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

Anonymous Author(s)
Affiliation
Address
email

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, which features are shared between brain and LM representations and which are selectively used in brains and LMs have remained underspecified. We propose Brain-LM crosscoders, which decompose brain responses and LM representations into shared sparse features and label each feature as being shared, brain-specific, or LM-specific based on its predictive contribution to each representation. Experiments on naturalistic language listening fMRI data 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, which will contribute to scientific discovery in both neuroscience and artificial neural network research. Our code is available at `https://anonymous.4open.science/r/brain-lm-crosscoder/`

1 Introduction

Language neuroscience and language model (LM) interpretability studies share an interest in how information of natural-language stimuli is represented inside the brain and models, respectively. Encoding models that predict brain responses from stimulus features have served as a standard tool in neuroscience to understand the nature of internal representations (Mitchell et al., 2008; Huth et al., 2016). Subsequent studies revealed that contextual LM representations predict brain responses better than static word embeddings (Jain and Huth, 2018), and that LM representations predict brain responses across layers, model scales, and training objectives (Schrimpf et al., 2021; Caucheteux and King, 2022; Toneva and Wehbe, 2019; Hosseini et al., 2024). These results indicate that LM representations and brain responses use the shared information extracted from the 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 those that are clearly represented in one but not the other. Studies analyzing prediction errors can identify aspects of brain responses that LMs fail to capture (Zhou et al., 2024), but this approach is inherently asymmetric because it cannot extract features that are clearly represented in LMs but absent in the brain. A symmetric and feature-level comparison 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 with a separate decoder (Figure 1). Each latent coordinate represents a candidate linguistic feature. For each feature, we measure the explained variance (EV) in both fMRI responses and LM representations. Feature

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

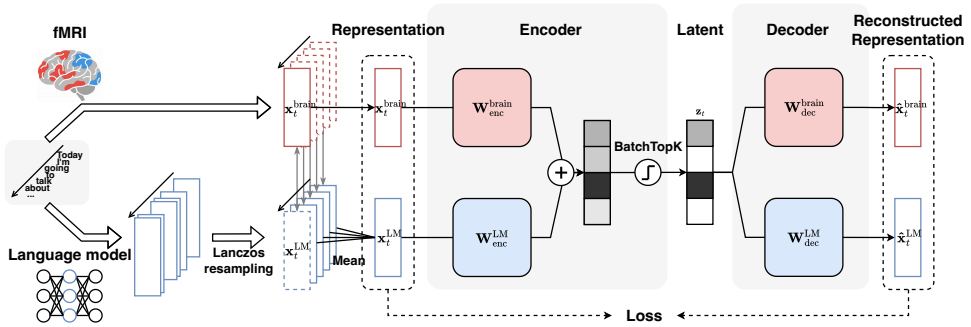

Figure 1: **Overview of Brain-LM crosscoder.** Time-aligned fMRI responses and language model (LM) representations for the same linguistic stimuli are encoded into a shared sparse feature vector. Token-level LM representations are resampled to the fMRI repetition-time (TR) grid with Lanczos resampling 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 ($\Delta^{\mathrm{EV}} < -0.9$).

predominance is defined as the relative difference between these two EV values, indicating whether a feature predicts fMRI responses or LM representations better. A brain-specific feature indicates more variance in fMRI responses than in LM representations, while an LM-specific feature shows the reverse pattern. This feature-level comparison between the brain and LMs stands in contrast to the layer-level comparisons that are typical in recent encoding models.

