (Table 2). After transfer, mean changes in both metrics are negative. Three of the four transfer intervals exclude zero. The added predictive value within a description therefore does not transfer with the fitted readouts. Strong regularization may contribute: source-selected penalties and block weights can shrink the 1,280-column SAE block more than the smaller control blocks. This comparison cannot distinguish a lack of additional information from difficulty estimating it.

Table 2: Change in prediction after adding the SAE block to the enhanced baseline at $K = 256$. Paired differences are averaged across six models within each participant, then across 59 participants. We bootstrap participants for the 95% intervals.

Regime	Description or direction	Δr	ΔR^2_{oos}
Within	Physical	$-.0001$ $[-.0045, .0049]$	$.0010$ $[.0003, .0018]$
Within	Social	$.0047$ $[-.0007, .0104]$	$.0034$ $[.0022, .0046]$
Transfer	Physical→Social	$-.0118$ $[-.0170, -.0069]$	$-.0019$ $[-.0036, -.0004]$
Transfer	Social→Physical	$-.0108$ $[-.0163, -.0055]$	$-.0006$ $[-.0027, .0015]$

4.4 The empirical correspondence remains inconclusive

The twelve transfer correlations range from $-.626$ to $.335$, with a mean of $-.056$. Shifting the LM targets in time produces means at least this high in 160 of 199 draws ($p = .805$). Only Gemma 2 9B P→S passes both the temporal and feature-identity null tests at uncorrected $p < .05$. Its reverse direction does not. No model therefore passes both tests in both directions (Figure 4D). The prespecified second bank construction, ranked by source prevalence, also falls within its temporal null (mean $\rho = -.093$, $p = .810$; Table 3). Given the limited simulation recovery, these failures cannot establish that brain and model lack shared organization.

Feature prevalence, or how often a feature activates, is positively associated with BOLD reliance in all twelve cells. Its association with LM reliance varies by family and direction. Selection by $p(1-p)$ already favors features of intermediate source prevalence. Adjusting for prevalence after examining the results raises mean correspondence to $+.109$, but the matched temporal test gives $p = .060$, and no identity test was run for this adjustment. Prevalence changes the sign of the mean comparison without establishing shared reliance. Appendix I and Figure 4 give the full empirical results.

4.5 Bank changes do not resolve the empirical comparison

The empirical mean stays within its temporal null across all five nested bank sizes, including the smallest bank of 64 features (Table 3). Prevalence ranking, correlation pruning, and grouped ablation give the same qualitative result. Grouping is a limited change here: within-layer clustering at $|\rho| = .80$ leaves 239–254 groups from 256 features. It therefore does little to reduce the number of contributions being estimated.

Across the 120-cell bank-size and bank-variant sweep, seven cells have uncorrected $p < .05$, but five involve the same model and direction across nested variants. The smallest Benjamini–Hochberg q is .20. The empirical result persists across bank choices, but remains inconclusive: the exploratory simulations use a different fitting procedure and do not establish adequate sensitivity for these transfer tests.

4.6 Repeatability does not establish recovery

The transferred BOLD reliance profile repeats across participant groups, with a split-half correlation of .773, but both groups could share the same ranking error. Repeatability alone cannot establish recovery. We have no corresponding transfer reliability estimate for the LM profile.

Averaging BOLD across participants before fitting raises within-description prediction to $r = .164$–$.267$, and three of twelve cells pass both null tests. These fits hold out time points from the same

Table 3: Empirical correspondence under changes to the feature bank. Each row averages six models and two directions. All p values use 199 pooled temporal-null draws that shift both source and held-out LM targets, with offsets shared across models. Grouping subtracts every cluster member's contribution; pruning keeps one representative per cluster.

Sensitivity	Setting	Mean ρ	p_{temporal}
Bank size	$K = 64$	$-.106$	.825
	$K = 128$	$-.112$	.885
	$K = 256$	$-.056$	.805
	$K = 512$	$-.075$	.755
	$K = 1024$	$-.063$	.725
Bank ranking	source prevalence, $K = 256$	$-.093$	.810
Grouping	clusters at $\lvert\rho\rvert = .80$, $K = 256$	$-.053$	.760
Pruning	representatives at $\lvert\rho\rvert = .80$, $K = 256$	$-.055$	.780

listeners and do not test generalization to new listeners or transfer between narrations (Appendix J.1). They therefore do not resolve the cross-description comparison.

5 Discussion

The feature banks predict both targets across narrations, and BOLD reliance profiles repeat across participant groups. Yet simulations recover little known correspondence at the empirical bank size. Prediction and repeatability therefore do not establish that these profiles measure shared reliance.

Smaller banks improve recovery in a sweep with few seeds; adequate transfer sensitivity remains untested. Matrix dimensions alone do not explain BOLD self-recovery in the Gaussian controls. Stacking listeners who hear the same narration repeats stimulus predictors without adding stimulus time points. Independent Gaussian entries change this repetition, temporal structure, and feature correlations together. These controls therefore cannot isolate the cause of poor recovery or define a universal sample-size requirement.

The first calibration specified agreement between signed coefficients, which did not control agreement between noiseless ablation losses. The corrected calibration plants the loss rankings directly (Appendix K). Recovery checks must test the quantity being interpreted: here, agreement between two fitted reliance profiles.

6 Limitations

The study covers two narrations, uses only the LM's final hidden state as its model target, and calibrates two model families. Next-token prediction could give different feature rankings. BOLD is a vascular signal averaged within 200 parcels; the lag window ends near the hemodynamic peak, and individual reliance profiles agree weakly across listeners. SAE features also depend on the model, layer, and release.

In every cell, at least 84% of selected features remain active in more than 5% of held-out TRs (Appendix E), although activation frequency and magnitude may change. Run order was counterbalanced but not modelled, and fitting included the shared 26-TR opening jingle.

7 Conclusion

The cross-description analysis finds no consistent feature-reliance correspondence, but simulations recover little known correspondence at the same bank size. Smaller banks recover more in exploratory tests. The empirical null therefore cannot rule out shared brain–model organization. Before interpreting such a null, the analysis needs to show that it can recover agreement between known feature-reliance profiles.

REPRODUCIBILITY STATEMENT

The appendix documents the data, models, analysis, and calibration. The anonymous code release contains the analysis specification, code, eight corrected calibration aggregates, the superseded first grid with its defect notice, and exploratory simulation scripts and outputs. It provides deterministic reports and checksum catalogs. The 192 calibration shards are retained with SHA-256 digests but are not redistributed. The released aggregates support checking reported detection rates, not individual shard contents.

The accompanying supplementary archive includes the saved sign and penalty diagnostics, bank-size and predictor-control outputs, and code to reproduce the numerical summaries in Table 1 and Appendix K without running simulations. Diagnostic R^2_{oos} values are replicate-level means over folds. Individual fold scores were not saved in that diagnostic file. Fold-specific penalty choices are included.

Code:

https://anonymous.4open.science/r/A-Cross-Description-Test-of-Sparse-Feature-Correspondence-Between-Language-Models-and-fMRI-Responses-512A/

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation and provide feedback on methodology and interpretation. We did not use these tools to generate synthetic datasets. Additionally, we used them for literature searches, manuscript drafting and editing, reference formatting, and figure preparation. We manually reviewed and revised all AI-assisted material. AI-assisted code was tested, and reported values were checked against saved outputs. We take responsibility for the final content of this work, including its text, claims, code, and results.

REFERENCES

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14, 2014. doi: 10.3389/fninf.2014.00014.

Soroosh Afyouni and Thomas E. Nichols. Insight and inference for DVARS. *NeuroImage*, 172:291–312, 2018. doi: 10.1016/j.neuroimage.2017.12.098.

Richard Antonello and Alexander G. Huth. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79, 2024. doi: 10.1162/nol_a_00087.

Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T. Liu. A component based noise correction method (CompCor) for BOLD and perfusion based fmri. *NeuroImage*, 37(1):90–101, 2007. doi: 10.1016/j.neuroimage.2007.04.042.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

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

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.48550/arXiv.2409.14507. URL `https://arxiv.org/abs/2409.14507`. Oral presentation; arXiv:2409.14507.

Jingyuan E. Chen and Gary H. Glover. Functional magnetic resonance imaging methods. *Neuropsychology Review*, 25(3):289–313, 2015. doi: 10.1007/s11065-015-9294-9.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. doi: 10.48550/arXiv.2309.08600. URL `https://arxiv.org/abs/2309.08600`.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022. doi: 10.1016/j.neuroimage.2022.119728.

Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019. doi: 10.1038/s41592-018-0235-4.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. URL `https://arxiv.org/abs/2408.00118`.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL `https://arxiv.org/abs/2503.19786`.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4.

Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL `https://arxiv.org/abs/2407.21783`.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. Sparse autoencoders map brain–LLM alignment onto cortical semantic topography. *arXiv preprint arXiv:2605.23035*, 2026. doi: 10.48550/arXiv.2605.23035. URL `https://arxiv.org/abs/2605.23035v1`. Accepted at CoNLL 2026; version 1.

Nima Hadidi, Ebrahim Feghhi, Bryan H. Song, Idan A. Blank, and Jonathan C. Kao. 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.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama Scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024. doi: 10.48550/arXiv.2410.20526. URL https://arxiv.org/abs/2410.20526.

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.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. doi: 10.1038/nature17637.

Shailee Jain and Alexander G. Huth. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6629–6638, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html.

Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15:5523, 2024. doi: 10.1038/s41467-024-49173-5. URL https://doi.org/10.1038/s41467-024-49173-5.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2502.04878. URL https://arxiv.org/abs/2502.04878. arXiv:2502.04878.

Michael A. Lepori, Kendrick Kay, and Greta Tuckute. Interpreting brain responses to language with sparse features from language models. *arXiv preprint arXiv:2606.06857*, 2026. doi: 10.48550/arXiv.2606.06857. URL https://arxiv.org/abs/2606.06857v2. Version 2.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Ziming Mao, Jia Xu, Zeqi Zheng, Haofang Zheng, Dabing Sheng, Yaochu Jin, and Guoyuan Yang. Sparse autoencoders bridge the deep learning model and the brain. *arXiv preprint arXiv:2506.11123*, 2025. doi: 10.48550/arXiv.2506.11123. URL https://arxiv.org/abs/2506.11123v1. Version 1.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma Scope 2 technical paper. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf.

Samuel A. Nastase, Valeria Gazzola, Uri Hasson, and Christian Keysers. Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, 14(6):667–685, 2019. doi: 10.1093/scan/nsz037. URL https://doi.org/10.1093/scan/nsz037.

Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, Gina Choe, Ariel Goldstein, Tamara Vanderwal, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8(1):250, 2021. doi: 10.1038/s41597-021-01033-3.

Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L. Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, 197:482–492, 2019. doi: 10.1016/j.neuroimage.2019.04.012.

OpenNeuro. Narratives fmri dataset, ds002345, version 1.1.4. OpenNeuro dataset repository, 2021. URL https://openneuro.org/datasets/ds002345/versions/1.1.4.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025. URL https://arxiv.org/abs/2501.16615.

Jonathan D. Power, Kelly A. Barnes, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154, 2012. doi: 10.1016/j.neuroimage.2011.10.018.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. URL https://arxiv.org/abs/2407.14435.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018. doi: 10.1093/cercor/bhx179.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 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.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32, 2019. doi: 10.48550/arXiv.1905.11833. URL https://arxiv.org/abs/1905.11833. arXiv:1905.11833.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024. doi: 10.1038/s41562-023-01783-7.

Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrës Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017. doi: 10.1016/j.neuroimage.2016.10.038. URL https://doi.org/10.1016/j.neuroimage.2016.10.038.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9(11): e112575, 2014. doi: 10.1371/journal.pone.0112575.

A Stimuli, cohort, and BOLD targets

A.1 Dataset and stimulus selection

We use the Narratives naturalistic-language fMRI dataset (OpenNeuro ds002345; Nastase et al. 2021), specifically the audio-only shapesphysical and shapessocial runs. Each of the 59 participants heard both narrations in counterbalanced order. Both audio files contain 21,961,728 samples at 48,000 Hz: 457.536 seconds, including the opening jingle and silence. Thus 7.6 minutes describes the full recording, not the narration alone. The runs share TR, scanner, and acquisition parameters. Transfer tests generalization between these two stimuli. It does not establish generalization across narrative domains.

We obtain word timings from a Whisper transcription aligned to the audio, with a stimulus-onset correction of 4.5 seconds. The words of the shared opening jingle occupy the first 26 TRs of both runs, which we retain in fitting, ablation, and the nulls. Although the source dataset counterbalanced run order, we do not model each participant's order. The content-vocabulary overlap reported in the main text is the Jaccard index of word types after removing stop words: .13 for the full transcripts and .10 after removing the jingle. Shared words mainly describe spatial and temporal relations ("around", "towards", "slowly", "two").

Each primary BOLD target contains 200 cortical parcels, and the primary analyses require all 59 eligible participants (Table 4). The Physical target combines derivatives for 47 participants with

Table 4: Evaluation cohorts. The full row count is the number of participants multiplied by the number of story TRs. Flagged rows remain on the saved timing grid but are excluded as responses during fitting.

Description	TR (s)	TRs	Participants	Full rows	Flagged	Retained
PHYSICAL	1.5	303	59	17,877	487	17,390
SOCIAL	1.5	305	59	17,995	488	17,507

derivatives recovered separately for 12 others. We checked that each participant/run key was unique, each array had the expected 303 rows and identical parcel labels, all values were finite, and every file had a recorded hash.

A.2 WHAT A BOLD TARGET REPRESENTS

The BOLD target is averaged within cortical parcels. BOLD reflects vascular changes coupled to aggregate neural activity and blurs events over time (Chen & Glover, 2015). The predictions therefore concern this measured response rather than neuronal firing.

B FMRI PREPROCESSING AND TARGET CONSTRUCTION

For each participant-run, Nilearn's `NiftiLabelsMasker` (Abraham et al., 2014) extracts parcel time series from fMRIPrep derivatives in `MNI152NLin6Asym` space (Esteban et al., 2019), using the Schaefer-2018 200-parcel, seven-network cortical atlas (Schaefer et al., 2018). Extraction applies no standardization, detrending, smoothing, or filtering. Cleaning follows as a separate step. Both descriptions retain all 200 parcel labels.

We clean the time series using 24 motion regressors (six rigid-body estimates, their derivatives, squares, and squared derivatives) and the first six anatomical CompCor components (Behzadi et al., 2007). Cleaning includes detrending, standardized confounds, and a 0.008-Hz high-pass filter, with no low-pass filter or global-signal regressor. We then standardize each parcel within each run using only uncensored TRs,

$$z_{s,t,p} = \frac{y_{s,t,p} - \mu_{s,p}^{(U)}}{\sigma_{s,p}^{(U)}}, \tag{4}$$

where U denotes uncensored TRs. Cleaning and filtering use the full retained run grid, including flagged time points. Only the subsequent run-wise scaling uses uncensored TRs. These operations precede the fold split. The separate scaling applied during ridge fitting uses only the training fold.

A TR is flagged when framewise displacement exceeds 0.5 mm, standardized DVARS exceeds 1.5, or an fMRIPrep non-steady-state indicator is active (Power et al., 2012; Afyouni & Nichols, 2018). We omit flagged rows only when building a design matrix, so earlier uncensored stimulus lags remain valid predictors for later responses. No censored target contributes to any loss or metric.

C TRANSCRIPT, TOKEN, AND TR ALIGNMENT

We assign each word to the TR containing its midpoint after the onset correction. For onset o_i and offset d_i, the assignment is $t_i = \lfloor ((o_i + d_i)/2 + \delta_{\text{onset}})/1.5 \rfloor$. An out-of-range index stops validation rather than being clipped. We tokenize the transcript separately for each base model without special tokens, then group subword tokens back into words. Any mismatch between these reconstructed words and the timing table or TR-aligned transcript stops the analysis.

We run the model on consecutive, non-overlapping causal windows of at most 1,024 tokens, with context limited to each window. At each selected layer, the corresponding SAE encodes the residual stream in its released JumpReLU form (Llama Scope publishes JumpReLU-form parameters derived from its TopK training). The TR predictor for feature k is its mean non-negative activation over all

Table 5: Model and SAE configurations. Each model's feature bank draws from the six depth-spaced residual-stream layers listed here.

Model	Model ID	SAE release	Dictionary per layer	Layers
Gemma 2 2B	`google/gemma-2-2b`	Gemma Scope PT residual canonical	16,384	4, 8, 13, 17, 22, 25
Gemma 2 9B	`google/gemma-2-9b`	Gemma Scope PT residual canonical	16,384	7, 13, 21, 27, 36, 40
Llama 3.1 8B	`meta-llama/Llama-3.1-8B`	Llama Scope LXR	32,768	5, 10, 16, 21, 27, 31
Gemma 3 1B	`google/gemma-3-1b-pt`	Gemma Scope 2 residual, big-L_0	16,384	4, 8, 13, 17, 22, 25
Gemma 3 4B	`google/gemma-3-4b-pt`	Gemma Scope 2 residual, big-L_0	16,384	5, 10, 17, 22, 29, 33
Gemma 3 12B	`google/gemma-3-12b-pt`	Gemma Scope 2 residual, big-L_0	16,384	7, 15, 24, 31, 41, 47

tokens assigned to that TR, including tokens at which it is silent,

$$x_{t,k} = \frac{1}{|B_t|} \sum_{i \in B_t} \max(a_{i,k}, 0), \tag{5}$$

with empty bins set to zero.

