Crosscoders Identify Shared or Specific Features between the Human Brain and Language Models

Koshiro Aoki [1] **Itsuki Hamada** [1] **Naho Orita** [1] **Daisuke Kawahara** [1] **Hiromu Sakai** [1]

Abstract

To what extent do human brains and language models (LMs) share internal representations of language, and how do these representations differ? Prior work has shown that LM representations can predict brain responses to naturalistic language stimuli, suggesting that the two systems encode common information. However, it has remained unclear which features are shared between brain and LM representations and which are specific to one of them. We propose Brain-LM crosscoders, which decompose brain responses and LM representations into a common set of sparse features and label each feature as shared, brain-specific, or LM-specific based on its predictive contribution to each representation. Experiments on fMRI data recorded during naturalistic story listening show that language associated with body, family, and action tends to be brain-specific, whereas colloquial expressions tend to be LM-specific. Brain-LM crosscoders compare biological and artificial language representations at the feature level and provide a method for scientific discovery in both neuroscience and LM interpretability research.[1]

1. Introduction

Language neuroscience and language model (LM) interpretability studies share an interest in how language stimuli are represented inside brains and models, respectively. Encoding models that predict brain responses from stimulus features have been a standard method in neuroscience for understanding the nature of internal representations. In early work, models used text corpus statistics to predict fMRI activation patterns for individual words (Mitchell et al., 2008), and voxelwise models mapped semantic selectivity across cortex during natural speech (Huth et al., 2016). Contextual representations from recurrent networks further improved prediction (Jain & Huth, 2018). Subsequent studies showed that pretrained language models predict brain responses to naturalistic stimuli (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022), even when trained on a developmentally realistic amount of data (Hosseini et al., 2024). These results indicate that LM representations and brain responses encode shared information extracted from the same linguistic stimuli.

Encoding models are effective for measuring the degree to which LM representations approximate brain responses. However, it remains difficult to distinguish features that are shared between the brain and LMs from features that are represented in only one of them. Studies analyzing prediction errors can identify aspects of brain responses that LMs fail to capture (Zhou et al., 2024). This approach is inherently asymmetric because it cannot extract features that are represented in LMs but absent in the brain. A symmetric comparison at the feature level is therefore needed to identify where the two representations align and where they diverge.

We propose *Brain-LM crosscoder*, a sparse decomposition that maps fMRI responses and LM representations to a shared sparse latent vector and reconstructs each representation (Figure 1). Each latent coordinate represents a candidate linguistic feature. For each feature, we measure the explained variance (EV) in fMRI responses and in LM representations. Brain-LM predominance is defined as the relative difference between these two EV values and indicates whether a feature predicts fMRI responses or LM representations better. A brain-specific feature explains more variance in fMRI responses than in LM representations, and an LM-specific feature shows the reverse pattern. Because brain-specific and LM-specific features are recovered in the same decomposition, this approach provides a symmetric, feature-level comparison that encoding models, which measure aggregate prediction performance in one direction, cannot offer.

We apply this method to naturalistic fMRI data from LeBel et al. (2023) with four LMs and three subjects (Section 4)

[1]Waseda University, Tokyo, Japan. Correspondence to: Koshiro Aoki <aokikoshiro@akane.waseda.jp>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

[1]Our code is available at `https://github.com/AokiKoshiro/brain-lm-crosscoder/`.

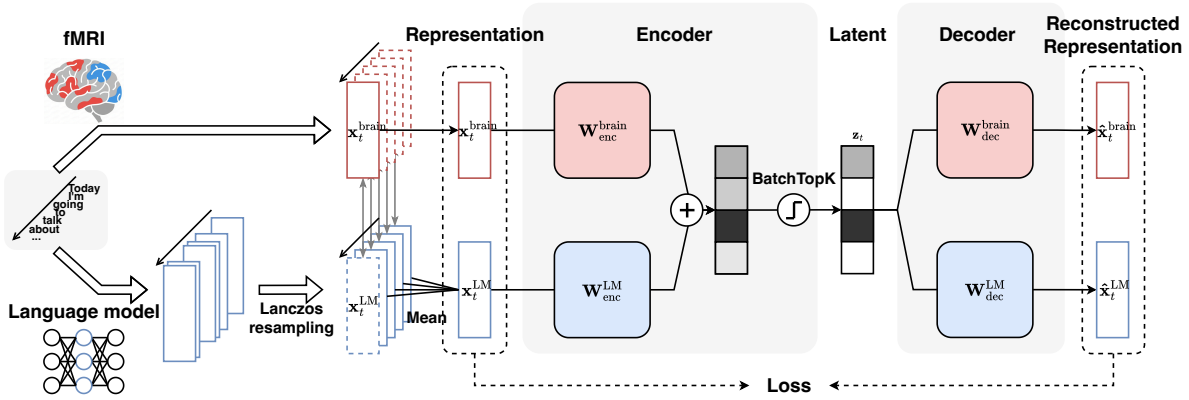


Figure 1. **Overview of Brain-LM crosscoder.** Time-aligned fMRI responses and language model (LM) representations for the same linguistic stimuli are encoded into a shared sparse feature vector. Token-level LM representations are resampled to the fMRI repetition-time (TR) grid with Lanczos resampling and averaged over four preceding TRs to account for hemodynamic delay. The crosscoder reconstructs each representation with a separate decoder. For each feature, leave-one-out explained variance (EV) in brain responses and LM representations is computed, and the relative difference Δ^{EV} classifies features as shared ($|\Delta^{\mathrm{EV}}| \leq 0.9$), brain-specific ($\Delta^{\mathrm{EV}} > 0.9$), or LM-specific ($\Delta^{\mathrm{EV}} < -0.9$).

and report three findings. First, we found that most shared features disappear when the stimulus alignment between brain responses and LM representations is randomly shuffled, which suggests that these features reflect shared subspaces of the representations rather than artifacts. Second, an automatic interpretability analysis shows that around 40% of the learned features have explanations that predict their activations better than chance level. Third, we found that language related to body, family, and action is associated with brain-specific features, whereas colloquial markers and fillers are associated with LM-specific features.

2. Related work

Brain encoding models. An encoding model is a computational model that describes how input stimuli determine the output response of cognitive systems. In language neuroscience, encoding models predict brain responses from stimulus features (Mitchell et al., 2008; Huth et al., 2016). Contextual LM representations improved prediction over static word embeddings (Jain & Huth, 2018). Subsequent work studied layerwise alignment, scaling, and the relation between next-word prediction and brain predictivity (Toneva & Wehbe, 2019; Antonello et al., 2023; Caucheteux et al., 2023; Merlin & Toneva, 2024; Lei et al., 2025; Lopez-Cardona et al., 2025). Our method builds on these approaches by decomposing brain and LM representations into sparse latent features and characterizing each feature as brain-specific or LM-specific based on its predictive contribution to each representation.

Divergences between brains and language models. Zhou et al. (2024) analyzed words with large encoding errors and found that LMs fail to capture social, emotional, and physical aspects of brain responses. Xu et al. (2025) compared conceptual word ratings between humans and large language models (LLMs) and found that the similarity between humans and LLMs decreases from non-sensorimotor to sensory to motor domains, a result that suggests sensorimotor grounding is a major source of the gap. Mahowald et al. (2024) argued that LLMs show strong formal linguistic competence, whereas functional competence remains uneven and often requires augmentation. However, these approaches characterize divergences along predefined categories or only in the direction from LMs to the brain. They do not identify features that LMs capture but the brain does not, or the reverse. By treating the two representations symmetrically, our approach, the Brain-LM crosscoder, identifies both brain-specific and LM-specific features.

Dictionary learning. Dictionary learning represents observed values as sparse linear combinations of basis vectors (Olshausen & Field, 1996; Aharon et al., 2006). In neuroscience, this technique has been applied to extract resting-state functional networks and brain regions from fMRI data in a data-driven manner (Eavani et al., 2012; Varoquaux et al., 2011; Abraham et al., 2013; Lee et al., 2011; Dohmatob et al., 2016). In LM interpretability, sparse autoencoders (SAEs), a form of dictionary learning, are used to obtain monosemantic, interpretable features from internal representations (Bricken et al., 2023; Huben et al., 2024). Recent work has proposed variants of SAEs, such as crosscoders, for identifying differences between layers or models (Lindsey et al., 2024; Minder et al., 2025; Thasarathan et al., 2025; Kassem et al., 2025; Boughorbel et al., 2025; Jiralerspong & Bricken, 2026). Other recent work has also proposed encoding models based on interpretable embeddings using dictionary learning (Zeng & Gallant, 2026; Lepori et al.,

2026) or yes-no question answering (Benara et al., 2024; Singh et al., 2025), but they are still encoding models and do not directly interpret the differences between brain and LM representations.

3. Method

Our method encodes brain responses and LM representations into a shared sparse latent space, from which both shared and representation-specific features can be extracted (Figure 1). Notation used in this paper is summarized in Appendix A.

3.1. Preparing brain and language model representations

For each time point t on the repetition-time grid, the method uses a brain response $\mathbf{x}_t^{\text{brain}} \in \mathbb{R}^{d_{\text{brain}}}$ and an LM representation $\mathbf{x}_t^{\text{LM}} \in \mathbb{R}^{d_{\text{LM}}}$ derived from the same stimulus transcript. Here d_{brain} is the number of voxels and d_{LM} is the dimensionality of the LM representation.

Representation of brain and language model. In this paper, the brain representation is the blood-oxygen-level-dependent (BOLD) response measured while a subject listens to stories. The response is sampled at every repetition time (TR) and treated as a vector over voxels. The LM representation is the residual stream (i.e., the output of the transformer block) at a specific layer when the same transcript of the story is provided as input to the LM.

Temporal alignment. Brain responses and LM representations are observed on different time grids. We resample token-level LM representations to the TR grid with Lanczos resampling, following prior work on encoding models (LeBel et al., 2023). This gives an LM vector aligned to each BOLD sample.

Hemodynamic delay. The BOLD response reflects neural activity after a delay. For each BOLD sample, we therefore use four preceding LM representations sampled on the TR grid, corresponding to approximately 2, 4, 6, and 8 seconds before the BOLD sample. These delayed LM representations are averaged. We hereafter refer to this averaged vector as the LM representation $\mathbf{x}_t^{\text{LM}}$.

Normalization. Brain and LM representations are centered and scaled with statistics from the training set so that both have a root sum square (RSS) of 1. This keeps the reconstruction objective from being dominated by the raw norm of either representation.

3.2. Brain-LM crosscoder

Crosscoders have been proposed for analyzing differences between models in LM interpretability, but prior work has been limited to comparisons between language models (Lindsey et al., 2024; Minder et al., 2025; Kassem et al., 2025; Boughorbel et al., 2025; Jiralerspong & Bricken, 2026). *Brain-LM crosscoder* extends this method to brain responses and LM representations as shown in Figure 1.