We apply this method to naturalistic fMRI data from LeBel et al. (2023) with four LMs and three subjects (Section 4) and report three findings. First, we found that most shared features disappear when stimulus alignment between brain and LM is randomly shuffled, suggesting 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 are interpretable in comparison with a random explanation. Third, we found that body, family, and action-related language 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 and Huth, 2018). Subsequent work studied layerwise alignment, scaling, and the relation between next-word prediction and brain predictivity (Toneva and Wehbe, 2019; Antonello et al., 2023; Caucheteux et al., 2023; Merlin and Toneva, 2024; Lei et al., 2025; Lopez-Cardona et al., 2025; Singh et al., 2025). Building on these approaches, our proposed method identifies which sparse latent features jointly reconstruct brain and LM representations and characterizes 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 and physical aspects of brain responses. Xu et al. (2025) compared conceptual word ratings between humans and LLMs and found that LLM–human similarity 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 acquire formal linguistic competence but not functional competence. However, these approaches characterize divergences along predefined categories or in the LM-to-brain direction, and do not address which features LMs capture that the brain does not and vice versa. By treating the

two representations symmetrically, our approach, the Brain-LM crosscoder, aims to identify both brain-specific and LM-specific features.

Dictionary learning. Dictionary learning represents observed value as sparse linear combinations of basis vectors (Olshausen and 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 been proposed crosscoders, SAE variants designed to identify 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). These methods have so far been applied within either brain data or LM activations alone, but we apply them across both domains to investigate correspondences between the two representations.

3 Method

Our method encodes brain responses and LM representations into a shared sparse latent space. This enables the extraction of both shared features and representation-specific features within a single unified framework (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 and an LM representation derived from the same stimulus transcript. In this study, we use fMRI BOLD signals as the brain responses. We denote these time-aligned observations by $\mathbf{x}_t^{\mathrm{brain}} \in \mathbb{R}^{d_{\mathrm{brain}}}$ and $\mathbf{x}_t^{\mathrm{LM}} \in \mathbb{R}^{d_{\mathrm{LM}}}$. Here d_{brain} is the number of voxels and d_{LM} is the LM representation dimension.

Representation of brain and language model. In this paper, the brain representation is the blood-oxygen-level-dependent (BOLD) response measured while a participant listens to natural-language stories. The response is sampled at every repetition time (TR) and treated as a vector over voxels. The LM representation is the hidden-state activation at a specific layer when the same transcript 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 (Huth et al., 2016). 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 TR-sampled LM representations, 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 $\mathbf{x}_t^{\mathrm{LM}}$.

Normalization. Brain and LM representations are centered and scaled with training-set statistics 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). *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 $\mathbf{x}_t^{\text{lm}_{\text{delay}}}$ at time t into a shared d_{latent}-dimensional latent space. Applying a sparse activation function to this representation yields 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 delayed LM representation from the latent variable, as follows.

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

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

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

Here $\mathbf{W}_{\mathrm{enc}}^{\mathrm{brain}}$ and $\mathbf{W}_{\mathrm{enc}}^{\mathrm{LM}}$ are modality-specific encoder weights, $\mathbf{W}_{\mathrm{dec}}^{\mathrm{brain}}$ and $\mathbf{W}_{\mathrm{dec}}^{\mathrm{LM}}$ are modality-specific decoder weights, and $\mathbf{b}_{\mathrm{enc}}$ is the encoder bias. The parameter k is the target number of active features. The model uses tied weights ($\mathbf{W}_{\mathrm{enc}} = \mathbf{W}_{\mathrm{dec}}^{\top}$). 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

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

Here, $\mathrm{TV}(\cdot)$ is total variation (Rudin et al., 1992; Michel et al., 2011) to penalize large differences between adjacent voxel weights in the brain decoder and reduce overfitting (Section C.2). $\mathcal{L}_{\mathrm{aux}}$ 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, and use the average of their weights at inference time (Section C.4). This can be regarded as a form of cluster ensemble and consensus clustering (Strehl and Ghosh, 2003; Monti et al., 2003; Fred and Jain, 2005).

3.3 Brain-LM predominance

Prior studies using crosscoders identified model-specific features by comparing the norms of per-feature decoder weight vectors (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 of the target representations on the scalar activation $z_{t,i}$ with an intercept 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 estimate 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

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

$\Delta_i^{\mathrm{EV}} > 0$ indicates that feature i predicts brain responses better than LM representations, and $\Delta_i^{\mathrm{EV}} < 0$ 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, the predominance ratio 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 $\Delta_i^{\mathrm{EV}} < -0.9$.