We store this representation for every dictionary feature, allowing later feature selection without either target. The LM target averages the final hidden state before the unembedding over the same tokens, $\boldsymbol{h}_t = |B_t|^{-1} \sum_{i \in B_t} \boldsymbol{h}_i^{\text{final}}$. Mean and summed next-token surprisal and dense layer means are also stored for the control analyses. The sensitivity analysis uses maximum, sum, and active-token-count pooling, which require token-level records. These were produced by a separate outcome-blind extraction over the union of the selected banks.

D Models and SAE releases

We use model-matched residual-stream dictionaries from Gemma Scope and Llama Scope (Gemma Team, 2024; Grattafiori et al., 2024; Lieberum et al., 2024; He et al., 2024), with Gemma Scope 2 releases for the listed Gemma 3 checkpoints (McDougall et al., 2025). We load the official parameter files directly and check each layer's encoder against the reference implementation. Feature indices apply only within a model, layer, and release. Matching indices do not identify shared features across models. Gemma 2 and Llama 3.1 form the confirmatory set, and Gemma 3 is a prespecified replication extension.

E Source-description feature banks

For Physical→Social, we select the feature bank from the Physical predictor matrix before reading Social. The reverse direction uses a separately selected Social bank. Within-description analyses instead rebuild the bank from each fold's purged outer-training TRs, excluding held-out TRs from selection.

To select a bank, we exclude features active in fewer than five source TRs or more than 95% of source TRs. We rank eligible features within each layer by $V_k = p_k(1 - p_k)$, where p_k is feature k's fraction of source TRs with nonzero activation. This score does not favour activation magnitude. Ties are broken by positive-activation RMS and then feature index. We interleave the six layer rankings round-robin, distributing the K features as evenly as possible across the six layers when enough are eligible. This defines nested banks at $K \in \{64, 128, 256, 512, 1024\}$. A bank ranked by activation prevalence provides a sensitivity analysis.

For correlation-pruned banks, we use complete-linkage clustering within each layer, based on absolute Spearman correlation at thresholds .70, .80, and .90. Each cluster is represented by its highest-ranked

Table 6: Eligible features and retained clusters for each source bank. Feature counts follow the layer order in Table 5. The last column counts complete-linkage clusters in the $K = 256$, $|\rho| = .80$ bank.

Model	Source	Eligible features by layer	Clusters
Gemma 2 2B	Physical	3488/1874/2053/1624/1612/119	246
	Social	3981/2120/1969/1762/1959/115	254
Gemma 2 9B	Physical	2178/2643/1928/1811/938/1253	250
	Social	2567/3249/2221/1995/1024/1368	254
Llama 3.1 8B	Physical	1479/1373/1398/1250/1216/11805	240
	Social	1701/1507/1515/1627/1693/18553	249
Gemma 3 1B	Physical	3840/4300/4315/3704/2365/45	239
	Social	4395/5088/5179/4787/3275/46	241
Gemma 3 4B	Physical	3248/3005/2507/3029/2847/34	248
	Social	3626/3386/2836/3814/3817/34	249
Gemma 3 12B	Physical	2865/2618/1832/1898/1939/32	244
	Social	3356/3009/1858/2127/2488/32	244

feature. The exact cluster memberships over the $K = 256$ variance bank define the groups for grouped ablation.

Four models fill the balanced layer allocation (Table 6). For Gemma 3 4B and 12B, the deepest sampled layer has only 34 and 32 eligible features against a quota of 43 at $K = 256$, so the other layers supply the remaining features. At $K = 1024$, the deepest layer runs out of eligible features for Gemma 2 2B and all three Gemma 3 models (Table 7). Llama 3.1 8B has the opposite pattern: its deepest layer has 11,805 and 18,553 eligible features, against roughly 1,500 elsewhere. The number of features shuffled within each layer therefore varies across models and bank sizes.

Table 7: Selected features per layer in the source-selected variance banks, following the layer order in Table 5.

Model	Source	Per-layer count, $K = 256$	Per-layer count, $K = 1024$
Gemma 2 2B	Physical	43/43/43/43/42/42	181/181/181/181/181/119
Gemma 2 2B	Social	43/43/43/43/42/42	182/182/182/182/181/115
Gemma 2 9B	Physical	43/43/43/43/42/42	171/171/171/171/170/170
Gemma 2 9B	Social	43/43/43/43/42/42	171/171/171/171/170/170
Llama 3.1 8B	Physical	43/43/43/43/42/42	171/171/171/171/170/170
Llama 3.1 8B	Social	43/43/43/43/42/42	171/171/171/171/170/170
Gemma 3 1B	Physical	43/43/43/43/42/42	196/196/196/196/195/45
Gemma 3 1B	Social	43/43/43/43/42/42	196/196/196/195/195/46
Gemma 3 4B	Physical	45/45/44/44/44/34	198/198/198/198/198/34
Gemma 3 4B	Social	45/45/44/44/44/34	198/198/198/198/198/34
Gemma 3 12B	Physical	45/45/45/45/44/32	199/199/198/198/198/32
Gemma 3 12B	Social	45/45/45/45/44/32	199/199/198/198/198/32

Most selected features remain active in the held-out description. Across the twelve $K = 256$ cells, median activation prevalence ranges from .38–.50 on the source description and .23–.51 on the held-out description. The median ratio of held-out to source activation standard deviation is .81–.96. Between 84.8% (Llama 3.1 8B, SOCIAL→PHYSICAL) and 96.9% of features are active in more than 5% of held-out TRs; at most 2% are silent throughout the held-out description in any cell.

An earlier bank, selected by reciprocal-rank aggregation over five other Narratives stories, is retained only as a control and does not enter any reported result.

F PREDICTION MODELS AND SCALING

For K features, the BOLD predictor uses the current and four preceding TRs, $\widetilde{\boldsymbol{x}}_t = [\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-4}] \in \mathbb{R}^{5K}$, with all five columns of each SAE feature forming one group. The predictor of the LM hidden state uses the unlagged $\boldsymbol{x}_t \in \mathbb{R}^K$. We fit ridge regression (Hoerl & Kennard, 1970) in dual form, standardizing predictors and targets using only training-row statistics and replacing zero training scales with one. We return predictions to target units before scoring. Predictors are standardized but not high-pass filtered. The BOLD target is filtered as described in Appendix B.

We select a separate penalty from $\{10^{-1}, 10^{0}, \ldots, 10^{10}\}$ for Pearson r and for out-of-sample

$$R^2_{\text{oos}} = 1 - \frac{\sum_{i \in \text{te}} (y_i - \widehat{y}_i)^2}{\sum_{i \in \text{te}} (y_i - \bar{y}_{\text{tr}})^2} \tag{6}$$

The denominator uses the training mean, so R^2_{oos} can be negative. Figure 4A reports Pearson r from the r-tuned models. Table 8 reports R^2_{oos} from the regressions used for the primary reliance profiles. In each transfer cell, 16 to 33 of 59 participants' r-tuned BOLD models select the largest penalty in the grid. Their predictions approach the training mean, with held-out R^2_{oos} near zero. The cell means across all participants range from $-.44$ to $-.10$.

Within-description evaluation. We evaluate each participant's predictor in five contiguous outer test blocks. Training excludes four TRs on either side of each test block, preventing training and test responses from sharing a lagged predictor. Inner folds split the remaining training rows into contiguous blocks, apply the same exclusion, and select the penalty (Varoquaux et al., 2017). We rebuild the feature bank in each fold and compute correspondence as the mean of the five fold-level Spearman correlations. We never average profiles by position across different feature identities.

Cross-description transfer. We choose the penalty most often selected across the five source folds, fit the final predictor on all source TRs, and apply it with the same scaling to the other description. Both descriptions use the same 59 participants. Transferred BOLD prediction reaches $r = .068$–$.092$ for r-tuned models and $R^2_{\text{oos}} = .003$–$.008$ for R^2_{oos}-tuned models. Predictors of the LM hidden state reach $r = .302$–$.475$ and $R^2_{\text{oos}} = .134$–$.271$.

The transferred BOLD r values exceed the within-description fold values of $.023$–$.062$, but the procedures differ: transfer trains on a full description and tests on all retained TRs of the other, whereas each within-description fold trains on four-fifths of a description, with purged boundaries and a fold-specific bank. The purge does not remove dependence from whole-run BOLD preprocessing (Appendix B).

