# EXPLORING DIVERGENCE BETWEEN BRAIN RESPONSES AND LANGUAGE MODEL REPRESENTATIONS USING CROSSCODERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

To what extent do human brains and language models (LMs) share internal representations of language, and how do these representations differ? Prior work has shown that LM representations can predict brain responses to naturalistic language stimuli, suggesting that the two systems encode common information. However, it has remained unclear which features are shared between brain responses and LM representations and which are specific to one of them. We propose Brain-LM crosscoders, which decompose brain responses and LM representations into a common set of features and label each feature as shared, a brain-specific candidate, or an LM-specific candidate based on its predictive contribution to brain responses and to LM representations. We apply Brain-LM crosscoders to open fMRI data while participants listened to naturalistic stories. Generated explanations of brain-specific and LM-specific candidate features pose hypotheses about information that only brain responses or only LM representations encode. Brain-LM crosscoders identify the similarities and differences between brain responses and LM representations through individual features instead of a single prediction score, and could help discover concepts of LMs that diverge from those of humans.

## 1 INTRODUCTION

Language neuroscience and language model (LM) interpretability share an interest in how brains and models represent language. Encoding models, which predict brain responses from stimulus features, have long served neuroscience as a tool for characterizing these representations. In functional magnetic resonance imaging (fMRI) studies, early encoding models predicted brain responses to individual words from text corpus statistics (Mitchell et al., 2008), and voxelwise models mapped semantic selectivity across cortex during natural speech (Huth et al., 2016). Contextual representations from recurrent networks improved prediction (Jain & Huth, 2018), and representations of pretrained LMs now predict brain responses to naturalistic stimuli (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022), even when the LM is trained on a developmentally realistic amount of data (Hosseini et al., 2024). These results indicate that LM representations and brain responses encode shared information about the same linguistic stimuli.

Predictive success alone, however, does not guarantee scientific understanding, and moving from prediction to explanation remains a challenge (Serre & Pavlick, 2025). An encoding model compresses the relationship between LM representations and brain responses into a single prediction score, and this score does not reveal which features they share and which features only one of them encodes.

Sparse dictionary learning offers a way to find such features. In LM interpretability, sparse autoencoders (SAEs) decompose internal representations into sparse and interpretable features (Bricken et al., 2023; Huben et al., 2024). Because SAEs enumerate features without supervision, they are particularly suited to discovering concepts that researchers did not specify in advance (Peng et al., 2026; Movva et al., 2025; Adams et al., 2025). Crosscoders extend SAEs to several representations: a single dictionary reconstructs the activations of multiple models and separates features that the models share from features specific to one model (Lindsey et al., 2024; Minder et al., 2025).