Architecture. Our main crosscoder architecture follows Minder et al. (2025). The Brain-LM crosscoder maps the brain response $\mathbf{x}_t^{\text{brain}}$ and the LM representation $\mathbf{x}_t^{\text{LM}}$ at time t into a shared d_{latent}-dimensional latent space. Applying a sparse activation function to this representation outputs a latent $\mathbf{z}_t \in \mathbb{R}^{d_{\text{latent}}}$, whose i-th coordinate $z_{t,i}$ corresponds to the activation of feature i at stimulus time t. The decoder has separate weights for each modality and reconstructs the brain response and the LM representation from the latent $\mathbf{z}_t$, as follows.

$$\mathbf{z}_t = \text{BatchTopK}\big(\mathbf{W}_{\text{enc}}^{\text{brain}}\mathbf{x}_t^{\text{brain}} + \mathbf{W}_{\text{enc}}^{\text{LM}}\mathbf{x}_t^{\text{LM}} + \mathbf{b}_{\text{enc}},\, k\big), \tag{1}$$

$$\hat{\mathbf{x}}_t^{\text{brain}} = \mathbf{W}_{\text{dec}}^{\text{brain}}\mathbf{z}_t, \tag{2}$$

$$\hat{\mathbf{x}}_t^{\text{LM}} = \mathbf{W}_{\text{dec}}^{\text{LM}}\mathbf{z}_t. \tag{3}$$

Here $\mathbf{W}_{\text{enc}}^{\text{brain}}$ and $\mathbf{W}_{\text{enc}}^{\text{LM}}$ are modality-specific encoder weights, $\mathbf{W}_{\text{dec}}^{\text{brain}}$ and $\mathbf{W}_{\text{dec}}^{\text{LM}}$ are modality-specific decoder weights, and $\mathbf{b}_{\text{enc}}$ is the encoder bias. The parameter k is the target number of active features. The model uses tied weights ($\mathbf{W}_{\text{enc}} = \mathbf{W}_{\text{dec}}^{\top}$). BatchTopK is the sparse activation function that retains only the Bk largest pre-activations across a batch of B examples and zeros out the rest, so that each example has k active features on average (Bussmann et al., 2024) (Section C.1).

Loss function. For one pair of training examples, the training loss is

$$\mathcal{L} = \|\mathbf{x}_t^{\text{brain}} - \hat{\mathbf{x}}_t^{\text{brain}}\|_2^2 + \|\mathbf{x}_t^{\text{LM}} - \hat{\mathbf{x}}_t^{\text{LM}}\|_2^2 + \text{TV}(\mathbf{W}_{\text{dec}}^{\text{brain}}) + \alpha\mathcal{L}_{\text{aux}}. \tag{4}$$

Here, $\text{TV}(\cdot)$ is a total variation penalty (Rudin et al., 1992; Michel et al., 2011) that penalizes large differences between adjacent voxel weights in the brain decoder and reduces overfitting (Section C.2). The term $\mathcal{L}_{\text{aux}}$ is an auxiliary loss that encourages inactive features to explain residual error, and α is its coefficient (Gao et al., 2025) (Section C.3). To reduce overfitting and make the learned feature vectors less dependent on weight initialization, we jointly train multiple crosscoders with different weight initializations while sharing their latent activations. At inference time, we use the average of their weights (Section C.4). This can be regarded as a form of cluster ensemble and consensus

clustering (Strehl & Ghosh, 2003; Monti et al., 2003; Fred & Jain, 2005).

3.3. Brain-LM predominance

Prior studies using crosscoders identified model-specific features by comparing the norms of the decoder weight vectors for each feature (Lindsey et al., 2024; Minder et al., 2025). However, decoder norm does not directly measure a feature's contribution to reconstruction. A feature that is present in both models but has a relatively small decoder norm in one model may be incorrectly classified as specific to the other model. This issue may be negligible when comparing closely related models such as a base model and its chat-tuned variant, but it requires careful consideration when comparing heterogeneous representations such as brain responses and LM activations. We therefore use cross-validated prediction performance to classify features as shared, brain-specific, or LM-specific.

Specifically, we use explained variance (EV) with leave-one-out cross-validation (LOOCV) for brain responses ($\mathrm{EV}_i^{\mathrm{brain}}$) and for LM representations ($\mathrm{EV}_i^{\mathrm{LM}}$) as the predictive performance of each feature i. Each EV measures the fraction of target variance that a one-dimensional linear projection from the feature activation recovers. We fit an ordinary least squares regression from the scalar activation $z_{t,i}$ to the target representations and compute the prediction residual for each hold-out data. The EV is defined as one minus the ratio of the total squared LOOCV residual to the predictable variance. The predictable variance is defined as the total variance minus the noise variance. For brain responses, the noise variance is estimated from repeated responses, so that the EV is normalized by the variance that is in principle predictable. For LM representations, noise correction is not applied because LM representations are deterministic. The closed-form expression and estimation details are given in Appendix D. A negative EV indicates that feature i does not generalize to hold-out time points.

When at least one of $\mathrm{EV}_i^{\mathrm{brain}}$ and $\mathrm{EV}_i^{\mathrm{LM}}$ is positive, we define *Brain-LM predominance* $\Delta_i^{\mathrm{EV}} \in [-1, 1]$ of feature i as

$$\Delta_i^{\mathrm{EV}} = \frac{\max(\mathrm{EV}_i^{\mathrm{brain}}, 0) - \max(\mathrm{EV}_i^{\mathrm{LM}}, 0)}{\max(\mathrm{EV}_i^{\mathrm{brain}}, \mathrm{EV}_i^{\mathrm{LM}})}. \quad (5)$$

A positive Δ_i^{EV} indicates that feature i predicts brain responses better than LM representations, and a negative Δ_i^{EV} indicates the reverse. When both $\mathrm{EV}_i^{\mathrm{brain}}$ and $\mathrm{EV}_i^{\mathrm{LM}}$ are non-positive, the feature does not generalize to either representation. In this case, the predominance is undefined, and we label the feature non-predictive.

We classify the remaining features into three categories: *shared* features satisfy $|\Delta_i^{\mathrm{EV}}| \leq 0.9$, *brain-specific* features satisfy $\Delta_i^{\mathrm{EV}} > 0.9$, and *LM-specific* features satisfy $\Delta_i^{\mathrm{EV}} < -0.9$.[2]

4. Experiments

We train Brain-LM crosscoders using BOLD responses of three subjects and internal representations of four LMs (Section 4.1), and analyze the learned features through the following experiments. First, we compare crosscoders trained on time-aligned representations of the brains and LMs with crosscoders trained on shuffled representations, to test whether shared features reflect shared information within representations derived from the same stimulus (Section 4.2). We then apply an automatic interpretation pipeline (Paulo et al., 2025) in which one LLM generates a description of each feature from stimulus transcripts that activate the feature and another LLM uses that description to predict hold-out activations (Section 4.3). Finally, we identify words in the stories that represent brain-specific and LM-specific features to characterize the two groups at a global level (Section 4.4).

4.1. Experimental setup

Dataset. We use the natural-language fMRI dataset of LeBel et al. (2023); Tang et al. (2023). Functional magnetic resonance imaging (fMRI) records brain activity by measuring changes in blood oxygenation. Three subjects (UTS01, UTS02, and UTS03)[3] listened to English podcast stories of roughly 10–15 minutes each while their brain responses were recorded. The BOLD signal is sampled every 2 seconds (TR = 2 s), yielding one brain response vector per time step. We use 58 stories for training, 8 for validation, and 17 for testing.

Language model representations. The language models used in our experiments are Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen3-8B, and Qwen3-32B (Team, 2025). For each model, we use the residual stream at the layer where the performance of a conventional encoding model is highest. The selected layers are 12 for Llama-3.1-8B-Instruct, 19 for Llama-3.1-70B-Instruct, 20 for Qwen3-8B, and 47 for Qwen3-32B (Appendix E).

Training configuration. Each crosscoder uses dictionary size $d_{\mathrm{latent}} = 128$ (equal to the number of features), BatchTopK sparsity $k = 4$, and five crosscoders with different initial weights for multi-start cluster ensemble. We train for 300 epochs with batch size 2048. Optimization uses Adam (Kingma & Ba, 2015) with learning rate 10^{-4}. The coeffi-

[2]These labels are operationally defined. They indicate which representation a feature predicts better under this pipeline, not that the feature is exclusively present in one system.

[3]UTS01 is a 24-year-old female, UTS02 is a 34-year-old male, and UTS03 is a 21-year-old male.

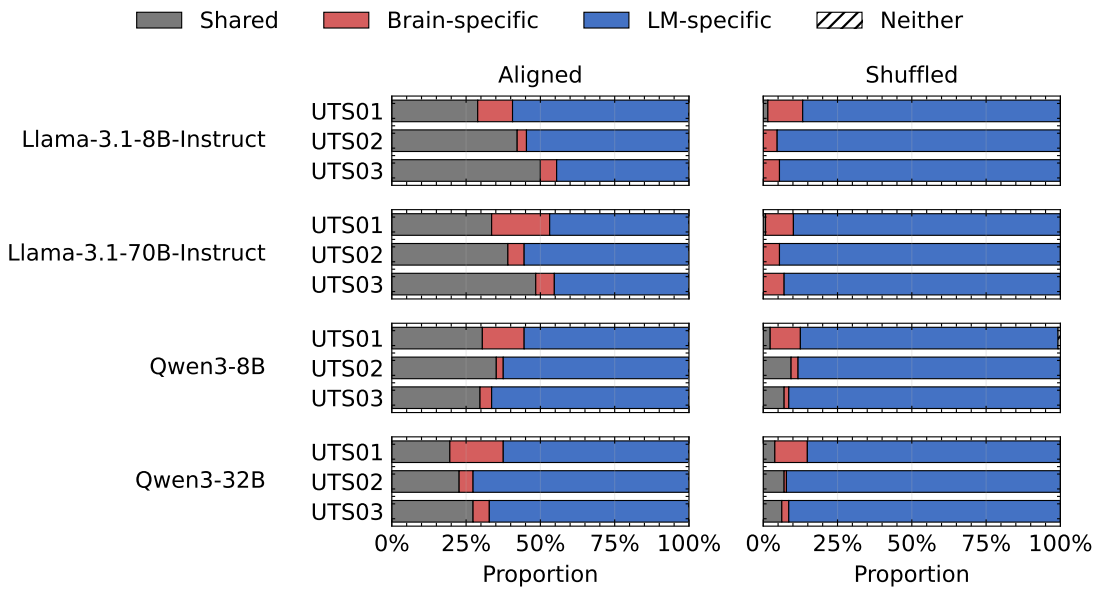


Figure 2. **Ratios of feature categories for Brain-LM crosscoders trained on three subjects and four language models.** Each bar shows the fraction of 128 features classified as shared ($|\Delta^{\mathrm{EV}}| \leq 0.9$), brain-specific ($\Delta^{\mathrm{EV}} > 0.9$), LM-specific ($\Delta^{\mathrm{EV}} < -0.9$), or neither (both EV values non-positive), where Δ^{EV} is the relative difference between leave-one-out explained variance for brain responses and for LM representations. In the time-aligned condition, fMRI and LM time points are stimulus-aligned. In the shuffled condition, the time points of brain responses are randomly permuted before training, destroying stimulus alignment. The fraction of shared features drops from approximately 34% (time-aligned) to approximately 3% (shuffled).

cient of the auxiliary loss is $\alpha = 0.03125$. Appendices B to F provide the preprocessing, training, and analysis details.