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 transcript words that characterize brain-specific and LM-specific time points 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), specifically subjects UTS01, UTS02, and UTS03. BOLD responses are sampled at TR=2 seconds. The stimuli are English podcast stories of roughly 10–15 minutes each. 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 $d_{\text{latent}} = 128$ (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 with learning rate 10^{-4} (Kingma and Ba, 2015). The auxiliary-loss coefficient 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 brain-response time points before training. The shuffle preserves the marginal distribution of brain responses and LM representations, while removing their time alignment between the two. If 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 falls from approximately 34% in the time-aligned condition to approximately 3% in the shuffled condition, while the fraction of LM-specific features grows from approximately 58% to approximately 91%.

The contrast between the two conditions indicates that shared features reflect information that is jointly carried by stimulus-aligned brain responses and LM representations. This validates the use of shared, brain-specific, and LM-specific feature groups in the subsequent analyses, where each group is treated as a meaningful feature set that depends on cross-modal time alignment.

4.3 To what extent are the learned features interpretable?

We then examine how often learned features can be described in 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

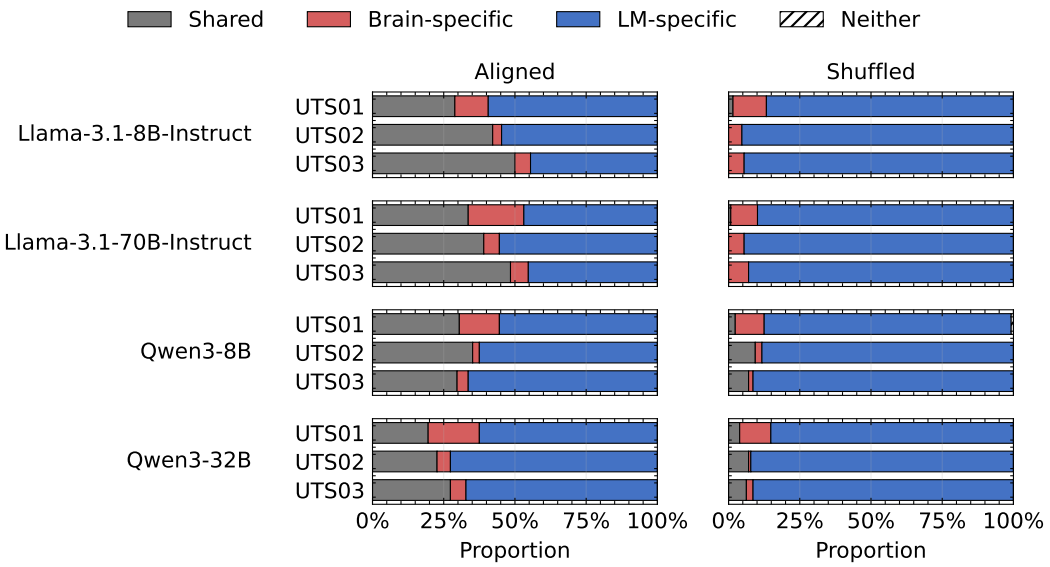

Figure 2: **Feature-category ratios for Brain-LM crosscoders trained on three subjects and four language models.** Each bar shows the fraction of 128 features classified as shared ($|\Delta^{\mathrm{EV}}| \leq 0.9$), brain-specific ($\Delta^{\mathrm{EV}} > 0.9$), LM-specific ($\Delta^{\mathrm{EV}} < -0.9$), 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, brain-response time points are randomly permuted before training, destroying stimulus alignment. The fraction of shared features drops from approximately 34% (time-aligned) to approximately 3% (shuffled).

before t to t. For each feature, an explainer LLM receives transcript windows where the feature activates and transcript windows where it does not, and generates a short description of the difference between the two groups. The top-activating transcript window means the transcript window at time points where the feature activation $z_{t,i}$ is largest. A detector LLM then 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 on these test windows. Detection accuracy is calculated as the fraction of transcript windows that the detector LLM correctly classifies as activating or not. A one-sided binomial test with FDR correction at $q = 0.05$ identifies descriptions that predict activations in the test windows better than random choice.