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

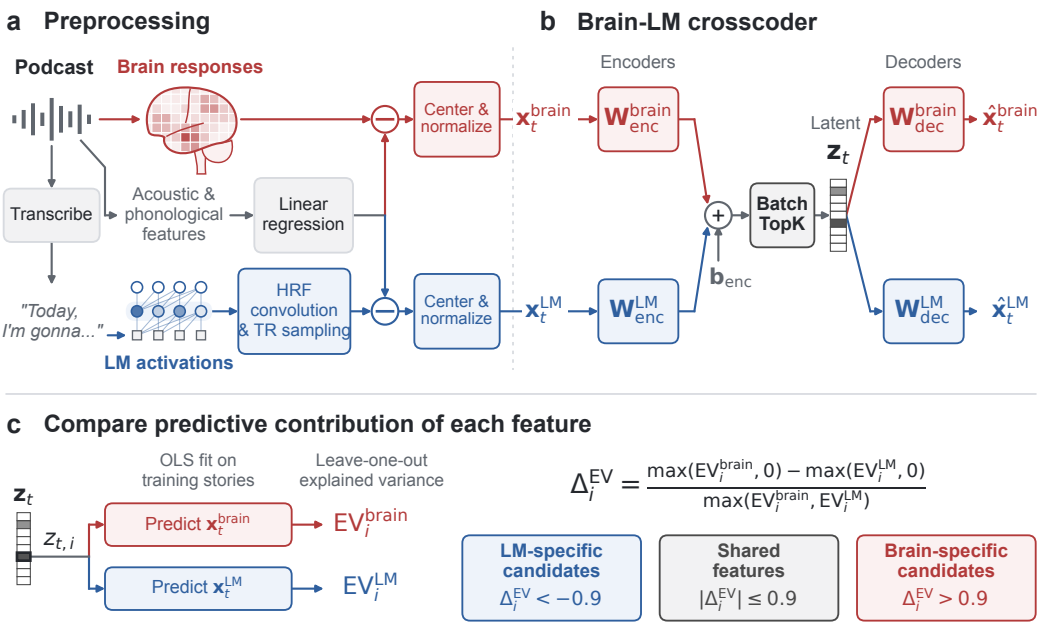



Figure 1: **Overview of the Brain-LM crosscoder. (a)** Language model (LM) activations are convolved with a hemodynamic response function (HRF) and sampled at the fMRI observation times. Components linearly predicted by acoustic and phonological features are subtracted from brain responses and LM representations, which are then centered and normalized. **(b)** The crosscoder encodes the paired vectors into a shared sparse feature vector and reconstructs each with a separate decoder. **(c)** For each feature, leave-one-out explained variance (EV) measures its predictive contribution to brain responses and LM representations. The relative difference $\Delta^{\mathrm{EV}}$ classifies features as shared ($|\Delta^{\mathrm{EV}}| \leq 0.9$), brain-specific candidates ($\Delta^{\mathrm{EV}} > 0.9$), or LM-specific candidates ($\Delta^{\mathrm{EV}} < -0.9$).

We extend crosscoders to human brain responses measured with fMRI and LM representations, which arise in different neural substrates. A *Brain-LM crosscoder* maps brain responses and LM representations to a shared sparse latent vector and reconstructs each with a separate decoder (Figure 1). Each latent coordinate represents a candidate linguistic feature. For each feature, we measure the leave-one-out explained variance (EV) in brain responses and in LM representations. *Brain-LM predominance*, the relative difference between these two EV values, indicates whether a feature predicts brain responses or LM representations better. A brain-specific candidate explains more variance in brain responses than in LM representations, and an LM-specific candidate shows the reverse pattern. Because a single decomposition recovers shared features and both kinds of candidates, the comparison treats the brain and the LM symmetrically.

We apply Brain-LM crosscoders to naturalistic fMRI data from LeBel et al. (2023), covering three subjects and four LMs. An automatic interpretation procedure (Paulo et al., 2025) shows that, for 41% of the evaluated features, generated explanations predict held out feature activations better than chance, and these features are more interpretable than those of a baseline method, ShIndICA (Pandeva & Forré, 2023) (Sections 4.3 and 4.4). For example, the explanation of one brain-specific candidate reads "moments of gaining empathy or understanding others' suffering and pain," and that of one LM-specific candidate reads "vivid descriptions of sudden violent physical impacts or collisions" (Section 4.5). Although these explanations do not necessarily imply a difference in concepts between humans and LMs (Section 5.1), they describe features that encoding models ignore and pose hypotheses for future intervention experiments (Section 5.2). Furthermore, this comparison between brains and LMs matters for AI safety because it could help detect concepts in which AI systems diverge from humans (Section 5.3).

## 2 Related work

**Brain encoding models.** An encoding model is a computational model that describes how input stimuli determine the output response of a cognitive system. In language neuroscience, encoding models predict brain responses from stimulus features (Mitchell et al., 2008; Huth et al., 2016). Contextual LM representations improved prediction over static word embeddings (Jain & Huth, 2018). Subsequent work studied layerwise alignment, scaling, and the relation between next-word prediction and brain predictivity (Toneva & Wehbe, 2019; Antonello et al., 2023; Caucheteux et al., 2023; Merlin & Toneva, 2024; Lei et al., 2025; Lopez-Cardona et al., 2025). These studies evaluate an LM by how well its representations predict brain responses. Brain-LM crosscoders use the same paired data to decompose brain responses and LM representations into sparse features and characterize each feature by its predictive contribution to each of the two.

**Divergences between brains and language models.** Zhou et al. (2024) analyzed words with large encoding errors and found that LMs fail to capture social, emotional, and physical aspects of brain responses. Xu et al. (2025) compared conceptual word ratings between humans and large language models (LLMs) and found that the similarity between humans and LLMs decreases from non-sensorimotor to sensory to motor domains, a result that suggests sensorimotor grounding is a major source of the gap. Mahowald et al. (2024) argued that LLMs show strong formal linguistic competence, whereas functional competence remains uneven and often requires augmentation. These approaches characterize divergences along predefined categories or only in the direction from LMs to the brain, so they do not identify features that LMs capture and the brain does not, or the reverse. By treating the brain and the LM symmetrically, Brain-LM crosscoders identify both brain-specific and LM-specific candidates without predefined categories.