G Controls

We compare a baseline with two predictor blocks against a model that adds the SAE bank as a third block. The first block contains word and phoneme counts, audio-envelope RMS and its first difference, and story-position ramp bases with plateaus from 3 to 51 TRs. The second contains mean and summed model-specific token surprisal per TR. Banded ridge assigns each block an independent penalty (Nunez-Elizalde et al., 2019; Dupré la Tour et al., 2022). A fixed random search chooses block weights within the same nested folds. We measure each participant's improvement from adding the SAE block, using paired sign-flip tests and bootstrap intervals. Benjamini–Hochberg correction covers the six models within each story or transfer direction.

Table 2 in the main text reports the paired increments and participant-bootstrap intervals.

Representation controls are dense residual-stream layer means and 99 random feature banks. The random banks match the selected bank on layer, activity-prevalence decile, and log-variance decile. Their results are included in the release but do not support any primary claim.

H Reliance and nulls

Reliance. We subtract each feature group's contribution from the fitted regression prediction, keeping the weights fixed. This is equivalent to setting the group's standardized held-out inputs to zero, their training mean. Reliance is the original R^2_{oos} minus the adjusted R^2_{oos}, averaged over target dimensions. Negative values mean the adjustment improves prediction. For transfer, BOLD reliance is averaged over participants; within a description, it is computed per fold. We compare the paired BOLD and LM reliance profiles primarily with Spearman correlation, and record Pearson correlation and top-12.5% overlap as sensitivity checks. For grouped reliance, we subtract the combined contribution of all features in a source-defined cluster, including every lag of every member for BOLD. Conditions with fewer than three groups are reported as unresolved.

Temporal null. We circularly shift the source and held-out LM targets by independent fixed offsets, each at least ten TRs from the original alignment. We draw offsets without replacement and never give the two targets the same signed offset. For each of 199 shift pairs, we select the penalty again, fit the LM predictor, transfer it, and ablate its features. We compare the resulting reliance profile with the fixed observed BOLD profile. Figure 3 shows the twelve per-cell distributions. For within-description analyses, we shift the single LM target and refit every fold.

We draw offsets once per direction and bank size and share them across the six models. Pooled statistics average null draws with the same shift index across cells. Shared offsets coordinate retiming across models within each direction and do not affect per-cell p values. Pooling nulls also share offsets across models, but shift only source LM targets. Primary nulls shift both source and held-out targets (Tables 3 and 13). With 199 shifts, the smallest one-sided p is $.005$.

Identity null. We shuffle the LM reliance profile's feature labels within each sampled layer for 10,000 draws. This preserves layer allocation but breaks exact feature pairing. Preserved between-layer structure can shift the null mean from zero: it ranges from $-.35$ to $+.02$ across cells (Table 10). Even a negative observed ρ can therefore pass this test. Passing shows that feature labels matter, but does not rule out agreement caused by shared timing.

The temporal null's minimum $p = .005$ limits multiple-testing correction. For ties at this minimum, two cells out of twelve or twelve out of 120 suffice for Benjamini–Hochberg $q \leq .05$; thirteen out of 120 suffice for strict $q < .05$. These are sufficient counts for tied minimum p values, not necessary counts when other small p values also contribute to the correction. We correct across the twelve primary cells and treat the 120-cell sweep as a robustness analysis.

Decision rule. The prespecified rule requires the global cross-description statistic to pass both null tests and be positive in both directions. It must persist at three adjacent bank sizes including 256, appear in at least two model families after correction, and agree between the variance-selected bank and a second construction. To interpret a non-significant result, conditional and grouped analyses must also remain null across bank size, selection, and pooling choices, with 80% detection at planted $\rho = .40$. Otherwise, a small effect cannot be distinguished from insufficient sensitivity. After examining the first results, we added a requirement that the bank improve held-out prediction beyond the stimulus baseline in the evaluation setting under study. This requirement was not prespecified.

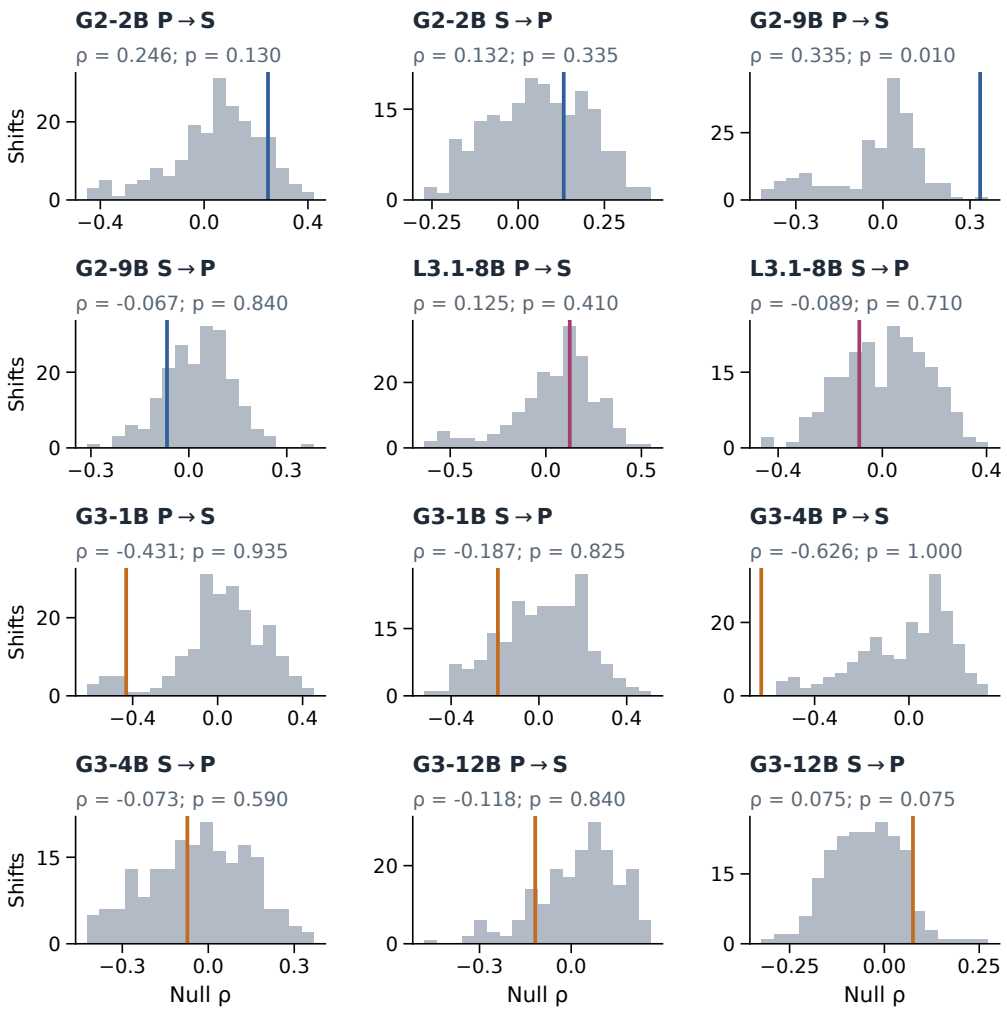

Figure 3: Temporal null distributions at $K = 256$. Each panel shows 199 correlations after retiming the LM targets and retuning the predictor for one model and direction. The vertical line marks the observed correlation (blue for Gemma 2, magenta for Llama 3.1, orange for Gemma 3).

I Detailed empirical results

I.1 SAE features do not improve BOLD prediction under transfer

Source-selected features predict both targets in twelve transfer cells. Predictors tuned for r reach mean parcelwise BOLD r of .07–.09 and LM r of .30–.47 (Figure 4A). Table 8 reports prediction from the regressions used to measure reliance. Their BOLD predictors explain less than one percent of held-out variance ($R^2_{\text{oos}} = .003$–$.008$), limiting the precision of hundreds of feature-ablation losses. Appendix F reports the negative R^2_{oos} of the r-tuned predictors and compares transfer with within-description evaluation.

Table 8: Prediction and correspondence at $K = 256$. BOLD and LM columns give held-out R^2_{oos} from the R^2_{oos}-tuned regressions used to measure reliance; Spearman ρ compares their reliance profiles. Temporal p values use 199 retimed refits; identity p values use 10,000 within-layer permutations (null centres in Table 10). The final two columns give post-hoc prevalence-adjusted correspondence and its matched temporal p (Section 4.4). All p values are one-sided.