4.2. Are shared features derived from shared stimuli?

This experiment asks whether shared features arise because fMRI responses and LM representations carry information about the same stimulus at the same time, or simply because each representation has similar statistical structure on its own. We compare a time-aligned condition that trains on representations derived from the same stimulus with a shuffled condition that permutes the time points of brain responses before training. The shuffle preserves the marginal distributions of brain responses and LM representations, while removing the time alignment between the two. If the Brain-LM crosscoder relies only on these marginal distributions, the shuffled condition should produce nearly the same fractions of shared, brain-specific, and LM-specific features as the time-aligned condition. If the learned features depend on aligned stimulus content, the fraction of shared features should drop under the shuffle.

The fraction of shared features decreases under the shuffled condition for all three subjects and all four LMs (Figure 2). Averaged over the four LMs, the fraction of shared features decreases from 34% in the time-aligned condition to 3% in the shuffled condition, while the fraction of LM-specific features increases from 58% to 91%.

The contrast between the two conditions indicates that shared features reflect information that is jointly carried by brain responses and LM representations aligned to the same stimulus. This result supports treating the shared, brain-specific, and LM-specific feature groups as meaningful feature sets in the subsequent analyses.

4.3. To what extent are the learned features interpretable?

We then examine to what extent the learned features can be interpreted in terms of the stimulus transcripts. We use an automatic interpretation pipeline adapted from recent SAE work (Paulo et al., 2025). Each time point t is represented by a transcript window, the segment of the stimulus transcript from 8 seconds before t to t. For each feature, an explainer LLM receives the top-activating transcript windows, namely the windows at time points where the feature activation $z_{t,i}$ is largest, and transcript windows where the feature does not activate. The explainer LLM then generates a short description of the difference between the two groups. Next, a detector LLM receives this description and hold-out transcript windows, and classifies each window as one in which the feature activates or not. We measure interpretability by detection accuracy, the fraction of hold-out windows that the detector LLM classifies correctly. A one-sided binomial test with false discovery rate (FDR) correction at $q = 0.05$ identifies descriptions that predict activations in the hold-out

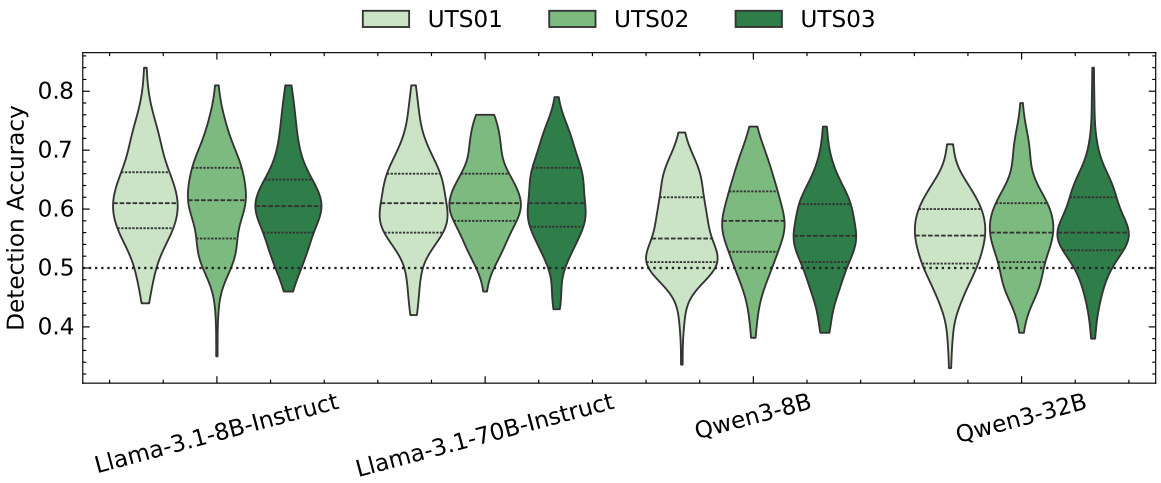


Figure 3. **Detection accuracy of automatically generated feature explanations, grouped by subject and language model.** For each of the 128 features per crosscoder, an explainer LLM receives the top-activating and non-activating transcript windows and generates a short description of the feature. A separate detector LLM uses this description to classify hold-out test transcript windows as activating or non-activating. Detection accuracy is the fraction of correct classifications. The dotted line (accuracy = 0.5) marks chance level.

Table 1. **Detection accuracy of automatically generated feature explanations for each of the 12 subject–language model pairs, all trained on stimulus-aligned (paired) observations.** Each crosscoder learns 128 features. For each feature, an explainer LLM generates a description from the top-activating transcript windows, and a detector LLM classifies held-out test windows using this description. Mean detection accuracy is the average fraction of correctly classified windows across all 128 features. FDR-significant features is the number of features whose detection accuracy is above chance (0.5) by a one-sided binomial test after Benjamini–Hochberg correction at $q = 0.05$.

Language model	Subject	Mean detection accuracy	FDR-significant features
Llama-3.1-8B-Instruct	UTS01	0.618	77
Llama-3.1-8B-Instruct	UTS02	0.614	81
Llama-3.1-8B-Instruct	UTS03	0.612	75
Llama-3.1-70B-Instruct	UTS01	0.611	66
Llama-3.1-70B-Instruct	UTS02	0.622	82
Llama-3.1-70B-Instruct	UTS03	0.617	76
Qwen3-8B	UTS01	0.563	33
Qwen3-8B	UTS02	0.578	41
Qwen3-8B	UTS03	0.555	20
Qwen3-32B	UTS01	0.551	18
Qwen3-32B	UTS02	0.565	30
Qwen3-32B	UTS03	0.570	33

windows better than random choice.

Across all 1,536 features learned from time-aligned brain and LM representations, the mean detection accuracy is 0.589, which is above the chance level of 0.5. In total, 632 explanations pass the binomial test after FDR correction, corresponding to 41.1% of all features. These significant features have a mean detection accuracy of 0.663 (Figure 3 and Table 1).

Figure 4 shows three representative features from the Brain-LM crosscoder trained on UTS03 and Llama-3.1-70B-Instruct. The brain-specific feature (Feature 101, $\Delta^{\mathrm{EV}} = +0.98$) activates on passages of personal reflection and emotional connection. The shared feature (Feature 11, $\Delta^{\mathrm{EV}} = +0.66$) activates on geographic locations and place names, a pattern that both the brain and the LM encode. The LM-specific feature (Feature 45, $\Delta^{\mathrm{EV}} = -1.00$) activates on fragmented speech with repeated fillers and self-corrections. Explanations of all other features for UTS03 and Llama-3.1-70B-Instruct are provided in Appendix K.

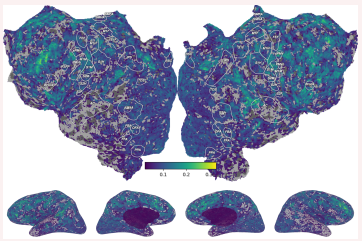
4.4. What distinguishes brain-specific and LM-specific features?

Finally, we globally investigate which transcript words are associated with brain-specific and LM-specific features. We aggregate feature predominance at each TR into a scalar score $D_t = \sum_i z_{t,i} \Delta_i^{\mathrm{EV}}$, where $z_{t,i}$ is the latent activation of feature i at TR t. We define text windows with $D_t > 0.9$ as the brain-specific text group and text windows with

BRAIN-SPECIFIC
Feature 101
$\Delta^{\mathrm{EV}} = +0.98$ Acc. = **0.68**

Explanation:
Reflection on meaningful personal relationships and emotional connections

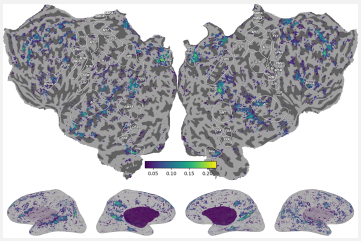
Top activating text:
"like a cloud of black robes you know the priests move off into the dark a a and and we're free to go and i just you…"
"america i'm strong and today i'm an a a sergeant in the army national guard because i can belong to an organization…"
"and i find myself missing the conversations that we never had the questions i never asked him but i"



SHARED
Feature 11
$\Delta^{\mathrm{EV}} = +0.66$ Acc. = **0.76**

Explanation:
Specific geographic locations and place names throughout the narration.

Top activating text:
"middle of the anza borrego desert which is the northeast end of san diego county and it's a it's a it's an amazing…"
"highway into ithaca and um it's on the southern shore of lake cayuga there's flatland and these hills and it has a a…"
"the foothills of the uh atlas mountains not far from my home base in marrakesh the family i stayed with there"



LM-SPECIFIC
Feature 45
$\Delta^{\mathrm{EV}} = -1.00$ Acc. = **0.79**

Explanation:
Fragmented, stammering speech with repeated filler words and self-corrections.

Top activating text:
"like the the the you like my when my mother would go she had to go outa town a couple times and he would take me on…"
"take a picture i'm like six years old and crying and stuff and uh and and i'm thinking of this shit as i'm pulling into…"
"i never made love with the art counselor again but i you know we parted amicably it was very sweet and i go back to…"

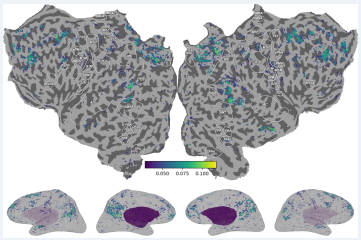


Figure 4. **Three representative features from the Brain-LM crosscoder trained on subject UTS03 with Llama-3.1-70B-Instruct.** Δ^{EV} is the Brain-LM predominance: positive values indicate that the feature explains more leave-one-out variance in fMRI responses than in LM representations, and negative values indicate the reverse. "Acc." is the detection accuracy of the automatically generated explanation. The 3 examples per card are the top-activating transcript windows. Cortical maps show voxelwise Pearson correlation between the feature activation and the fMRI response on held-out test data; only voxels with FDR-corrected $p < 0.05$ are shown.