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

Figure 4 shows three representative features from the Brain-LM crosscoder using UTS03 and Llama-3.1-70B-Instruct. The brain-specific feature (Feature 101, $\Delta_i^{\mathrm{EV}} = +0.98$) responds to passages of personal reflection and emotional connection. The shared feature (Feature 11, $\Delta_i^{\mathrm{EV}} = +0.66$) activates on geographic locations and place names, a pattern that both the brain and the LM encode. The LM-specific feature (Feature 45, $\Delta_i^{\mathrm{EV}} = -1.00$) fires 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 time points?

Finally, we investigate which transcript words are associated with brain-specific and LM-specific time points. We aggregate feature predominance at each TR into a scalar score $D_t = \sum_i z_{t,i} \Delta_i^{\mathrm{EV}}$, which sums the predominance values of active features after scaling them by their activations. TRs with $D_t > 0.9$ form the brain-specific group, and TRs with $D_t < -0.9$ form the LM-specific group. For

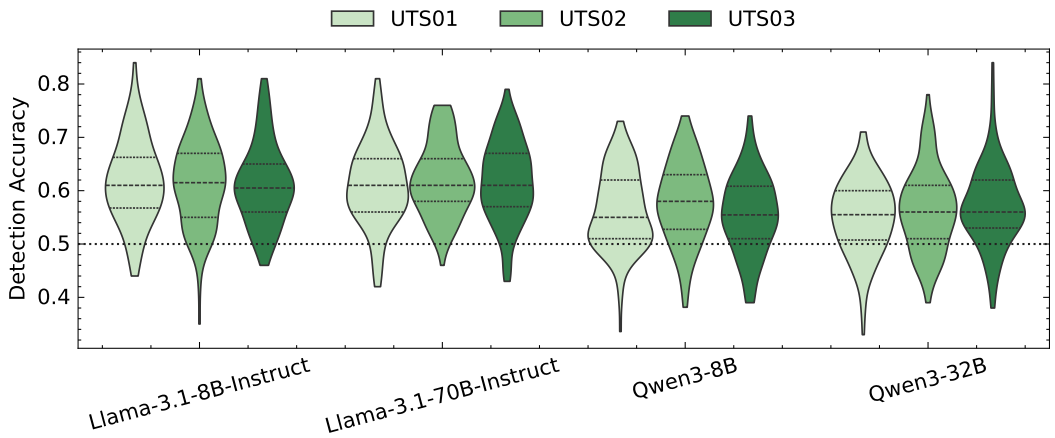

Figure 3: **Detection accuracy of automatically generated feature explanations, grouped by subject and language model.** For each of the 128 features per crosscoder, an explainer LLM receives the top-activating and non-activating transcript windows and generates a short description of the feature. A 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

each group, we lemmatize the corresponding word counts and rank words by the log-odds ratio with an informative Dirichlet prior (Monroe et al., 2017). This statistic measures whether a word occurs more frequently in one predominance group than the other, controlling for its overall frequency in the dataset. Appendix J gives the exact calculation.

Figure 5 shows that LM-specific word clouds include colloquial expressions such as *like*, *gonna*, and *know*. They are consistent across all 12 configurations. Brain-specific word clouds are more heterogeneous, but they repeatedly include body, family, and action terms such as *hand*, *mother*, and *walk*.