Model	Direction	R^2_{oos}		ρ	p_{temporal}	p_{identity}	ρ |prev.	p_{matched}
		BOLD	LM					
Gemma 2 2B	P→S	.0045	.237	.246	.130	$< .001$	.277	.040
Gemma 2 2B	S→P	.0067	.191	.132	.335	.029	.153	.290
Gemma 2 9B	P→S	.0068	.271	.335	.010	$< .001$	.195	.005
Gemma 2 9B	S→P	.0084	.203	$-.067$	.840	.836	.031	.510
Llama 3.1 8B	P→S	.0034	.233	.125	.410	.004	.121	.340
Llama 3.1 8B	S→P	.0046	.209	$-.089$	.710	.864	$-.124$	.760
Gemma 3 1B	P→S	.0072	.158	$-.431$	.935	.965	$-.002$	.610
Gemma 3 1B	S→P	.0067	.157	$-.187$	.825	.128	.070	.485
Gemma 3 4B	P→S	.0080	.134	$-.626$	1.000	1.000	$-.032$	.660
Gemma 3 4B	S→P	.0081	.150	$-.073$	.590	.012	.208	.020
Gemma 3 12B	P→S	.0069	.210	$-.118$	.840	.051	.190	.060
Gemma 3 12B	S→P	.0073	.162	.075	.075	$< .001$	.216	.005
All twelve	mean	.0065	.193	$-.056$	.805	–	.109	.060

To test whether the SAE features improve prediction beyond simpler stimulus features, we add them to a banded-ridge model alongside lexical, acoustic, positional, and surprisal covariates. Surprisal measures how unexpected a token is under the LM. Averaging across models and participants, adding the SAE features within a description increases R^2_{oos} by .0010 for PHYSICAL (95% participant-bootstrap interval $[.0003, .0018]$) and .0034 for SOCIAL ($[.0022, .0046]$). Changes in r are indistinguishable from zero in both descriptions ($[-.0045, .0049]$ and $[-.0007, .0104]$).

After transfer, adding the SAE features reduces mean prediction: $\Delta r = -.012$ and $-.011$ in the two directions, and $\Delta R^2_{\text{oos}} = -.002$ and $-.001$. Three of the four intervals lie entirely below zero (Appendix G). The features therefore show no added predictive value under this transfer procedure. Penalties and block weights chosen on the source can shrink the 1,280-column SAE block more than the smaller control blocks. This analysis cannot distinguish a lack of additional information from difficulty estimating it.

I.2 CORRESPONDENCE VARIES WITH MODEL FAMILY AND FEATURE PREVALENCE

At $K = 256$, the twelve reliance correlations range from $-.626$ to .335, with a mean of $-.056$ (Figure 4B). Their signs vary across model families and transfer directions. These correlations describe predictors whose SAE features do not improve transferred BOLD prediction beyond the stimulus baseline.

In the confirmatory set, Gemma 2 and Llama 3.1, mean correspondence is $+.114$, with four of six cells positive. This mean is not unusually high relative to the pooled temporal null ($p = .160$; Appendix H). The Gemma 3 extension has a mean of $-.227$, with five of six cells negative. Its mean is unusually low under its pooled null ($p = .030$ in the lower tail). Removing the two Gemma 3 4B cells changes the all-condition mean to $+.002$.

Feature prevalence is the fraction of held-out TRs in which a feature is active. It correlates positively with BOLD reliance in all twelve cells ($\rho = .025$–$.868$). LM associations are near zero in Llama, range from $-.761$ to .291 in Gemma 2 (including $-.383$ for 2B S→P), and are negative throughout Gemma 3 ($-.729$ to $-.283$). Selection by $p(1-p)$ already depends on source prevalence and may contribute to these held-out associations. Gemma 3's identity-null means are also negative ($-.15$ to $-.35$; Table 10). These systems use a different SAE release. In the 4B and 12B models, the deepest sampled layer has too few eligible features to fill its allocation (Appendix E).

Adjusting for prevalence raises mean correspondence to $+.109$ (range $-.12$ to .28; Table 8, Figure 4E). We remove prevalence's association with both observed reliance profiles and every temporal-null pro-

file before computing correspondence. The matched pooled null has a 95% interval of $[-.121, .125]$, with $p = .06$. Four cells have matched-null $p < .05$: Gemma 2 2B and Gemma 2 9B P→S, and Gemma 3 4B and 12B S→P. No model passes in both directions. This post-hoc adjustment used only $K = 256$ and was not tested against the identity null. It shows that prevalence affects the comparison, but does not establish shared reliance.

Correspondence also changes with predictor tuning and token pooling. Predictors tuned for r give a mean of $+.097$, but have no matched temporal null and predict worse than the training mean in R^2_{oos} terms (Appendix I). We therefore report their reliance profiles descriptively. Replacing mean activation within each TR with the maximum, sum, or active-token count raises the BOLD profile's mean correlation with prevalence from .53 to .71, while the LM profile's falls from $-.29$ to $-.40$. Mean correspondence becomes $-.102$, $-.189$, and $-.207$, respectively (Figure 4F). Jointly ablating correlated features gives a mean of $-.053$. Few features merge at $|r| = .80$, leaving 239–254 groups from 256 features.

I.3 No model passes both null tests in both directions

Shifting the LM targets in time produces mean correlations at least as high as the observed mean ($-.056$) in 160 of 199 draws ($p = .805$; Figure 4C). For each draw, we average correlations across the twelve model–direction pairs using the same shift index. These null means average .011, with a 95% interval of $[-.214, .157]$. The observed correspondence therefore provides no evidence of stronger agreement than this timing control. Grouped ablation gives $p = .760$. Across the 120-cell bank-size and bank-variant sweep, seven cells have uncorrected $p < .05$. Five involve the same model and direction across nested variants. The smallest Benjamini–Hochberg q is .20. We treat this sweep as a robustness analysis (Appendix H).

Six of twelve cells pass the identity null. One, Gemma 3 4B S→P, has a negative observed ρ but a still more negative null mean (Appendix H). Only Gemma 2 9B P→S passes both null tests ($\rho = .335$, $p_{\text{identity}} < .001$, $p_{\text{temporal}} = .010$). Its reverse direction does not ($\rho = -.067$, $p_{\text{identity}} = .836$, $p_{\text{temporal}} = .840$). No model therefore meets the requirement in both directions (Figure 4D).

Transfer directions use different feature banks, sharing 87–117 of 256 features per direction (Jaccard overlap: 20–30%). Restricting the comparison to shared features gives positive correlations in both directions for Gemma 2, near-zero correlations for Llama 3.1 8B, and negative correlations in both directions for every Gemma 3 model (Table 12). This descriptive analysis has no matched null, but suggests that bank composition contributes to the Gemma 2 9B reversal.

I.4 Sensitivity and reliability analyses

Changing the feature-bank size does not change the main finding. Across five nested banks, mean correspondence ranges from $-.112$ to $-.056$ and falls within the pooled temporal-null interval at every size. Ranking features by source prevalence instead gives mean $\rho = -.093$ ($p = .810$) at $K = 256$. Pruning correlated features gives a mean of $-.055$ ($p = .780$). Alternative token-pooling estimates also fall within the primary $K = 256$ null interval. Their own nulls shift only source LM targets and test a different hypothesis (Figure 4F, Table 13).

Inter-subject correlation (Nastase et al., 2019) measures response agreement across listeners. Its median is .16 and .20 for the two descriptions. Median between-listener correlations of reliance are .08 within a description and .31 under transfer (Appendix J). The underlying procedures differ, so these values do not directly compare reliability. The average BOLD transfer profile has a split-half correlation of .77. Agreement between participant groups does not show whether the ranking is accurate, and we have no corresponding transfer reliability estimate for the LM profile.

Averaging BOLD across participants before fitting raises mean-fold prediction to $r = .164$–$.267$. In this analysis, three of twelve within-description cells pass both null tests. These evaluations hold out time points but use the same listeners for training and testing, so they do not show whether the result generalizes to new listeners (Appendix J.1).