$D_t < -0.9$ as the LM-specific text group. For each group, we count lemmatized words in the transcript and rank them by the log odds ratio with an informative Dirichlet prior (Monroe et al., 2017). This method compares the frequency of each word in the brain-specific group with its frequency in the LM-specific group by computing the log of the ratio of their odds, after adding a Dirichlet prior whose counts equal the overall corpus frequency of each word to stabilize estimates for rare words. The log odds ratio is then divided by its approximate standard deviation to yield a standardized score. Appendix J gives the exact calculation.

In Figure 5, each word cloud displays the top-ranked words for the corresponding group, with word size proportional to the standardized log odds ratio. LM-specific word clouds include colloquial expressions such as *like*, *gonna*, and *know*. These expressions appear consistently across all 12 pairs of subject and LM. Brain-specific word clouds are more heterogeneous, but they repeatedly include body, family, and action terms such as *hand*, *mother*, and *walk*.

5. Discussion

Interpretation of Brain-LM predominance. A brain-specific feature predicts fMRI responses better than LM representations. This means that the brain response encodes information along that feature dimension that the LM does not capture. The log odds word ranking (Section 4.4) shows that brain-specific time points are characterized by body, family, and action expressions (Figure 5). Body and action terms are grounded in sensorimotor experience, and family terms involve social relationships; both types of knowledge are part of the brain's response to language but are not directly available from text alone. This result is consistent with Zhou et al. (2024), who found that LMs under-explain physical and social aspects of brain responses.

Conversely, the same log odds word ranking associates LM-specific time points with colloquial markers, fillers, and contractions (Figure 5). LM training corpora consist primarily of written text, in which fillers, contractions, and other markers of spoken language are rare. When the LM processes the podcast transcripts used in this experiment, these colloquial tokens are outliers relative to the training distribution, so the LM is likely to represent them with distinctive activation patterns. The brain, by contrast, processes spoken language routinely and may not produce as distinctive a response to these tokens.

Brain-LM predominance thus captures two complementary gaps: the brain encodes sensorimotor and social dimensions of language absent from text, while the LM produces distinctive representations for spoken-language markers that

Figure 5. **Predominance word clouds for all pairs of subject and language model.** Rows are language models; columns are subjects. For each crosscoder, a TR-level predominance score $D_t = \sum_i z_{t,i} \Delta_i^{\mathrm{EV}}$ is computed by summing the feature predominance values weighted by their activations over the active features. TRs with $D_t > 0.9$ form the brain-specific group (left, red), and TRs with $D_t < -0.9$ form the LM-specific group (right, blue). Word size is proportional to the log odds ratio with an informative Dirichlet prior, reflecting how characteristic a lemmatized word is of one group relative to the other.

are rare in its written training data.

Asymmetry between brain-specific and LM-specific feature counts. Across all three subjects and four LMs, LM-specific features outnumber brain-specific features (Figure 2). A likely explanation is that the effective dimensionality of fMRI responses is lower than that of LM representations. The BOLD signal has limited temporal resolution, which constrains the number of linearly independent patterns recoverable from each time point. The LM hidden states have 4096 to 8192 dimensions and are computed deterministically from the input, so they carry many more distinguishable patterns per time point. Features whose activation time courses correlate with these LM-specific patterns but lack a detectable fMRI counterpart are classified as LM-specific. This asymmetry may therefore reflect a limitation of fMRI measurement rather than a genuine imbalance in the information each system encodes. The brain may encode linguistic distinctions that BOLD signals from fMRI lack the resolution to detect. Higher-field imaging, finer spatial sampling, or complementary recording modalities with higher temporal resolution could make additional patterns encoded by the brain accessible.

Limitations.

- The present experiments are correlational. They do not establish that the brain and LM perform the same causal computation, even when a feature is shared.
- The LM uses text transcripts, whereas the participants heard speech. Auditory and prosodic features are therefore only indirectly represented in the LM input.
- The dictionary size of 128 is small compared with those used in common LM interpretability studies (Bricken et al., 2023; Gao et al., 2025). This is due to the limited information content of fMRI BOLD signals and the limited amount of available training data.
- The automatic interpretation accuracy depends on another LLM and should be treated as an automatic filter over features, not as a substitute for human and experimental validation.

Broader impact. This work is basic research that compares internal representations of human brains and language models. Applications may include improved brain-computer interfaces and more biologically informed language model design. One risk is that decoded neural representations could reveal private cognitive or emotional states if applied without informed consent.

References

Abraham, A., Dohmatob, E., Thirion, B., Samaras, D., and Varoquaux, G. Extracting brain regions from rest fmri with total-variation constrained dictionary learning. In Mori, K., Sakuma, I., Sato, Y., Barillot, C., and Navab, N. (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pp. 607–615, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40763-5.

Aharon, M., Elad, M., and Bruckstein, A. K-svd: An algo-

rithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. doi: 10.1109/TSP.2006.881199.

Antonello, R., Vaidya, A., and Huth, A. Scaling laws for language encoding models in fMRI. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=2W4LxJbgec.

Benara, V., Singh, C., Morris, J. X., Antonello, R., Stoica, I., Huth, A., and Gao, J. Crafting interpretable embeddings for language neuroscience by asking LLMs questions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mxMvWwyBWe.

Boughorbel, S., Dalvi, F., Durrani, N., and Hawasly, M. Beyond the leaderboard: Understanding performance disparities in large language models via model diffing. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31360–31371, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1598. URL https://aclanthology.org/2025.emnlp-main.1598/.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bussmann, B., Leask, P., and Nanda, N. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

Caucheteux, C. and King, J.-R. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Caucheteux, C., Gramfort, A., and King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.

Dohmatob, E., Mensch, A., Varoquaux, G., and Thirion, B. Learning brain regions via large-scale online structured sparse dictionary learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/130f1a8e9e102707f3f91b010f151b0b-Paper.pdf.

Eavani, H., Filipovych, R., Davatzikos, C., Satterthwaite, T. D., Gur, R. E., and Gur, R. C. Sparse dictionary learning of resting state fmri networks. In *2012 Second International Workshop on Pattern Recognition in NeuroImaging*, pp. 73–76, 2012. doi: 10.1109/PRNI.2012.25.

Fred, A. L. and Jain, A. K. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005. doi: 10.1109/TPAMI.2005.113.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., and others, . The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., and Fedorenko, E. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63, 04 2024. ISSN 2641-4368. doi: 10.1162/nol_a_00137. URL https://doi.org/10.1162/nol_a_00137.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Jain, S. and Huth, A. Incorporating context into language encoding models for fmri. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and

Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

Jiralerspong, T. and Bricken, T. Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between llms. *arXiv preprint arXiv:2602.11729*, 2026.

Kassem, A. M., Shi, Z., Rostamzadeh, N., and Farnadi, G. REVIVING YOUR MNEME: Predicting the side effects of LLM unlearning and fine-tuning via sparse model diffing. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32250–32263, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1641. URL https://aclanthology.org/2025.emnlp-main.1641/.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., Tang, J., Xu, L., and Huth, A. G. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023.

Lee, K., Tak, S., and Ye, J. C. A data-driven sparse glm for fmri analysis using sparse dictionary learning with mdl criterion. *IEEE Transactions on Medical Imaging*, 30(5): 1076–1089, 2011. doi: 10.1109/TMI.2010.2097275.

Lei, Y., Ge, X., Zhang, Y., Yang, Y., and Ma, B. Do large language models think like the brain? sentence-level evidences from layer-wise embeddings and fmri. *arXiv preprint arXiv:2505.22563*, 2025. URL https://arxiv.org/abs/2505.22563.

Lepori, M. A., Kay, K., and Tuckute, G. Interpreting brain responses to language with sparse features from language models. *arXiv preprint arXiv:2606.06857*, 2026. URL https://arxiv.org/abs/2606.06857.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse Crosscoders for Cross-Layer Features and Model Diffing. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Lopez-Cardona, A., Idesis, S., Bruns, M. M., Abadal, S., and Arapakis, I. Brain–language model alignment: Insights into the platonic hypothesis and intermediate-layer advantage. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*, 2025. URL https://openreview.net/forum?id=k57mkhxyLA.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2024.01.011. URL https://www.sciencedirect.com/science/article/pii/S1364661324000275.

Merlin, G. and Toneva, M. Language models and brains align due to more than next-word prediction and word-level information. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18431–18454, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1024. URL https://aclanthology.org/2024.emnlp-main.1024/.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. Total variation regularization for fmri-based prediction of behavior. *IEEE Transactions on Medical Imaging*, 30(7):1328–1340, 2011. doi: 10.1109/TMI.2011.2113378.

Minder, J., Dumas, C., Juang, C., Chughtai, B., and Nanda, N. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=yFdNygEryH.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. Predicting human brain activity associated with the meanings of nouns. *Science*, 320 (5880):1191–1195, 2008. doi: 10.1126/science.1152876. URL https://www.science.org/doi/abs/10.1126/science.1152876.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2017. doi: 10.1093/pan/mpn018.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data.

Mach. Learn., 52(1–2):91–118, July 2003. ISSN 0885-6125. doi: 10.1023/A:1023949509487. URL https://doi.org/10.1023/A:1023949509487.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Paulo, G. S., Mallen, A. T., Juang, C., and Belrose, N. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EemtbhJOXc.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(92)90242-F. URL https://www.sciencedirect.com/science/article/pii/016727899290242F.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. 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. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105646118.

Singh, C., Antonello, R. J., Guo, S., Mischler, G., Gao, J., Mesgarani, N., and Huth, A. G. Evaluating scientific theories as predictive models in language neuroscience. *bioRxiv*, 2025. doi: 10.1101/2025.08.12.669958. URL https://www.biorxiv.org/content/early/2025/08/12/2025.08.12.669958.

Strehl, A. and Ghosh, J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3(null):583–617, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897735. URL https://doi.org/10.1162/153244303321897735.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Team, Q. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Thasarathan, H., Forsyth, J., Fel, T., Kowal, M., and Derpanis, K. G. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UoaxRN88oR.

Toneva, M. and Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In Székely, G. and Hahn, H. K. (eds.), *Information Processing in Medical Imaging*, pp. 562–573, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-22092-0.

Xu, Q., Peng, Y., Nastase, S. A., Chodorow, M., Wu, M., and Li, P. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, 9(9):1871–1886, 2025.

Zeng, A. and Gallant, J. L. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=3aNvX9TQTo.

Zhou, Y., Liu, E., Neubig, G., Tarr, M. J., and Wehbe, L. Divergences between language models and human brains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=DpP5F3UfKw.