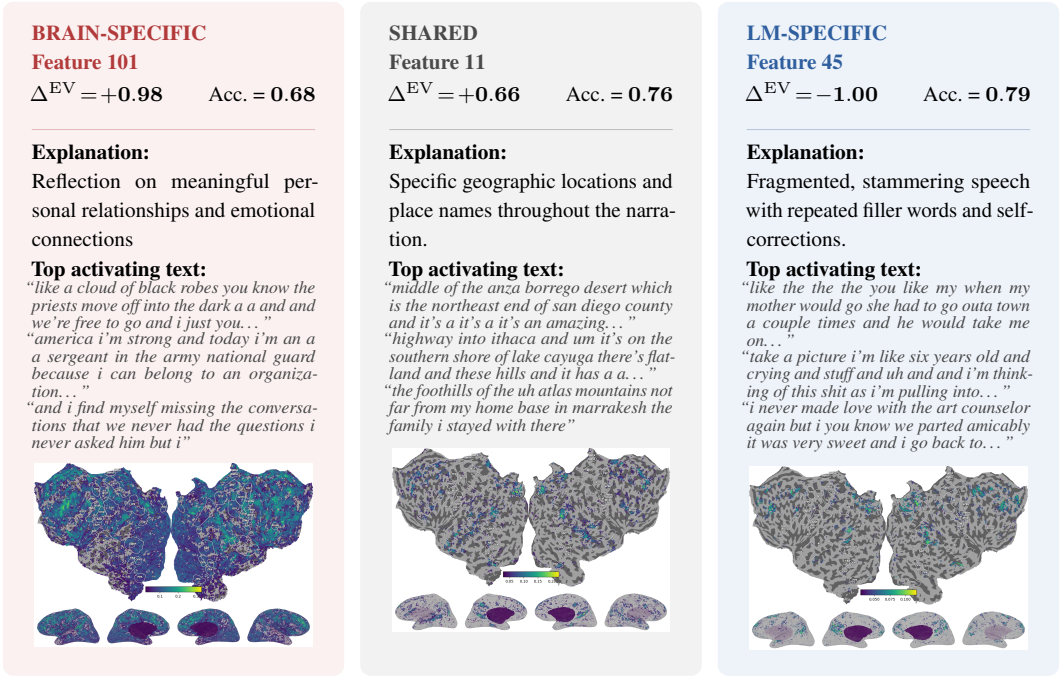

Figure 4: **Three representative features from the Brain-LM crosscoder trained on subject UTS03 with Llama-3.1-70B-Instruct.** Δ^{EV} is the Brain-LM predominance: positive values indicate that the feature explains more leave-one-out variance in fMRI responses than in LM representations, and negative 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.

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). These semantic categories are grounded in sensory and motor experience that the brain encodes during language comprehension but that text-only LMs lack direct access to. This result is consistent with Zhou et al. (2024), who found that LMs under-explain social, emotional, and physical aspects of brain responses. Brain-specific features may therefore capture embodied and social knowledge not available in text corpora.

LM-specific features predict LM representations better than fMRI responses. 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 spoken-language markers are rare. When the LM processes the podcast transcripts used in this experiment, these colloquial tokens are distributional 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 reflects the difference between the embodied semantics that the brain encodes and the representations that the LM derives from text input alone.

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

Figure 5: **Predominance word clouds for all subject–language model pairs.** 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 activation-weighted feature predominance values over 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.

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 is a property of the measurement rather than of crosscoders: the brain may encode linguistic distinctions that BOLD fMRI lacks the resolution to detect. Higher-field imaging, finer spatial sampling, or complementary recording modalities with higher temporal resolution could make additional brain-encoded patterns 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 and amount of training data available of fMRI BOLD signals.
- The automatic interpretation accuracy depends on another LLM and should be treated as an automatic first-pass 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. Possible positive applications include improved brain-computer interfaces and more biologically informed design of language models. A potential 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. (2013). Extracting brain regions from rest fmri with total-variation constrained dictionary learning. In Mori, K., Sakuma, I., Sato, Y., Barillot, C., and Navab, N., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 607–615, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aharon, M., Elad, M., and Bruckstein, A. (2006). K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322.

Antonello, R., Vaidya, A., and Huth, A. (2023). Scaling laws for language encoding models in fMRI. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Boughorbel, S., Dalvi, F., Durrani, N., and Hawasly, M. (2025). Beyond the leaderboard: Understanding performance disparities in large language models via model diffing. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31360–31371, Suzhou, China. Association for Computational Linguistics.