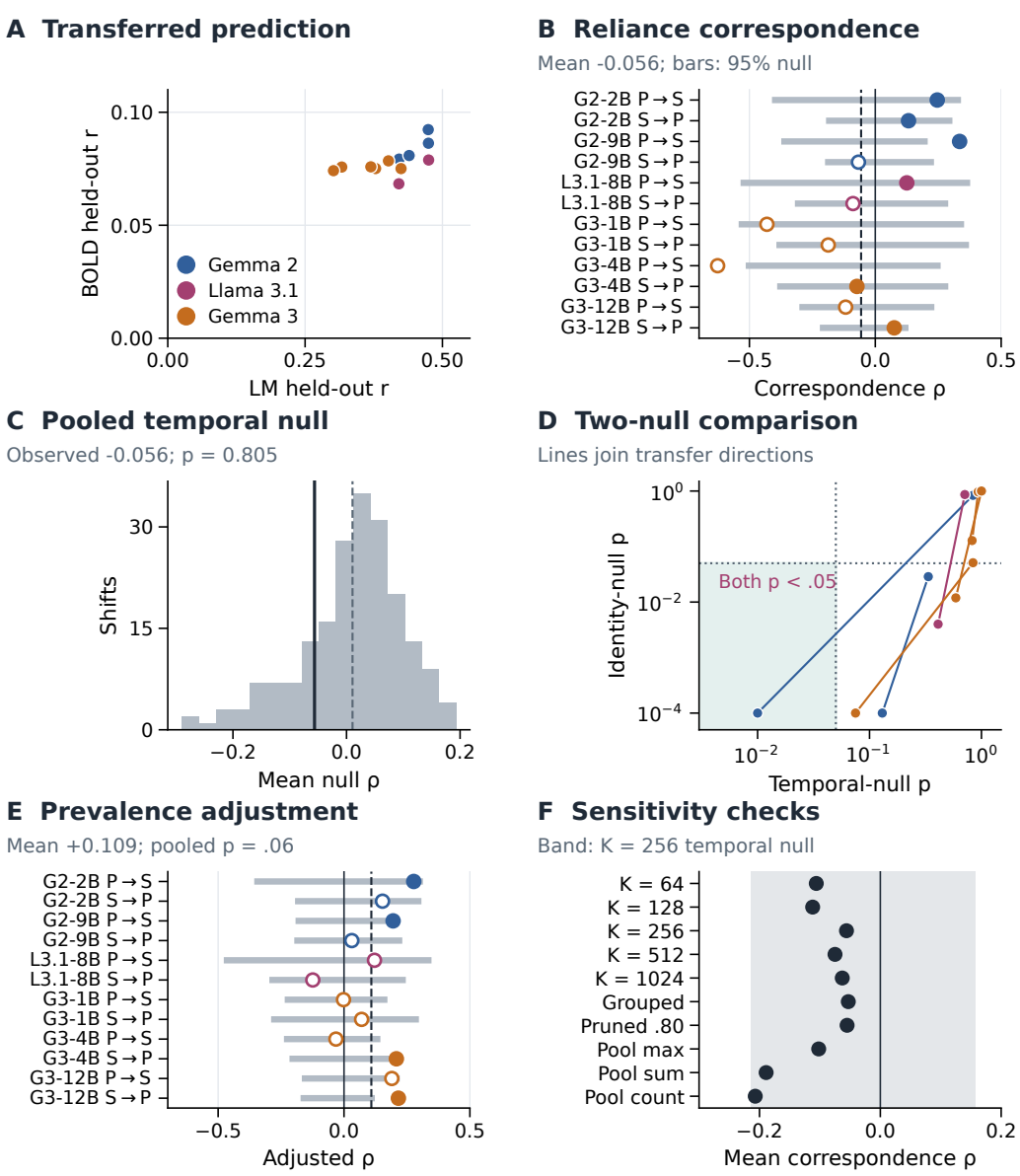

Figure 4: Prediction and reliance correspondence under transfer. G2/G3 denote Gemma 2/3; L3.1 denotes Llama 3.1. Blue, magenta, and orange distinguish Gemma 2, Llama 3.1, and Gemma 3. **A:** held-out BOLD and LM r from predictors tuned for r. **B:** per-cell ρ and 95% temporal-null intervals; filled points exceed their within-layer identity null. **C:** 199 pooled null means; the solid line marks the observed mean and the dashed line the null mean. **D:** temporal and identity p-values; lines join directions within a model; the shaded region has $p < .05$ on both axes. **E:** prevalence-adjusted ρ and matched 95% temporal-null intervals; filled points have $p < .05$ (post hoc, without an identity test). **F:** mean ρ under alternative settings; the band shows the $K = 256$ pooled 95% temporal-null interval.

I.5 Additional cell-level statistics

Table 10 gives the full $K = 256$ results, including alternative tuning, Pearson correlation, top-12.5% overlap, identity-null means, and associations with prevalence. Tuning for r changes mean Spearman correspondence to $+.097$, with ten of twelve cells positive. These predictors, however, perform

Table 9: Correspondence within each sampled layer (L1 shallowest to L6 deepest, in the order of Table 5), the median absolute BOLD ablation loss, and the fraction of features whose ablation reduces held-out R^2_{oos}.

Model	Dir.	L1	L2	L3	L4	L5	L6	median $\lvert I_k \rvert$	$I_k > 0$
Gemma 2 2B	P→S	.228	.272	−.029	.317	.406	.426	1.7×10^{-5}	0.50
Gemma 2 2B	S→P	.187	.252	.117	−.174	.080	.229	2.7×10^{-5}	0.37
Gemma 2 9B	P→S	.393	.433	.590	.119	.315	.544	1.8×10^{-5}	0.64
Gemma 2 9B	S→P	−.034	.136	−.191	−.167	−.073	−.086	2.2×10^{-5}	0.33
Llama 3.1 8B	P→S	.121	.345	.086	−.239	.090	.386	1.0×10^{-5}	0.64
Llama 3.1 8B	S→P	.249	.059	−.126	−.384	−.216	.410	1.3×10^{-5}	0.86
Gemma 3 1B	P→S	−.085	−.039	−.248	−.152	.005	−.782	1.7×10^{-5}	0.57
Gemma 3 1B	S→P	.299	.179	−.123	.111	.032	.509	2.4×10^{-5}	0.41
Gemma 3 4B	P→S	−.412	−.627	−.503	−.289	−.440	−.616	1.1×10^{-5}	0.71
Gemma 3 4B	S→P	.550	.021	.029	−.103	.140	−.382	1.9×10^{-5}	0.56
Gemma 3 12B	P→S	.237	−.241	.319	−.039	.128	−.428	1.4×10^{-5}	0.61
Gemma 3 12B	S→P	.285	.416	.071	.315	.281	.175	1.7×10^{-5}	0.43

Table 10: Per-cell statistics at $K = 256$. "prev." is the fraction of held-out TRs in which a feature is active.

Model	Dir.	ρ	$\rho_{r\text{-tuned}}$	Pearson	top-12.5%	id.-null mean	ρ(BOLD, prev.)	ρ(LM, prev.)
Gemma 2 2B	P→S	.246	.025	.139	.208	−.004	.619	.046
Gemma 2 2B	S→P	.132	.022	.208	.123	.019	.025	−.383
Gemma 2 9B	P→S	.335	−.142	.290	.143	−.048	.681	.291
Gemma 2 9B	S→P	−.067	.259	−.181	.032	−.008	.114	−.761
Llama 3.1 8B	P→S	.125	−.032	.185	.123	−.033	.535	.043
Llama 3.1 8B	S→P	−.089	.180	.058	.185	−.029	.464	.047
Gemma 3 1B	P→S	−.431	.170	−.362	.000	−.353	.868	−.495
Gemma 3 1B	S→P	−.187	.189	−.039	.143	−.243	.560	−.427
Gemma 3 4B	P→S	−.626	.219	−.499	.000	−.306	.843	−.729
Gemma 3 4B	S→P	−.073	.100	.106	.103	−.191	.522	−.444
Gemma 3 12B	P→S	−.118	.094	−.205	.000	−.189	.693	−.355
Gemma 3 12B	S→P	.075	.075	.193	.032	−.154	.405	−.283
Mean		−.056	.097	−.009	.091	−.128	.527	−.287

worse than the training mean in R^2_{oos} terms (Appendix F), and their correspondence has no matched temporal null. Pearson correlations average $-.009$, and top-12.5% overlap ranges from .000 to .208.

Table 9 reports correspondence within each sampled layer, along with the magnitude and sign of BOLD ablation losses. Within-layer correlations average $+.05$ and range from $-.78$ to .59. The deepest layer has the largest correlation in four of twelve cells. Median absolute BOLD ablation loss is only 1–3 $\times 10^{-5}$ in R^2_{oos} units. The fraction of features with positive reliance ranges from .33 to .86. In four cells, most ablations improve prediction. The profiles thus rank small changes of both signs, making their ordering difficult to estimate. Calibration tests recovery under these conditions. We have no transfer reliability estimate for the LM profile because only one profile is produced per direction.

BOLD reliance correlates positively with prevalence in all twelve cells (ρ = .025–.868). LM associations are near zero in Llama, mixed in Gemma 2 ($-.761$ to .291, including $-.383$ for 2B S→P), and negative throughout Gemma 3. We adjust both profiles for prevalence using rank-based partial correlation, then apply the same adjustment to each of the 199 stored retimed LM profiles to form a matched temporal null (Table 11). Mean correspondence rises from $-.056$ to $+.109$, and the two extreme Gemma 3 cells move to $-.002$ and $-.032$. The adjusted mean remains within the matched pooled 95% null interval $[-.121, .125]$: eleven of 199 shifts equal or exceed it, giving $p = .06$.

Four cells have matched-null $p < .05$: Gemma 2 2B and Gemma 2 9B P→S, and Gemma 3 4B and 12B S→P. None passes in both directions. This post hoc adjustment used only $K = 256$ and was not tested against the identity null, so it does not satisfy the decision rule. The sign change shows that prevalence affects the comparison. It does not establish that the adjustment reveals shared reliance.