A. Notation

Table 2. Notations used in this paper.

Symbol	Description
t	time point on the repetition-time grid
n	number of training TRs
d_{brain}	dimensionality of the brain response, equal to the number of fMRI voxels
d_{LM}	dimensionality of the LM representation
d_{latent}	dimensionality of the sparse latent space, equal to the dictionary size
k	target number of active features in BatchTopK sparsification
R	number of decoder starts used for multi-start cluster ensemble
$\mathbf{x}_t^{\text{brain}} \in \mathbb{R}^{d_{\text{brain}}}$	fMRI response at time t
$\mathbf{x}_t^{\text{LM}} \in \mathbb{R}^{d_{\text{LM}}}$	delay-averaged LM representation at time t
$\mathbf{z}_t \in \mathbb{R}^{d_{\text{latent}}}$	sparse latent feature vector at time t
$z_{t,i}$	activation of feature i at time t, the ith coordinate of $\mathbf{z}_t$
$\mathbf{Z} \in \mathbb{R}^{n \times d_{\text{latent}}}$	matrix of training feature activations
$\mathbf{X}^{\text{brain}} \in \mathbb{R}^{n \times d_{\text{brain}}}$	matrix of training brain responses
$\mathbf{X}^{\text{LM}} \in \mathbb{R}^{n \times d_{\text{LM}}}$	matrix of training LM representations
$\mathbf{W}_{\text{enc}}^{\text{brain}} \in \mathbb{R}^{d_{\text{latent}} \times d_{\text{brain}}}$	brain encoder weight matrix
$\mathbf{W}_{\text{enc}}^{\text{LM}} \in \mathbb{R}^{d_{\text{latent}} \times d_{\text{LM}}}$	LM encoder weight matrix
$\mathbf{W}_{\text{dec}}^{\text{brain}} \in \mathbb{R}^{d_{\text{brain}} \times d_{\text{latent}}}$	brain decoder weight matrix
$\mathbf{W}_{\text{dec}}^{\text{LM}} \in \mathbb{R}^{d_{\text{LM}} \times d_{\text{latent}}}$	LM decoder weight matrix
$\mathbf{W}_{\text{dec},i}^{\text{brain}} \in \mathbb{R}^{d_{\text{brain}}}$	ith column of $\mathbf{W}_{\text{dec}}^{\text{brain}}$, the brain decoder vector for feature i
$\mathbf{W}_{\text{dec},i}^{\text{LM}} \in \mathbb{R}^{d_{\text{LM}}}$	ith column of $\mathbf{W}_{\text{dec}}^{\text{LM}}$, the LM decoder vector for feature i
$\mathbf{b}_{\text{enc}} \in \mathbb{R}^{d_{\text{latent}}}$	encoder bias vector
$\text{EV}_i^{\text{brain}}$	brain explained variance of feature i
EV_i^{LM}	LM explained variance of feature i
σ_{noise}^2	brain noise variance estimate
Δ_i^{EV}	relative explained-variance difference of feature i
D_t	TR-level predominance statistic

B. Dataset and preprocessing details

The dataset is OpenNeuro ds003020, using the derivative fMRI responses and TextGrid forced alignments released with LeBel et al. (2023). We use subjects UTS01, UTS02, and UTS03. The Hugging Face model identifiers are `meta-llama/Llama-3.1-8B-Instruct`, `meta-llama/Llama-3.1-70B-Instruct`, `Qwen/Qwen3-8B`, and `Qwen/Qwen3-32B`. The fixed split contains 58 training stories, 8 validation stories, and 17 test stories. The validation split is used for layer selection and training monitoring. The crosscoder is trained on the training split, and automatic interpretation is evaluated on the test split.

For each story, LM representations are cached at TR times. Token times are computed from tokenizer character offsets and linearly interpolated word start and end times. Token hidden states are resampled to the TR grid with a Lanczos window of 3. For each retained BOLD sample, the LM input is the stack of the four preceding hidden states sampled on the TR grid, corresponding to approximately 2, 4, 6, and 8 seconds before the BOLD sample. The crosscoder averages these four delayed vectors before encoding and reconstruction.

Brain responses are centered and scaled by an RSS scale factor computed from the training TRs. For LM representations, the RSS scale factor is computed from the delay-averaged training vectors used by the crosscoder. Validation and test data use the training means and scale factors. The shuffled condition applies the same preprocessing after randomly permuting brain-response time points, thereby preserving marginal distributions while destroying stimulus alignment.

C. Crosscoder architecture details

C.1. BatchTopK activation function

BatchTopK (Bussmann et al., 2024) is a sparse activation function that enforces an approximate ℓ_0 sparsity constraint without an explicit ℓ_1 penalty on the activations. Let $\mathbf{a}_t = \mathbf{W}_{\text{enc}}^{\text{brain}}\mathbf{x}_t^{\text{brain}} + \mathbf{W}_{\text{enc}}^{\text{LM}}\mathbf{x}_t^{\text{LM}} + \mathbf{b}_{\text{enc}} \in \mathbb{R}^{d_{\text{latent}}}$ denote the pre-activation vector at time t. Each pre-activation $a_{t,i}$ is scaled by the factor $s_i = \|\mathbf{W}_{\text{dec},i}^{\text{brain}}\|_2 + \|\mathbf{W}_{\text{dec},i}^{\text{LM}}\|_2$ before ranking, so that features with larger decoder norms require proportionally larger encoder outputs to survive the selection. During training, for a batch of B examples, BatchTopK retains the Bk largest scaled pre-activations across the entire batch and sets all others to zero:

$$z_{t,i} = \text{ReLU}(a_{t,i}) \cdot \mathbf{1}[s_i \cdot a_{t,i} \geq \theta_B]\,, \tag{6}$$

where θ_B is the Bk-th largest value among the $B \cdot d_{\text{latent}}$ scaled pre-activations $\{s_i \cdot a_{t,i}\}$ in the batch. Because the budget Bk is shared across examples, each example has k active features on average, but the actual number varies per example.

At inference time, the batch-level selection is replaced by a per-example threshold θ, estimated during training as an exponential moving average of θ_B with decay γ:

$$\theta \leftarrow \gamma\,\theta + (1-\gamma)\,\theta_B. \tag{7}$$

The inference-time activation is

$$z_{t,i} = \text{ReLU}(a_{t,i}) \cdot \mathbf{1}[s_i \cdot a_{t,i} > \theta]\,. \tag{8}$$

We use $\gamma = 0.9$ in our experiments.

Unlike sparse autoencoders with ℓ_1 regularization, BatchTopK avoids feature suppression, in which the ℓ_1 penalty shrinks decoder norms toward zero even when the corresponding features contribute to reconstruction (Minder et al., 2025).

C.2. Total variation regularization

The isotropic total variation (TV) regularization penalizes large differences between adjacent voxel weights in the brain decoder to encourage spatially smooth decoder vectors and reduce overfitting (Rudin et al., 1992; Michel et al., 2011). Let V be the number of voxels in the subject mask and let E_x, E_y, and E_z be the sets of adjacent voxel pairs along the three spatial axes. For decoder start r and feature i, with brain-decoder weights $\mathbf{w}_{r,i} \in \mathbb{R}^V$, we compute

$$\text{TV}(\mathbf{w}_{r,i}) = \frac{1}{3} \sum_{a\in\{x,y,z\}} \frac{V}{|E_a|} \sum_{(u,v)\in E_a} (w_{r,i,u} - w_{r,i,v})^2. \tag{9}$$

The TV term in the loss is the mean of $\text{TV}(\mathbf{w}_{r,i})$ over decoder starts and features. Because this quantity is already normalized, it is added directly to the loss; this is equivalent to using a TV coefficient of 1.

C.3. Auxiliary loss

The auxiliary loss encourages inactive features to explain residual reconstruction error, preventing them from remaining permanently unused (Gao et al., 2025). A feature i is considered inactive (dead) if it has not fired within a sliding window of $W = 19{,}414$ TRs. Let $\mathcal{D} \subseteq \{1, \ldots, d_{\text{latent}}\}$ denote the set of dead features, and let $\mathbf{r}_t^{\text{brain}} = \mathbf{x}_t^{\text{brain}} - \hat{\mathbf{x}}_t^{\text{brain}}$ and $\mathbf{r}_t^{\text{LM}} = \mathbf{x}_t^{\text{LM}} - \hat{\mathbf{x}}_t^{\text{LM}}$ be the reconstruction residuals from the active features. For each dead feature $i \in \mathcal{D}$, the encoder pre-activation $a_{t,i}$ from the main forward pass is retained, and the auxiliary reconstruction is computed using only dead features:

$$\hat{\mathbf{x}}_t^{\text{aux},m} = \sum_{i\in\mathcal{D}} a_{t,i}\,\mathbf{W}_{\text{dec},i}^m, \tag{10}$$

where $m \in \{\text{brain}, \text{LM}\}$ and $\mathbf{W}_{\text{dec},i}^m$ is the i-th column of the decoder weight matrix for modality m. The auxiliary loss measures how well the dead features reduce the reconstruction residual:

$$\mathcal{L}_{\text{aux}} = \|\mathbf{r}_t^{\text{brain}} - \hat{\mathbf{x}}_t^{\text{aux,brain}}\|_2^2 + \|\mathbf{r}_t^{\text{LM}} - \hat{\mathbf{x}}_t^{\text{aux,LM}}\|_2^2. \tag{11}$$

The coefficient of the auxiliary loss is $\alpha = 0.03125$. Without this term, dead features receive no gradient signal from the reconstruction objective and remain unused, reducing the effective dictionary size.

C.4. Multi-start cluster ensemble

To reduce overfitting and make the learned feature vectors less sensitive to weight initialization, we train R decoder instances with different random initializations while sharing a single set of latent activations. After training, the R decoder vectors for each feature are averaged, and this mean decoder is used at inference time.

We use $R = 5$ decoder starts for each modality. For modality m and feature i, the model trains decoder vectors $\mathbf{w}^m_{1,i}, \ldots, \mathbf{w}^m_{R,i} \in \mathbb{R}^{d_m}$ that share the same latent activation $z_{t,i}$. During training, each target vector $\mathbf{x}^m_t$ is expanded to $[\mathbf{x}^m_t/\sqrt{R}, \ldots, \mathbf{x}^m_t/\sqrt{R}] \in \mathbb{R}^{Rd_m}$. This scaling preserves the target norm after expansion. The encoder receives the same expanded vector, and tied encoder weights are the transposes of the expanded decoder weights. For downstream feature analysis, we use the mean decoder vector

$$\bar{\mathbf{w}}^m_i = \frac{1}{R}\sum_{r=1}^{R} \mathbf{w}^m_{r,i}. \tag{12}$$

When computing latents after training, encoder weights derived from the mean decoders are scaled by R so that the averaged decoders operate on the original, unexpanded representation scale.