**Dictionary learning for discovery.** Dictionary learning represents observed signals as sparse linear combinations of basis vectors (Olshausen & Field, 1996; Aharon et al., 2006). In neuroscience, dictionary learning has extracted resting-state functional networks and brain regions directly from fMRI data (Eavani et al., 2012; Varoquaux et al., 2011; Abraham et al., 2013; Lee et al., 2011; Dohmatob et al., 2016). In LM interpretability, SAEs, a form of dictionary learning, recover monosemantic and interpretable features from internal representations (Bricken et al., 2023; Huben et al., 2024). SAEs also generate hypotheses for scientific discovery: features of SAEs trained on text embeddings identify concepts that predict a target variable (Movva et al., 2025), features of SAEs trained on a protein LM suggest roles in biological mechanisms (Adams et al., 2025), and Peng et al. (2026) argue that SAEs excel at discovering unknown concepts. Crosscoders and related variants learn one dictionary for several layers or models to identify their differences (Lindsey et al., 2024; Minder et al., 2025; Thasarathan et al., 2025; Kassem et al., 2025; Boughorbel et al., 2025; Jiralerspong & Bricken, 2026), and all of them compare artificial neural networks. Other recent work builds encoding models on interpretable embeddings obtained with dictionary learning (Zeng & Gallant, 2026; Lepori et al., 2026) or with yes-no question answering (Benara et al., 2024; Singh et al., 2025). These models predict brain responses from interpretable features, but they do not identify divergences between brains and LMs.

## 3 Method

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

### 3.1 Preparing brain responses and language model representations

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

**Brain responses and LM activations.** In this paper, a brain response is the blood-oxygen-level-dependent (BOLD) response measured while a subject listens to stories. The response is sampled

at every repetition time (TR) and treated as a vector over voxels. An LM activation is the residual stream vector (i.e., the output of the transformer block) of a subtoken at a specific layer when the same transcript of the story is provided as input to the LM.

**Temporal alignment.** We align LM activations with the BOLD response by convolving their time series with a hemodynamic response function (HRF). We estimate when each subtoken is spoken from word timings and its position within the word, and hold its LM activation constant during that interval. We apply a fixed statistical parametric mapping (SPM) canonical HRF shared across subjects and brain regions (Friston et al., 1998), then sample the convolved time series at the fMRI observation times to obtain LM representations (Appendix B.1).

**Removing acoustic and phonological features.** Human participants listen to spoken stories, whereas language models process text transcripts. To focus the comparison on meaning available from the transcript and its context, we subtract components that acoustic and phonological features linearly predict from both brain responses and LM representations, following the residualization approach of Toneva et al. (2022). We fit separate ridge regressions for brain responses and LM representations using features from the same speech stimuli, after applying the same HRF and sampling times (Appendices B.2 and B.3). The regression for LM representations accounts for information about sound that may also be encoded in representations of text.

**Normalization.** After temporal alignment and residualization, we center each dimension and scale the brain responses and the LM representations by one factor each, using statistics from the training set, so that both have a mean squared $\ell_2$ norm of 1. This prevents the larger of the two from dominating the reconstruction loss. The resulting vectors $\mathbf{x}_t^{\text{brain}}$ and $\mathbf{x}_t^{\text{LM}}$ serve as crosscoder inputs and reconstruction targets.

### 3.2 Brain-LM crosscoder

Crosscoders were proposed to analyze differences between models, and prior work has applied them only to artificial neural networks (Lindsey et al., 2024; Minder et al., 2025; Kassem et al., 2025; Boughorbel et al., 2025; Jiralerspong & Bricken, 2026). The *Brain-LM crosscoder* extends this method to brain responses and LM representations, which arise in different neural substrates (Figure 1).

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

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

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

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

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

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

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

Here, $\text{TV}(\cdot)$ is a total variation penalty (Rudin et al., 1992; Michel et al., 2011) that penalizes large differences between adjacent voxel weights in the brain decoder and reduces overfitting (Section C.2). The term $\mathcal{L}_{\text{aux}}$ is an auxiliary loss that encourages inactive features to explain residual error, and $\alpha$ is its coefficient (Gao et al., 2025) (Section C.3).