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. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bussmann, B., Leask, P., and Nanda, N. (2024). Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

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

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

Dohmatob, E., Mensch, A., Varoquaux, G., and Thirion, B. (2016). Learning brain regions via large-scale online structured sparse dictionary learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Eavani, H., Filipovych, R., Davatzikos, C., Satterthwaite, T. D., Gur, R. E., and Gur, R. C. (2012). Sparse dictionary learning of resting state fmri networks. In *2012 Second International Workshop on Pattern Recognition in NeuroImaging*, pages 73–76.

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

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. (2025). Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

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, . (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., and Fedorenko, E. (2024). 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.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. (2024). Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

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

Jain, S. and Huth, A. (2018). Incorporating context into language encoding models for fmri. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kassem, A. M., Shi, Z., Rostamzadeh, N., and Farnadi, G. (2025). 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., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32250–32263, Suzhou, China. Association for Computational Linguistics.

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

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

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

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

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. (2024). Sparse Crosscoders for Cross-Layer Features and Model Diffing. *Transformer Circuits Thread*.

Lopez-Cardona, A., Idesis, S., Bruns, M. M., Abadal, S., and Arapakis, I. (2025). 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*.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Merlin, G. and Toneva, M. (2024). 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., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18431–18454, Miami, Florida, USA. Association for Computational Linguistics.

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

Minder, J., Dumas, C., Juang, C., Chughtai, B., and Nanda, N. (2025). Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

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

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, 52(1–2):91–118.

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

Paulo, G. S., Mallen, A. T., Juang, C., and Belrose, N. (2025). Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Singh, C., Antonello, R. J., Guo, S., Mischler, G., Gao, J., Mesgarani, N., and Huth, A. G. (2025). Evaluating scientific theories as predictive models in language neuroscience. *bioRxiv*.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3(null):583–617.

Team, Q. (2025). Qwen3 technical report.

Thasarathan, H., Forsyth, J., Fel, T., Kowal, M., and Derpanis, K. G. (2025). Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*.

Toneva, M. and Wehbe, L. (2019). 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. Curran Associates Inc.

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

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

Zhou, Y., Liu, E., Neubig, G., Tarr, M. J., and Wehbe, L. (2024). Divergences between language models and human brains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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^2_{noise}	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 TR-sampled hidden states, 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.

Compared with ℓ_1-penalized sparse autoencoders, 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 auxiliary-loss coefficient 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 $\mathbf{w}_{1,i}^m, \ldots, \mathbf{w}_{R,i}^m \in \mathbb{R}^{d_m}$ that share the same latent activation $z_{t,i}$. During training, each target vector $\mathbf{x}_t^m$ is expanded to $[\mathbf{x}_t^m/\sqrt{R}, \ldots, \mathbf{x}_t^m/\sqrt{R}] \in \mathbb{R}^{Rd_m}$. 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

$$\bar{\mathbf{w}}_i^m = \frac{1}{R} \sum_{r=1}^{R} \mathbf{w}_{r,i}^m. \tag{12}$$

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

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

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

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

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

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

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

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

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

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, $\sigma^2 = \sigma_{\mathrm{noise}}^2$ 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.

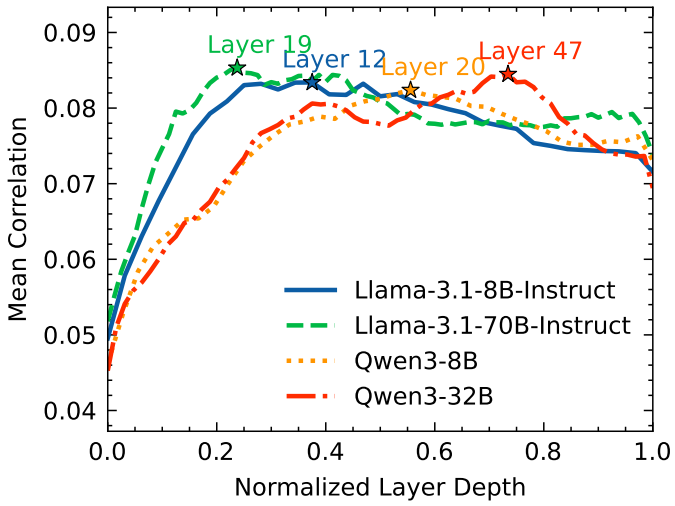

Figure 6: **Layer selection for language model representations used in crosscoder training.** For each language model layer, a kernel ridge regression encoding model predicts fMRI responses from delay-averaged LM representations. The plot shows validation-set mean Pearson correlation between predicted 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 subject-model configuration 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 subject-level view in Figure 3.