D. Closed-form explained variance with leave-one-out cross-validation

The leave-one-out cross-validated EV used in Section 3.3 has a closed-form expression that avoids retraining n separate projections. Let $\tilde{\mathbf{z}}_i \in \mathbb{R}^n$ be the column-centered activation of feature i, and let $\tilde{\mathbf{Y}} \in \mathbb{R}^{n\times m}$ be a column-centered target. The full-data one-dimensional least-squares coefficient is

$$\hat{\boldsymbol{\beta}}_i(\mathbf{Y}) = \frac{\tilde{\mathbf{Y}}^\top \tilde{\mathbf{z}}_i}{\|\tilde{\mathbf{z}}_i\|_2^2} \in \mathbb{R}^m. \tag{13}$$

For the regression of $\tilde{\mathbf{Y}}$ on $\tilde{\mathbf{z}}_i$ with intercept, the leverage of TR t is

$$h_{t,i} = \frac{1}{n} + \frac{\tilde{z}^2_{t,i}}{\|\tilde{\mathbf{z}}_i\|_2^2}, \tag{14}$$

and the LOOCV residual at TR t has the closed-form expression

$$\hat{\boldsymbol{\epsilon}}^{\mathrm{LOO}}_{t,i}(\mathbf{Y}) = \frac{\tilde{\mathbf{y}}_t - \tilde{z}_{t,i}\hat{\boldsymbol{\beta}}_i(\mathbf{Y})}{1 - h_{t,i}}, \tag{15}$$

where $\tilde{\mathbf{y}}_t \in \mathbb{R}^m$ is the tth row of $\tilde{\mathbf{Y}}$. The LOOCV EV is then

$$\mathrm{LOOEV}_i(\mathbf{Y}, \sigma^2) = \frac{\frac{1}{n}\|\tilde{\mathbf{Y}}\|_{\mathrm{F}}^2 - \frac{1}{n}\sum_{t=1}^{n}\|\hat{\boldsymbol{\epsilon}}^{\mathrm{LOO}}_{t,i}(\mathbf{Y})\|_2^2}{\frac{1}{n}\|\tilde{\mathbf{Y}}\|_{\mathrm{F}}^2 - \sigma^2} \tag{16}$$

$$= \frac{1 - \sum_{t=1}^{n}\|\hat{\boldsymbol{\epsilon}}^{\mathrm{LOO}}_{t,i}(\mathbf{Y})\|_2^2}{1 - n\sigma^2}. \tag{17}$$

The numerator is the variance of $\tilde{\mathbf{Y}}$ recovered by the hold-out predictions, and the denominator subtracts the noise variance σ^2 from the total target variance to obtain the predictable variance. The brain and LM explained variances of feature i are

$$\mathrm{EV}^{\mathrm{brain}}_i = \mathrm{LOOEV}_i(\mathbf{X}^{\mathrm{brain}}, \sigma^2_{\mathrm{noise}}), \qquad \mathrm{EV}^{\mathrm{LM}}_i = \mathrm{LOOEV}_i(\mathbf{X}^{\mathrm{LM}}, 0), \tag{18}$$

where $\mathbf{X}^{\mathrm{brain}} \in \mathbb{R}^{n\times d_{\mathrm{brain}}}$ and $\mathbf{X}^{\mathrm{LM}} \in \mathbb{R}^{n\times d_{\mathrm{LM}}}$ are the training brain responses and LM representations, and n is the number of training TRs. For brain EV, $\sigma^2 = \sigma^2_{\mathrm{noise}}$ is estimated from repeated responses to the story `wheretheressmoke`: after applying the same delay trimming and preprocessing, we compute voxelwise variance across repeats, average over time, and sum over voxels. For LM EV, $\sigma^2 = 0$ because LM representations are deterministic given the input.

E. Layer selection

Each LM layer is selected with a conventional encoding model before crosscoder training. For each candidate LM layer, delayed LM representations are used to predict fMRI responses with kernel ridge regression. The ridge coefficient is selected by five-fold cross-validation over $\{1, 10, 100, 1000, 10000, 100000\}$. The layer with the best validation mean correlation is used for crosscoder training. This procedure selects layer 12 for Llama-3.1-8B-Instruct, layer 19 for Llama-3.1-70B-Instruct, layer 20 for Qwen3-8B, and layer 47 for Qwen3-32B. Figure 6 shows the validation mean correlation across normalized layer depth for the language models used in the layer-selection analysis.

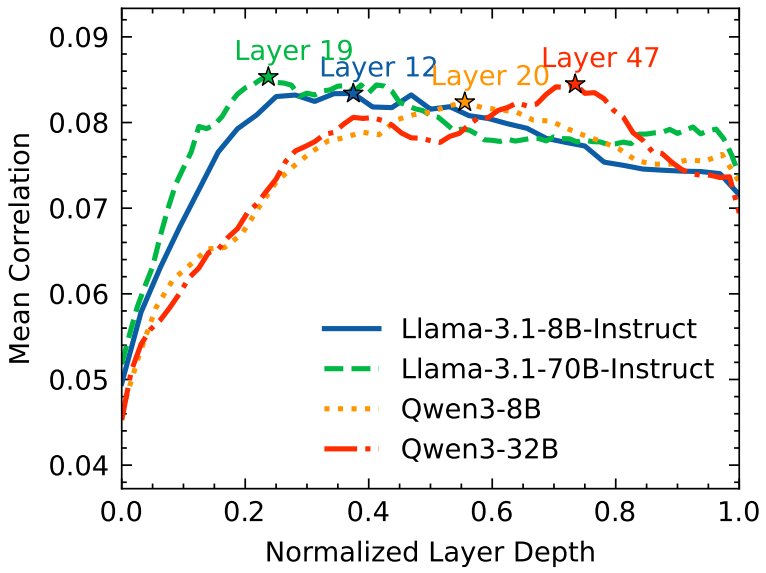


Figure 6. **Layer selection for language model representations used in crosscoder training.** For each language model layer, a kernel ridge regression encoding model predicts fMRI responses from delay-averaged LM representations. The plot shows validation-set mean Pearson correlation between predicted and observed fMRI responses as a function of normalized layer depth (layer index divided by total number of layers). Stars mark the selected layer for each model: layer 12 for Llama-3.1-8B-Instruct, layer 19 for Llama-3.1-70B-Instruct, layer 20 for Qwen3-8B, and layer 47 for Qwen3-32B.

F. Statistical testing details

Automatic interpretation uses a one-sided binomial test against the accuracy expected from random choice. For each feature, the detector classifies 50 positive and 50 negative hold-out transcript windows. The null probability is 0.5. P-values are corrected within each configuration of subject and model with Benjamini-Hochberg FDR at $q = 0.05$.

G. Detection accuracy by feature category

Figure 7 shows the automatic interpretation accuracy grouped by feature category, complementing the view grouped by subject in Figure 3.

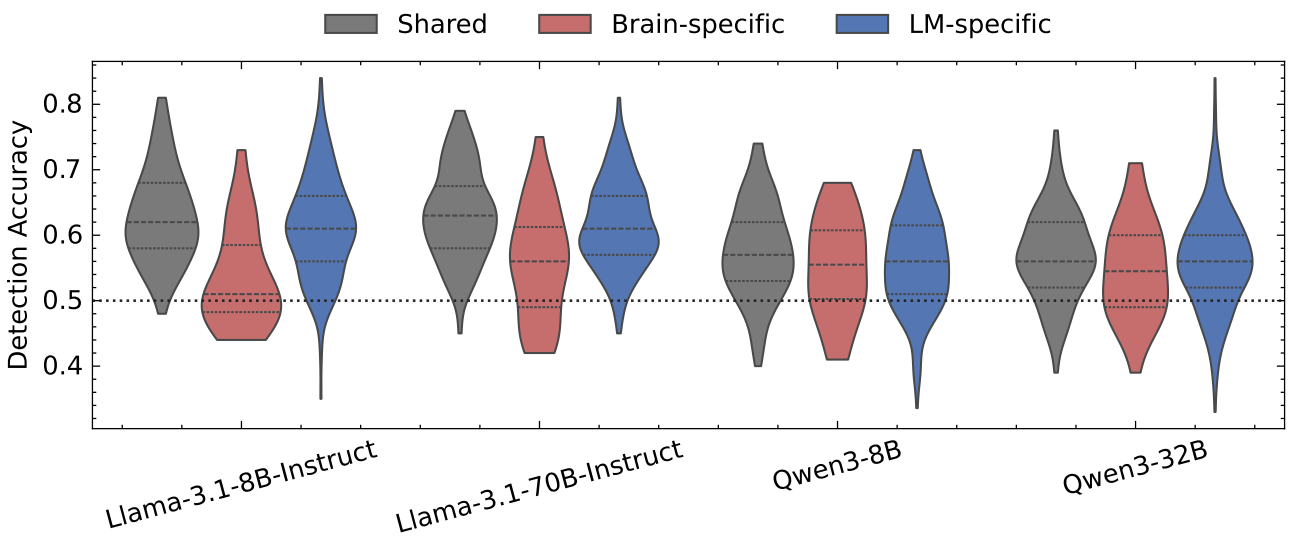


Figure 7. **Detection accuracy of automatically generated feature explanations, grouped by feature category.** Features are classified as shared ($|\Delta^{\mathrm{EV}}| \leq 0.9$), brain-specific ($\Delta^{\mathrm{EV}} > 0.9$), or LM-specific ($\Delta^{\mathrm{EV}} < -0.9$) based on the relative difference between leave-one-out explained variance for brain responses and for LM representations. The dotted line marks chance level (accuracy = 0.5). Features labeled "neither" (both EV values non-positive) are omitted because no feature from time-aligned representations falls in this category.

H. Distribution of feature predominance across subjects and language models

Figures 8 to 11 show the distribution of Brain-LM predominance Δ^{EV} for each language model across all three subjects. Features near $\Delta^{\mathrm{EV}} = 0$ explain comparable variance in fMRI responses and LM representations, whereas features near $+1$ or -1 are brain-specific or LM-specific, respectively.