For maximum, sum, and active-token-count pooling, the BOLD profile's mean correlation with prevalence is .64, .62, and .71, while the LM profile's is $-.28$, $-.39$, and $-.40$. These opposite

associations accompany more negative correspondence. Table 12 reports how many features the two directions' banks share and correspondence computed on those features alone.

Table 11: Correspondence after adjusting both reliance profiles for held-out feature prevalence using rank-based partial correlation. The matched null applies the same adjustment to each retimed LM profile. p is one-sided against 199 shifts.

Model	Dir.	ρ given prevalence	matched null 95%	p
Gemma 2 2B	P→S	.277	[−.344, .301]	.040
Gemma 2 2B	S→P	.153	[−.182, .295]	.290
Gemma 2 9B	P→S	.195	[−.180, .156]	.005
Gemma 2 9B	S→P	.031	[−.185, .219]	.510
Llama 3.1 8B	P→S	.121	[−.465, .335]	.340
Llama 3.1 8B	S→P	−.124	[−.284, .233]	.760
Gemma 3 1B	P→S	−.002	[−.223, .160]	.610
Gemma 3 1B	S→P	.070	[−.277, .285]	.485
Gemma 3 4B	P→S	−.032	[−.226, .132]	.660
Gemma 3 4B	S→P	.208	[−.204, .200]	.020
Gemma 3 12B	P→S	.190	[−.155, .209]	.060
Gemma 3 12B	S→P	.216	[−.160, .110]	.005
Mean		.109	[−.121, .125]	.060

Table 12: Features shared by the two directions' $K = 256$ banks and correspondence computed on those features alone. Counts are out of 256 per direction; the Jaccard index over the union of the two banks is .20–.30.

Model	Shared features	ρ P→S (shared)	ρ S→P (shared)	ρ full banks
Gemma 2 2B	117	.276	.236	.246, .132
Gemma 2 9B	94	.209	.212	.335, −.067
Llama 3.1 8B	106	.099	−.023	.125, −.089
Gemma 3 1B	103	−.589	−.394	−.431, −.187
Gemma 3 4B	87	−.573	−.604	−.626, −.073
Gemma 3 12B	107	−.257	−.111	−.118, .075
Mean		−.127		−.056

J RELIABILITY AND SENSITIVITY ANALYSES

Response reliability. Leave-one-participant-out correlations measure how much of the cleaned BOLD response is shared across listeners. The median across participant–parcel correlations is .162 for PHYSICAL and .197 for SOCIAL. For each of 10,000 splits, we average the split-half correlations across parcels. The medians across splits are .542 and .612 (Spearman–Brown corrected .703 and .759). These are not acquisition-level noise ceilings.

Reliance reliability. We compare profiles only when they contain the same features. At $K = 256$, median pairwise Spearman correlation across participants is .076 within description (mean .088) and .310 under transfer. Correlations between split-half group profiles are .373 and .773, respectively. These values are not directly comparable: within-description profiles come from five short folds with different banks, whereas transfer uses one bank and every held-out TR. The higher transfer correlation therefore does not establish greater reliability. The different fold-specific banks also prevent comparison of the LM predictor's profiles across within-description folds.

Within-description correspondence from participant-level fits. At $K = 256$, the twelve fold-mean correlations average −.028. Two cells pass the within-description temporal null: Llama 3.1 8B on SOCIAL ($\rho = .308$, $p = .005$) and Gemma 3 12B on SOCIAL ($\rho = .009$, $p = .020$). None replicates across descriptions or families.

Table 13: Token-pooling sensitivity. Each row averages six models and two directions. These pooled temporal nulls share offsets across models but shift only source LM targets, unlike the primary test in Table 3. The finite-sample denominator is 200.

Sensitivity	Setting	Mean ρ	p_{temporal}
Pooling	mean positive activation	$-.056$	.720
	maximum positive activation	$-.102$	.910
	summed positive activation	$-.189$	.970
	active-token count	$-.207$	.965

The pooling rows fit token-level records but shift only source LM targets. Their p values therefore test a different null and are not directly comparable with the primary p values, even though mean pooling matches the primary mean correspondence.

J.1 Participant-averaged BOLD

We average cleaned, run-standardized BOLD responses across uncensored participants at each TR and repeat the within-description analysis with the original predictors, fold-specific banks, a fully retuned temporal null, and the identity null. Training and testing use different time points from the same listeners. All twelve cells are exploratory: they test prediction at held-out times, not generalization to new listeners.

Table 14: Participant-averaged BOLD results at $K = 256$. Prediction is the mean held-out fold correlation with the averaged BOLD target, using the r-tuned predictor. R^2_{oos} uses the R^2_{oos}-tuned predictor, and ρ is the fold-mean reliance correspondence. The one-sided p values compare against 199 retuned shifts and 10,000 within-layer permutations. Time is held out; participants are not.

Model	Description	mean-fold r	R^2_{oos}	ρ	p_{temporal}	p_{identity}
Gemma 2 2B	Physical	.184	.013	.096	.095	.007
Gemma 2 2B	Social	.239	.056	.018	.075	.193
Gemma 2 9B	Physical	.171	$-.002$	$-.011$	.500	.870
Gemma 2 9B	Social	.241	.081	$-.098$	.980	.979
Llama 3.1 8B	Physical	.164	.010	.141	.005	$< .001$
Llama 3.1 8B	Social	.267	$-.033$	.283	.005	$< .001$
Gemma 3 1B	Physical	.168	.037	.006	.475	.243
Gemma 3 1B	Social	.242	.074	$-.073$	.830	.296
Gemma 3 4B	Physical	.165	.050	$-.025$	.760	.150
Gemma 3 4B	Social	.239	.068	.004	.435	.042
Gemma 3 12B	Physical	.166	.025	.078	.015	$< .001$
Gemma 3 12B	Social	.248	.041	$-.022$	.085	.336

Averaging raises prediction from $r = .023$–$.062$ to $.164$–$.267$ (Table 14), consistent with reducing listener-specific noise, but R^2_{oos} remains near or below zero in most cells. Three cells pass both null tests: Llama 3.1 8B on both descriptions ($\rho = .141$ and $.283$, $p_{\text{temporal}} = .005$), and Gemma 3 12B on PHYSICAL ($\rho = .078$, $p_{\text{temporal}} = .015$). The grouped Llama variant also passes. These within-description results do not meet the rule requiring transfer in both directions. One passing cell has $R^2_{\text{oos}} \leq 0$: its ablation losses compare predictors that do no better than the training mean.

Averaging retains responses locked to the stimulus, and every listener contributes to training and testing. The LM predictor's reliance profile does not vary by participant, so reusing listeners does not directly enter its rank agreement with the BOLD profile. Nevertheless, this analysis neither establishes generalization to new listeners nor tests whether the SAE bank improves on the stimulus baseline. A future analysis should fit to a BOLD average from training listeners and evaluate prediction on different listeners, testing improvement over the baseline first.

Table 15: Corrected calibration: detection of planted ablation-loss correspondence by design, direction, and signal scale (multiples of the residual standard deviation). Detection uses the 95th-percentile threshold of the $\rho = 0$ replicates. Realized planted correlations are .208–.233, .404–.425, .606–.621, and .804–.809 at the four nominal levels. Detection at $\rho = 0$ is .050 by construction.

Design	Direction	Scale	$\rho = .2$	.4	.6	.8	max	null 95th
Gemma 2 2B pruned	P→S	1	.076	.052	.072	.056	.076	.138
Gemma 2 2B pruned	P→S	2	.092	.076	.092	.080	.092	.123
Gemma 2 2B pruned	S→P	1	.076	.048	.080	.100	.100	.110
Gemma 2 2B pruned	S→P	2	.024	.052	.032	.064	.064	.122
Gemma 2 2B variance	P→S	1	.036	.036	.072	.052	.072	.140
Gemma 2 2B variance	P→S	2	.060	.076	.072	.076	.076	.133
Gemma 2 2B variance	S→P	1	.060	.060	.072	.048	.072	.124
Gemma 2 2B variance	S→P	2	.068	.080	.068	.076	.080	.114
Llama 3.1 8B pruned	P→S	1	.060	.044	.064	.044	.064	.141
Llama 3.1 8B pruned	P→S	2	.044	.076	.064	.064	.076	.128
Llama 3.1 8B pruned	S→P	1	.036	.040	.044	.060	.060	.157
Llama 3.1 8B pruned	S→P	2	.052	.056	.056	.104	.104	.144
Llama 3.1 8B variance	P→S	1	.040	.052	.040	.068	.068	.148
Llama 3.1 8B variance	P→S	2	.036	.016	.028	.040	.040	.146
Llama 3.1 8B variance	S→P	1	.072	.064	.064	.076	.076	.143
Llama 3.1 8B variance	S→P	2	.048	.060	.056	.092	.092	.143

K CALIBRATION DESIGN

The calibration combines Gemma 2 2B and Llama 3.1 8B, two $K = 256$ banks (variance-selected and correlation-pruned), and both transfer directions, giving eight designs. Each uses the held-out description's predictor matrix restricted to the source-selected bank. BOLD predictors are lag-expanded and repeated across 59 participants. LM predictors are unlagged. We retain the participant-by-time BOLD folds, TR-level LM folds, four-TR purge, censoring pattern, and feature groups. Residual pools contain actual responses minus transferred predictions.