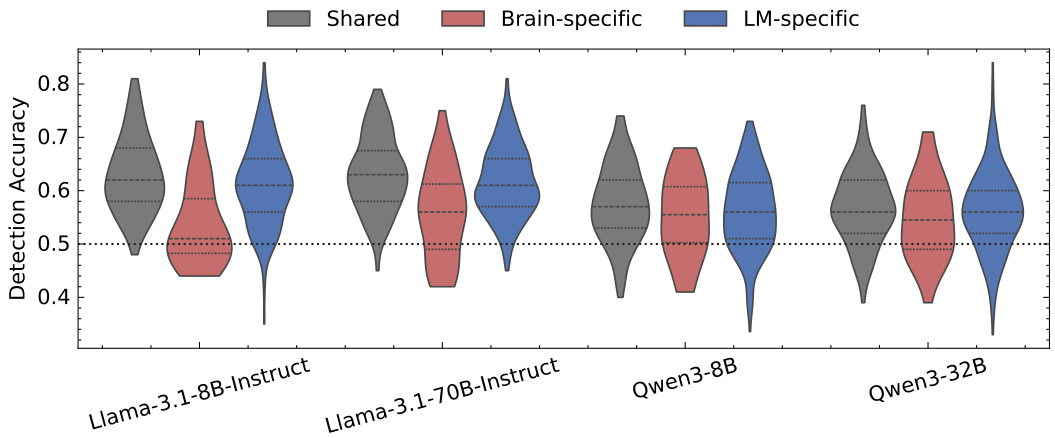

Figure 7: **Detection accuracy of automatically generated feature explanations, grouped by feature category.** Features are classified as shared ($|\Delta^{\mathrm{EV}}| \leq 0.9$), brain-specific ($\Delta^{\mathrm{EV}} > 0.9$), or LM-specific ($\Delta^{\mathrm{EV}} < -0.9$) based on the relative difference 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.

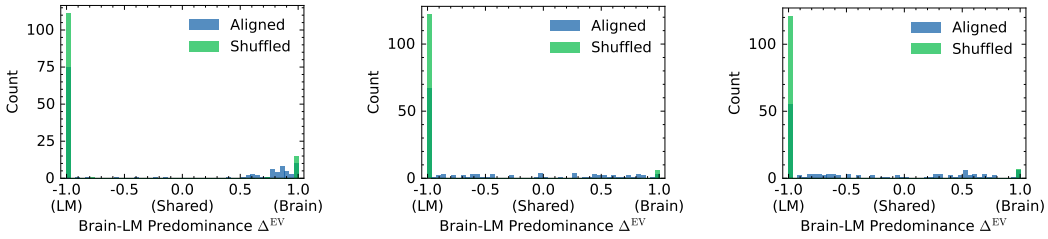

Figure 8: **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Llama-3.1-8B-Instruct layer 12.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive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

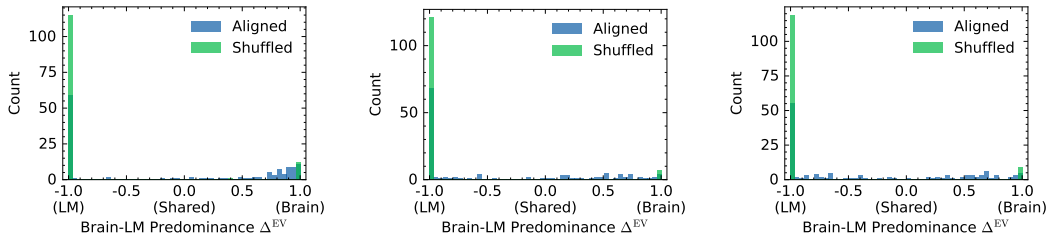

Figure 9: **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Llama-3.1-70B-Instruct layer 19.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive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