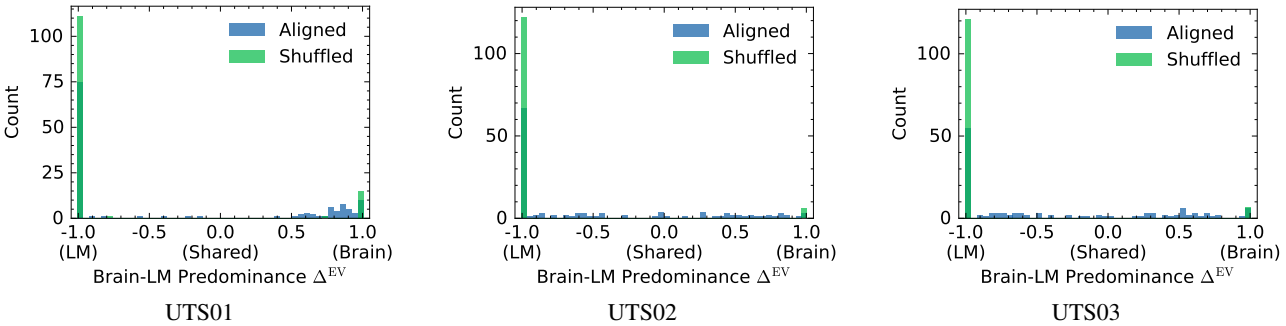


Figure 8. **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Llama-3.1-8B-Instruct layer 12.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive values indicate brain predominance and negative values indicate LM predominance. Crosscoders are trained on time-aligned representations.

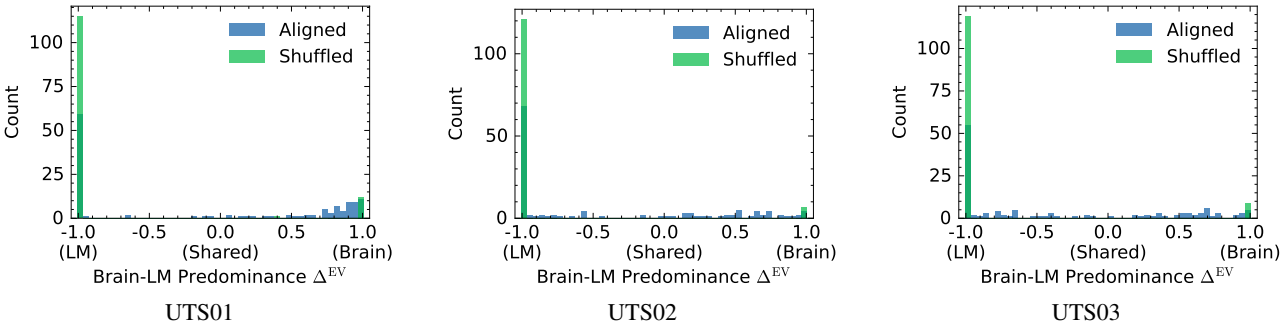


Figure 9. **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Llama-3.1-70B-Instruct layer 19.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive values indicate brain predominance and negative values indicate LM predominance. Crosscoders are trained on time-aligned representations.

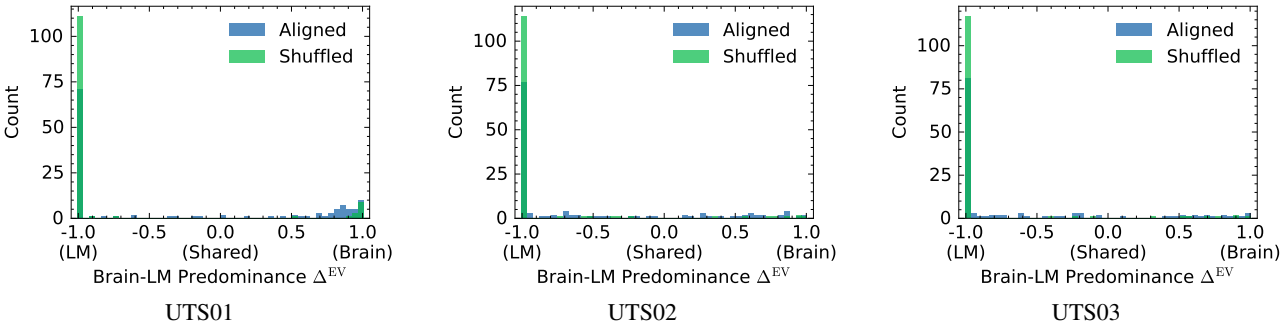


Figure 10. **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Qwen3-8B layer 20.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive values indicate brain predominance and negative values indicate LM predominance. Crosscoders are trained on time-aligned representations.

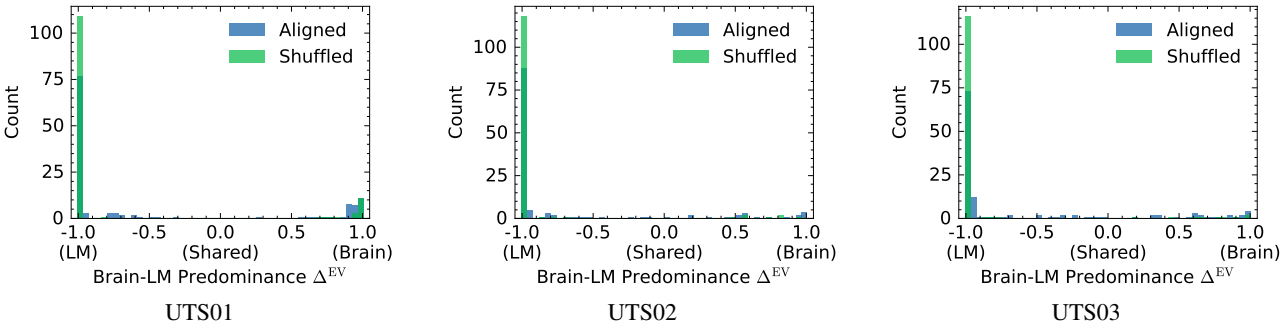


Figure 11. **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Qwen3-32B layer 47.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive values indicate brain predominance and negative values indicate LM predominance. Crosscoders are trained on time-aligned representations.

I. Automatic interpretation prompts

The explainer prompt presents positive and negative transcript windows for one feature and asks the LLM to generate a short description that distinguishes the positive windows from the negative windows. The detector prompt presents the generated description and a hold-out transcript window, then asks for a binary prediction of whether the feature should activate. Features are interpreted only when they activate at least once in both the explainer split and the detector split. For explanation, we use the training split. Positive examples are the highest-activation transcript windows, taking at most one window per story before selecting the top 10. Negative examples are 10 randomly selected zero-activation windows. For detection, we use the test split. Positive examples are 50 activating windows sampled across 50 activation quantiles, and negative examples are 50 randomly selected zero-activation windows. The detector receives shuffled batches of five windows. Both explainer and detector use `claude-haiku-4-5-20251001` with temperature 1. Detection accuracy is tested against random choice with a one-sided binomial test and FDR correction at $q = 0.05$.

I.1. Explainer prompt

System message.

```
## Purpose
- Identify patterns reflecting contrasts between positive and negative examples,
summarized in a single phrase.

## Input
- A list of text samples, each followed by an importance indicator.
- The text is part of a storytelling podcast narrated from a first-person perspective.

## Output
- Extract latent features and patterns that are common to positive examples but NOT
present in negative examples.
- Exclude attributes shared by virtually all samples (e.g., first-person perspective,
narrative style).
- Focus only on features that distinguish the positive examples from the negative ones.
- Write a concise single explanation, about 10 words in length.
- Do not make lists of possible explanations.
- Your explanation must be a noun phrase.
- End the response with `[EXPLANATION]:` followed by your explanation.

## Examples of explanation
- [EXPLANATION]: Words related to American football positions, specifically the tight end
position.
- [EXPLANATION]: The word "guys" in the phrase "you guys".
- [EXPLANATION]: "of" before words that start with a capital letter.
```

User message.

```
Positive Example 0: {{positive_text_0}}
Activation: {{latent_activation_0}}
...
Positive Example {{n}}: {{positive_text_n}}
Activation: {{latent_activation_n}}

Negative Example 0: {{negative_text_0}}
Activation: 0.00
...
Negative Example {{m}}: {{negative_text_m}}
Activation: 0.00
```

I.2. Detector prompt

System message.

```
You are an intelligent and meticulous linguistics researcher.
```

```
You will be given a certain latent of text, such as "male pronouns" or "text with
negative sentiment".

You will then be given several text examples. Your task is to determine which examples
possess the latent.

For each example in turn, return 1 if the sentence is correctly labeled or 0 if the
tokens are mislabeled. You must return your response in a valid Python list. Do not
return anything else besides a Python list.
```

User message.

```
Latent explanation: {{explanation}}

Text examples:

Example 0: {{test_example_0}}
Example 1: {{test_example_1}}
Example 2: {{test_example_2}}
Example 3: {{test_example_3}}
Example 4: {{test_example_4}}
```

J. Log odds word ranking

For each TR, we first aggregate feature-level predominance over the active feature vector:

$$D_t = \sum_i z_{t,i} \Delta_i^{\mathrm{EV}}. \tag{19}$$

Training TR windows with $D_t > 0.9$ form the brain-specific group, and windows with $D_t < -0.9$ form the LM-specific group. Representative words are ranked by applying the log odds ratio to lemmatized word counts from these two groups.

For group $g \in \{B, L\}$ and word w, let $c_{g,w}$ be the count of w and let $n_g = \sum_w c_{g,w}$. The prior count α_w is the total count of w across all training windows, with $\alpha_0 = \sum_w \alpha_w$. The brain-versus-LM log odds ratio with the informative Dirichlet prior is

$$\delta_w = \log \frac{c_{B,w} + \alpha_w}{n_B + \alpha_0 - c_{B,w} - \alpha_w} - \log \frac{c_{L,w} + \alpha_w}{n_L + \alpha_0 - c_{L,w} - \alpha_w}. \tag{20}$$

We standardize this value as

$$z_w = \frac{\delta_w}{\sqrt{(c_{B,w} + \alpha_w)^{-1} + (c_{L,w} + \alpha_w)^{-1}}}. \tag{21}$$

Brain-specific words are ranked by z_w, and LM-specific words are ranked by $-z_w$, so reported values are positive enrichment scores for the corresponding predominance group. Figure 5 shows the resulting word clouds.

K. Complete features for UTS03 and Llama-3.1-70B-Instruct

Table 3 lists all 128 features from the crosscoder trained on UTS03 and Llama-3.1-70B-Instruct.

Table 3. All 128 features from the Brain-LM crosscoder trained on subject UTS03, Llama-3.1-70B-Instruct, sorted by Brain-LM predominance Δ^{EV}. Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive values (red) indicate brain predominance and negative values (blue) indicate LM predominance. Acc. is the detection accuracy of the automatically generated explanation, measuring how well a separate LLM can use the explanation to classify held-out transcript windows.