### 3.3 BRAIN-LM PREDOMINANCE

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

Specifically, we use leave-one-out explained variance (EV) 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. For each feature, we fit an ordinary least squares regression with an intercept from the scalar activation $z_{t,i}$ to the brain responses or the LM representations on all training time points except one. We apply the fitted regression to the held out time point and compute the prediction residual. The predictions recover the total variance minus the mean squared residual. The EV divides this recovered variance by the predictable variance, 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. Appendix D gives the exact formula. A negative EV indicates that feature $i$ does not generalize to held 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

$$\Delta_i^{\mathrm{EV}} = \frac{\max(\mathrm{EV}_i^{\mathrm{brain}}, 0) - \max(\mathrm{EV}_i^{\mathrm{LM}}, 0)}{\max(\mathrm{EV}_i^{\mathrm{brain}}, \mathrm{EV}_i^{\mathrm{LM}})}. \tag{4}$$

A positive $\Delta_i^{\mathrm{EV}}$ indicates that feature $i$ predicts brain responses better than LM representations, and a negative $\Delta_i^{\mathrm{EV}}$ indicates the reverse. When both $\mathrm{EV}_i^{\mathrm{brain}}$ and $\mathrm{EV}_i^{\mathrm{LM}}$ are non-positive, the feature predicts neither brain responses nor LM representations at held out time points. In this case, the predominance is undefined, and we label the feature non-predictive.

We classify the remaining features into three categories: *shared* features satisfy $|\Delta_i^{\mathrm{EV}}| \leq 0.9$, *brain-specific candidate* features satisfy $\Delta_i^{\mathrm{EV}} > 0.9$, and *LM-specific candidate* features satisfy $\Delta_i^{\mathrm{EV}} < -0.9$.[1]

### 3.4 AUTOMATIC INTERPRETATION

We use an automatic interpretation procedure adapted from recent SAE work (Paulo et al., 2025). Each time point $t$ is represented by a transcript snippet, the segment of the stimulus transcript from 32 seconds before $t$ to $t$. For each feature, an explainer LLM receives the top-activating transcript snippets, namely the snippets at time points where the feature activation $z_{t,i}$ is largest, and transcript snippets where the feature does not activate. The explainer LLM then generates a short description of the difference between the two groups. Next, a detector LLM receives this description and held out transcript snippets, and classifies each snippet as one in which the feature activates or not. We measure interpretability by *detection accuracy*, the balanced accuracy of the detector LLM's classifications. A one sided binomial test with false discovery rate (FDR) correction at $q = 0.05$ identifies descriptions that predict activations in the snippets better than random choice.

## 4 EXPERIMENTS

We train Brain-LM crosscoders on brain responses of three subjects and internal representations of four LMs (Section 4.1). We first test how well generated explanations of the learned features

[1] A feature is a specific candidate if and only if the larger of its two EVs exceeds ten times the smaller. These labels are for convenience. They indicate whether a feature predicts brain responses or LM representations better under this evaluation, and Section 5.1 discusses how they relate to the concepts that each system holds.

predict held out feature activations (Section 4.3) and compare the results with a baseline (Section 4.4). We then examine representative features from each category and the hypotheses that they generate (Section 4.5).

### 4.1 Experimental setup

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

**Language model representations.** The language models used in our experiments are Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen3-8B, and Qwen3-32B (Team, 2025). For each model, we select the residual stream at the layer with the best validation performance in a conventional encoding model fitted before acoustic and phonological residualization. The selected layers are 13 for Llama-3.1-8B-Instruct, 33 for Llama-3.1-70B-Instruct, 16 for Qwen3-8B, and 42 for Qwen3-32B (Appendix E).

**Training configuration.** We choose the dictionary size $d_{\text{latent}}$ (equal to the number of features) from $\{32, 64, 128, 256, 512, 1024\}$ and the BatchTopK sparsity $k$ from $\{4, 8, 16, 32\}$ based on reconstruction loss and detection accuracy on the validation stories (Section 4.2). We train for 300 epochs with batch size 2048. Optimization uses Adam (Kingma & Ba, 2015) with learning rate $10^{-4}$. The coefficient of the auxiliary loss is $\alpha = 0.03125$. Appendices B to G provide the preprocessing, training, and analysis details.

**Baseline.** We use ShIndICA (Pandeva & Forré, 2023) as a baseline because its approach is similar to that of crosscoders. ShIndICA is a multi-view independent component analysis (ICA) that models each of several views as a linear mixture of shared sources and sources specific to that view. With brain responses and LM representations as the two views, each source is shared, specific to brain responses, or specific to LM representations by construction. We fit ShIndICA to the preprocessed training data of the crosscoders for Llama-3.1-8B-Instruct with $c$ shared sources and $(d_{\text{latent}} - c)/2$ sources specific to each view, so that both methods have $d_{\text{latent}} = 512$ features. For each subject, a binary search selects $c$ by the normalized reconstruction error (NRE) of Pandeva & Forré (2023) on the validation stories (Appendix H).