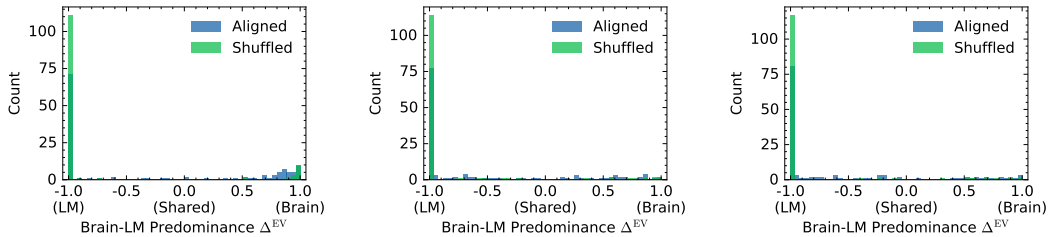

Figure 10: **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Qwen3-8B layer 20.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive values indicate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

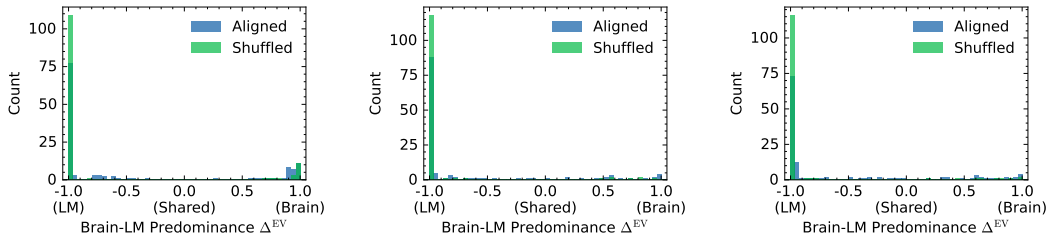

Figure 11: **Distribution of Brain-LM predominance Δ^{EV} for crosscoders trained with Qwen3-32B layer 47.** Each panel shows one subject (UTS01, UTS02, UTS03). Δ^{EV} is the relative difference between leave-one-out explained variance in brain responses and in LM representations; positive values indicate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

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

Feature	Δ^{EV}	Acc.	Explanation
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.
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.

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

Feature	Δ^{EV}	Acc.	Explanation
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.
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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction describe the method, the EV-based predominance score, and the empirical scope. Section 5 states the main limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Section 5 discusses the correlational design, the limits of BOLD fMRI, the use of text transcripts rather than speech, LLM-based interpretation, and the current placeholder cortical-map panels.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper introduces a modeling and analysis framework and evaluates it empirically. It does not state theoretical results that require formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Sections 3 and 4.1 and the appendix specify the dataset, preprocessing, model architecture, predominance score, layer selection, train/validation/test split, and training hyperparameters.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The paper uses a public fMRI dataset and open-weight LLMs.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Section 4.1 reports the data split, LM families and selected layers, dictionary size, sparsity, optimizer, learning rate, batch size, number of epochs, and regularization terms.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The feature interpretation analysis uses one-sided binomial tests with Benjamini-Hochberg FDR correction. Feature predominance is based on leave-one-out EV, and the paired versus unpaired comparison is reported across all subject–model configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: The draft specifies algorithmic hyperparameters but does not report GPU type, memory, runtime, or total compute budget.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work analyzes an existing public neuroimaging dataset and open-weight LMs.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Section 5 includes a paragraph of broader impact describing the work as basic research and noting possible downstream relevance to brain-computer interfaces.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a new pretrained model, image generator, scraped dataset, or other asset with a high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No].

Justification: The existing dataset, LM families, and methodological sources are cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not release a new dataset, model, or software artifact.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper performs secondary analysis of an existing public fMRI dataset and does not conduct new crowdsourcing or human-subject data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not collect new human-subject data. The original dataset paper documents the data collection protocol.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: LMs are used in the method as the representations compared with fMRI responses and as the automatic explainer and detector in Section 4.3.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.