Feature	Δ^{EV}	Acc.	Explanation
65	1.00	0.45	Narrative progression showing consequential events unfolding temporally.
46	0.99	0.45	Direct address to audience or reflective generalization about human experience.
67	0.99	0.55	Internal emotional reflection or introspective thought processes.
77	0.98	0.49	Fragmented speech with repeated conjunctions suggesting present-moment uncertainty.
101	0.98	0.68	Reflection on meaningful personal relationships and emotional connections
3	0.96	0.43	Fragmented speech with stutters, hesitations, and incomplete clauses.
68	0.95	0.58	Incomplete declarative phrases creating narrative suspense or emphasis.
30	0.94	0.56	Repeated casual dialogue markers with "like" and "and" filler words
123	0.89	0.52	Absurd hypothetical scenarios with specific, unexpected conditions.
29	0.89	0.65	Complete, extended narrative sequences with vivid sensory details and dialogue.
121	0.78	0.66	Specific named objects or food items with concrete sensory details.
4	0.78	0.64	Direct quoted dialogue exchanges with clear speaker attribution.
26	0.77	0.57	Presence of multiple people gathered together in group settings.
60	0.72	0.60	Intense physical collisions and high-impact bodily movements.
25	0.71	0.52	Specific, concrete details about narrator's personal experiences and locations.
41	0.69	0.57	Internal metacognitive commentary about one's own thoughts.
127	0.69	0.68	Vivid physical and personality descriptions of specific characters.
44	0.68	0.57	Concrete sequential descriptions of physical movement and actions.
80	0.67	0.65	Direct physical contact or gesture between two people in interaction.
122	0.67	0.60	References to educational institutions, degrees, or professional career accomplishments.
11	0.66	0.76	Specific geographic locations and place names throughout the narration.
18	0.65	0.58	Contradictory or paradoxical situations presented matter-of-factly.
100	0.65	0.54	References to serious medical diagnoses, diseases, or health conditions.
38	0.59	0.65	Specific descriptive details about people's identities, ages, or professions.
118	0.59	0.59	Introduction of named or described people with their defining roles or professions.
120	0.57	0.61	Descriptions of organized systems or structured group activities.
14	0.57	0.65	Narrator's self-reflection and judgment about their own feelings or beliefs.
51	0.54	0.57	Narrator's self-aware meta-commentary interrupting the story itself.
0	0.54	0.70	Sensory perception in moments of solitude or introspection.
43	0.52	0.62	Introspective reflection on emotions, character, and internal experience
115	0.51	0.67	Narrator acquiring or retrieving specific tangible objects.
48	0.49	0.72	Narrator performing active physical movements or deliberate actions.
88	0.48	0.64	Direct dialogue exchanges showing interpersonal negotiation or interaction.
47	0.40	0.61	Explicit time passage markers indicating narrative progression spanning weeks to years.
69	0.37	0.62	Narrative arcs of deliberately seeking out and meeting specific people.
36	0.35	0.71	Introduction of specific named individuals with personal details.
119	0.34	0.73	Internal self-doubt and psychological conflict exploration.
109	0.30	0.53	Direct dialogue or reported speech with conversational markers.
31	0.26	0.65	Transformative overcoming or reframing of difficult circumstances
21	0.26	0.67	Sequential descriptions of taking, moving, and placing objects.
112	0.22	0.62	Sensory descriptions of physical distress, pain, or discomfort.
89	0.19	0.69	Physical descriptions of emotional distress and bodily reactions.
61	0.19	0.61	Expressions of not knowing or discovering information gaps.
103	0.02	0.72	References to narrator's age or specific life stage periods.
117	-0.10	0.70	Metaphorical descriptions of internal bodily sensations and emotional intensity.
93	-0.14	0.72	Receiving unexpected communications or messages with significant news.
79	-0.16	0.59	Medical procedures, injuries, and surgical interventions with physical descriptions.
62	-0.24	0.66	Descriptions of the narrator's physical sensations and bodily states.
39	-0.38	0.64	Dialogue or statements about upcoming events and future plans.
59	-0.41	0.67	Direct dialogue or reported speech embedded in narrative.
114	-0.42	0.51	Direct quoted speech and dialogue from other characters.
86	-0.42	0.54	Emphatic first-person declarations of refusal or steadfast commitment.
12	-0.45	0.73	Vivid sensory and visual imagery describing scenes and moments.

Table 3. All 128 features for UTS03, Llama-3.1-70B-Instruct (continued).

Feature	Δ^{EV}	Acc.	Explanation
110	-0.49	0.60	Characterizations of people's personalities and social behaviors.
92	-0.52	0.74	Narrator actively initiating pursuit of new opportunities or goals.
40	-0.64	0.59	Completed action sequences describing arrival or resolution moments.
72	-0.65	0.45	Direct dialogue and personal interaction with specific named people.
42	-0.65	0.59	Direct speech and dialogue expressing reactions to others' statements.
87	-0.66	0.46	Specific vivid details and emotionally significant moments from personal experiences.
22	-0.66	0.72	Rhetorical questions and conditional phrases revealing internal anxiety and doubt.
64	-0.73	0.64	Repetitive phrases reflecting anxious internal monologue.
19	-0.73	0.58	Vivid descriptions of birth, infancy, or life-threatening medical situations.
52	-0.75	0.55	Hypothetical scenarios and conditional logical premises.
76	-0.77	0.70	Repetitive actions connected by coordinating conjunctions emphasizing obsessive effort.
84	-0.79	0.54	Narrative about deliberating on or transitioning between professional careers.
106	-0.79	0.65	Personal anxieties and uncertainties about romantic relationships and dating.
10	-0.80	0.61	Specific concrete objects or tangible physical spaces and environments.
27	-0.85	0.67	Depictions of individual's characteristic or habitual behavioral patterns.
8	-0.86	0.55	Vivid, specific details about objects, people, and concrete circumstances.
107	-0.87	0.63	Anaphoric repetition of clause beginnings across multiple parallel phrases.
85	-0.94	0.66	Revelations about relationships, emotions, or personal identity between people.
111	-0.97	0.67	Descriptions of vehicles, driving, and transportation journeys.
99	-0.97	0.58	Descriptive passages about external categories, groups, and environments rather than personal emotional reactions.
24	-1.00	0.68	Repetition of future intention statements using "gonna" or "will"
1	-1.00	0.52	Hypothetical or aspirational statements about future possibilities and desires.
2	-1.00	0.61	Direct communication or interaction between people in the narrative.
5	-1.00	0.57	Surreal or physically impossible spatial and object scenarios.
6	-1.00	0.73	Introspective reflection on interpersonal relationships and personal emotions.
7	-1.00	0.70	Personal reflection on unmet expectations or life trajectory decisions.
9	-1.00	0.52	Narrative pivot points marked by conjunction-led incomplete thoughts.
13	-1.00	0.71	Personal experiences involving deliberate lifestyle choices or geographic relocation.
15	-1.00	0.56	Dialogue or quoted speech from the narrator or other characters.
16	-1.00	0.73	References to historical events, deaths, or institutional conflicts beyond personal scope.
17	-1.00	0.65	Speaker's active choices and deliberate experiential adventures rather than family circumstances.
20	-1.00	0.58	Sudden realization or discovery moments with vivid sensory details.
23	-1.00	0.58	Narrative sequences with embedded decisions and their practical consequences.
28	-1.00	0.74	Vivid descriptions of physical actions and sensory experiences unfolding in sequence.
32	-1.00	0.67	Metacognitive reflection on personal thoughts, intentions, and decision-making processes.
33	-1.00	0.66	Narrative moments containing unexpected plot twists or complications.
34	-1.00	0.64	References to historically significant events or public figures.
35	-1.00	0.58	Direct questions expressing uncertainty about grave consequences or outcomes.
37	-1.00	0.61	Specific concrete activities and tangible action-based experiences described.
45	-1.00	0.79	Fragmented, stammering speech with repeated filler words and self-corrections.
49	-1.00	0.54	Acknowledgment of personal wrongdoing or internal moral conflict within oneself.
50	-1.00	0.67	Descriptions of pursuing formal education or career advancement opportunities.
53	-1.00	0.52	Direct dialogue or reciprocal interaction between narrator and another person.
54	-1.00	0.57	Descriptions of physical sensations and bodily states.
55	-1.00	0.62	References to formal ceremonies, weddings, anniversaries, or commemorative occasions.
56	-1.00	0.66	Explicit discussion of planning, preparing, or anticipating future needs.
57	-1.00	0.64	Descriptions of escalating interpersonal conflict or social tension.
58	-1.00	0.53	Narrator expressing confusion and active struggle to understand what's happening.
63	-1.00	0.58	Instructions or procedural steps describing sequential actions to perform.
66	-1.00	0.75	Repetitive stuttering and hesitations expressing emotional distress or vulnerability.
70	-1.00	0.55	Immediate physical danger or survival situations with active threats.
71	-1.00	0.60	Stories about people's difficult or contradictory behaviors and character traits.
73	-1.00	0.59	Narratives depicting romantic interest or courtship between people.
74	-1.00	0.67	Descriptions of acute physical danger or life-threatening emergency situations.
75	-1.00	0.67	Explicit descriptions of physical bodily states and sensations.
78	-1.00	0.58	Intimate moments of physical affection and chaotic group joy.
81	-1.00	0.64	Narrator actively searching, exploring, or investigating something specific.
82	-1.00	0.71	Emotional expressions of love and personal connection with individuals.
83	-1.00	0.51	Tension between personal desires and external circumstances or constraints.
90	-1.00	0.74	Repeated first-person plural pronouns describing shared group activities.

Table 3. All 128 features for UTS03, Llama-3.1-70B-Instruct (continued).

Feature	Δ^{EV}	Acc.	Explanation
91	-1.00	0.60	References to writing, letters, sending messages, or intentional communication exchanges.
94	-1.00	0.62	Discovery or encounter with a specific physical object or artifact.
95	-1.00	0.59	Speculative, hypothetical, or counterfactual scenarios rather than concrete past events.
96	-1.00	0.63	Narrator deliberating about future decisions and life choices.
97	-1.00	0.53	Narrative sequences with connected cause-and-effect storytelling chains.
98	-1.00	0.61	Concrete procedures or specific objects performing sequential actions.
102	-1.00	0.68	Self-aware commentary about the narrator's act of storytelling.
104	-1.00	0.60	Specific, tangible physical objects with concrete descriptive details.
105	-1.00	0.70	Specific names of real people, places, or brands mentioned.
108	-1.00	0.45	References to specific personal relationships and emotional bonds.
113	-1.00	0.56	Describing specific step-by-step procedures or technical instructions.
116	-1.00	0.59	Direct sensory observation or visual wonder at concrete phenomena.
124	-1.00	0.56	References to family members and household situations.
125	-1.00	0.64	Vivid sensory and physical details describing concrete actions.
126	-1.00	0.57	Spontaneous emotional interjections expressing immediate reactions to moments.