Planted correspondence. We plant correlation between ablation-loss profiles, the measured quantity. For the planted linear predictor, feature group k has noiseless loss $\beta_k^2\, c_k$, where $c_k = \|\sum_{j \in G_k} \tilde{x}_j\|^2$ on the standardized design. Loss depends on coefficient magnitude, not sign. We generate two positive loss profiles p, q with a prescribed Spearman correlation by reversing a rank block in a random ordering. Setting $\beta_k = s_k \sqrt{p_k / c_k}$ for the BOLD predictor and using q analogously for the LM predictor makes their planted noiseless loss profiles p and q. We draw signs s independently for the two systems.

We record the realized rank correlation of the planted profiles in every replicate. Features that are silent in the held-out description receive zero coefficients and tie at the bottom in both systems; this affects four of 256 features in the Gemma designs. Random loadings mix each signal into every target dimension. We scale the signals to one or two times each system's mean per-target residual standard deviation, then add them to resampled, recentred residual rows. Independent row resampling removes temporal autocorrelation. Within each outer fold, we reselect the ridge penalty, fit the predictor, and compute held-out grouped ablation losses. We then correlate the recovered profiles. All fitting occurs within this matrix, so the simulation does not reproduce training on one narration and testing on the other.

For each design and signal scale, we run 500 replicates at $\rho = 0$ and 250 at each of $\rho \in \{.2, .4, .6, .8\}$. Detection is the fraction of replicates at positive planted correlation that exceed the 95th percentile of $\rho = 0$ replicates for the same design and scale. The grid contains eight designs, 80 conditions, and 192 validated shards. It does not repeat bank selection or temporal-null testing, preserves no residual autocorrelation, and uses one planted profile per system across target dimensions and participants. These choices are intended to favor recovery; they do not establish a formal upper bound on the full test's sensitivity. Calibration covers only Gemma 2 2B and Llama 3.1 8B.

Result. No condition approaches the 80% detection criterion. Across the 64 positive conditions, detection ranges from .016 to .104 (Table 15). Mean recovered correlation ranges from $-.001$ to .029 at $\rho = 0$ and from .004 to .039 at $\rho = .80$, with replicate standard deviations of .06–.10 and null thresholds of .11–.16.

For Gemma 2 2B physical→social diagnostics at one residual standard deviation, mean R^2_{oos} across folds ranges from $-.04$ to .19 over the diagnostic replicates and targets. The true predictor would reach approximately .5 at this signal scale. Of 80 fold-specific penalty choices across eight replicates and two targets, 77 select 10^3 and three select 10^4, indicating strong shrinkage. The supplementary archive includes these summaries and code to reproduce the quoted numbers. The within-description results in Appendix J were not recalibrated.

Superseded first grid. The first calibration used two signed, centred rank-score profiles with a prescribed Spearman correlation directly as coefficients. Because ablation losses depend on coefficient magnitude, this construction did not control their correlation. The resulting noiseless loss correlation was .76 at nominal $\rho = 0$ and .55 at nominal $\rho = .80$.

Despite this defect, the first grid's detection curve rose from .05 to .91 for Gemma 2 2B. Fitted ablation losses include a signed cross-term, and reversing a rank block made coefficient signs agree more often as nominal correlation increased. The statistic therefore tracked sign agreement rather than recovery of the planted loss ordering. With corrected loss profiles and shared signs, two Gemma 2 2B physical→social diagnostic replicates using different random seeds gave correlations of .41 and .12 at nominal $\rho = 0$. Independent signs gave .00 and $-.11$ at $\rho = 0$, and .06 and $-.06$ at $\rho = .80$.

We retain the first grid with a defect notice but exclude it from the reported findings. The corrected calibration was introduced after review. It changes only how the profiles are planted and imports the original module for all remaining stages. The original shards and aggregates record both module hashes. Public copies omit these analysis-tree hashes and have separate file-checksum catalogs.

K.1 Recovery of planted reliance profiles (exploratory)

To examine the effect of bank size, we run an exploratory sweep at planted $\rho = .80$ and one residual standard deviation. It covers eight transfer designs and four within-description designs: Gemma 2 2B and Llama 3.1 8B on each description. At each size, we sample K features without replacement from the selected $K = 256$ bank and keep the fitting and ablation procedure unchanged. The smaller banks are random subsets, not nested prefixes.

We measure both how well each fitted profile recovers its own planted ranking (self-recovery) and how strongly the two fitted profiles correlate (recovered correspondence). Additional conditions replace the predictors with independent Gaussian entries or the empirical residuals with white noise. These controls retain matrix dimensions and the fitting procedure, but injected realizations also differ between predictor conditions. This sweep was not prespecified. Its scripts and outputs are released with the calibration files.

Table 16: Exploratory bank-size sweeps at one residual SD. Entries are medians; self-recovery compares fitted BOLD reliance with its planted ranking. Separate experiment: eight designs, five seeds. Paired transfer: eight designs, three seeds; its two columns share replicates. Within description: four designs, five seeds. Both paired experiments plant cross-system $\rho = .80$. Feature subsets and injections differ across experiments.

K	Separate experiment	Paired transfer experiment		Within description
	Self-recovery	Self-recovery	Recovered ρ	Recovered ρ
24	.990	.920	.556	.683
32	.938	.864	.449	.550
48	.813	.632	.394	.410
64	.601	.383	.182	.188
96	.385	.210	.119	.104
128	.278	.178	.115	.039
192	–	.096	.060	$-.003$
256	–	.040	.023	.037

Recovery declines as the bank grows. In the separate self-recovery experiment, BOLD self-recovery falls through .90 near $K = 38$, .80 near $K = 49$, and .50 near $K = 74$. LM self-recovery falls through .50 near $K = 58$. Recovering agreement between the two fitted profiles is harder: with planted .80, median recovered correspondence falls from .56 at $K = 24$ to .02 at $K = 256$ for source-selected transfer designs, and from .68 to .04 for within-description designs (Table 16).

Changing the predictors has a much larger effect than changing the residuals. Replacing empirical residuals with white noise changes brain-profile self-recovery by at most .01 in every design. Replacing the SAE predictors with independent Gaussian entries of the same shape gives median BOLD self-recovery of .998 with empirical residual noise and .999 with white noise at $K = 256$ (Table 1B). LM self-recovery remains lower, at .302 and .322, respectively. This control measures individual rankings, not cross-system correspondence. A separate surface sweep gives median Gaussian BOLD self-recovery of .95 with 3.5 rows per column, four times fewer than the full real design has. With real predictors, $K = 16$ at one-quarter of the rows has about 56 rows per column, whereas $K = 256$ at all rows has about 14. This comparison changes both bank size and rows per column.

The Gaussian comparison changes several properties at once. Although median pairwise SAE-column correlation is modest ($|r| \approx .05$), replacing the matrix also changes its temporal structure and repeated rows. We therefore cannot separate feature correlation from the number of independent observations, infer a universal feature-count limit, or conclude that recovery is independent of sample size. At $K = 256$, median recovered cross-system correlation is near zero despite a planted .80, leaving the empirical correspondence in Section 4.4 inconclusive.

L Cortical distribution of transferred prediction

For Gemma 2 2B at $K = 256$, transferred prediction follows nearly the same cortical pattern in both directions. The across-parcel Pearson correlation between the maps is .99 (Spearman .92), although the descriptions share about one content word in ten. Accuracy peaks in bilateral parcels that include early auditory cortex and falls to near zero in dorsal and medial parcels. These results describe prediction from the SAE bank, which does not show an improvement over the stimulus baseline in either transfer direction. Only Gemma 2 2B is included in this cortical analysis.

M Reproducibility checks and released files

Released tests cover fold purging, training-only scaling, the training-mean R^2_{oos} baseline, ridge fitting, grouped ablation, and the identity null. The released code also implements the temporal null. Checksum catalogs and numerical reports accompany the saved results. The supplementary reporting script verifies its evidence hashes and reproduces Table 1 and the diagnostic summaries without fitting models.