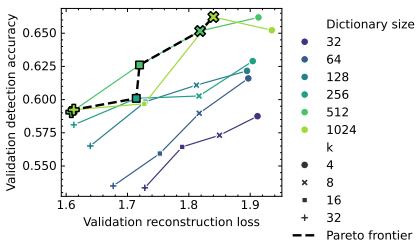



Figure 2: **Reconstruction loss and detection accuracy of crosscoders with different dictionary sizes and sparsity levels.** Each point corresponds to one crosscoder trained on UTS03 and Llama-3.1-8B-Instruct and shows its reconstruction loss and mean detection accuracy on the validation stories. Lower reconstruction loss and higher detection accuracy are better. Colors indicate the dictionary size $d_{\text{latent}}$, markers indicate the BatchTopK sparsity $k$, and lines connect crosscoders with the same dictionary size. The dashed black line traces the Pareto frontier, and markers with black outlines indicate the configurations on it.

### 4.2 Selecting dictionary size and sparsity

We select the dictionary size $d_{\text{latent}}$ and the BatchTopK sparsity $k$ using crosscoders trained on UTS03 and Llama-3.1-8B-Instruct. We train one crosscoder for each of the 24 combinations of the candidate values in Section 4.1 and keep the other training settings unchanged. For each crosscoder, we compute the reconstruction loss and the mean detection accuracy over up to 128 features on the validation stories. Figure 2 shows that the two metrics trade off: within each dictionary size, a larger $k$ generally lowers both the reconstruction loss and the detection accuracy. In particular, the curve for $d_{\text{latent}} = 512$ closely follows the Pareto frontier of the two metrics. Hereafter, we adopt $d_{\text{latent}} = 512$ and $k = 16$ from the Pareto frontier and use this configuration for all subjects and language models.

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

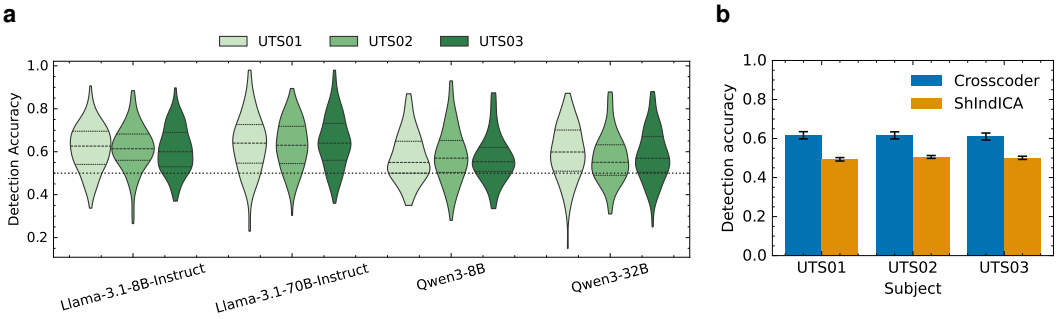



Figure 3: **Detection accuracy of automatically generated feature explanations.** For each evaluated feature, an explainer LLM receives the top-activating and non-activating transcript snippets and generates a short description of the feature. A detector LLM uses this description to classify held out test transcript snippets as activating or non-activating. Detection accuracy is the balanced accuracy of this classification, and chance level is 0.5. **(a)** Detection accuracy of the 128 evaluated features per crosscoder, grouped by subject and language model. **(b)** Detection accuracy of Brain-LM crosscoder features and ShIndICA sources. Both methods decompose the brain responses of each subject and the Llama-3.1-8B-Instruct representations into 512 features. Bars show the mean over up to 128 randomly sampled features on the test stories, and error bars show 95% confidence intervals.

### 4.3 Brain-LM crosscoders learn some interpretable features

We examine to what extent the learned features can be interpreted in terms of the stimulus transcripts with the automatic interpretation procedure (Section 3.4). Across the 1,536 evaluated features (128 per crosscoder), the mean detection accuracy is 0.603, which is above the chance level of 0.5. The explanations of 634 features (41.3%) pass the binomial test after FDR correction, and these features have a mean detection accuracy of 0.713 (Figure 3a and Table 2). Among the evaluated features, significant explanations occur in every category: 194 of 376 brain-specific candidates (51.6%), 63 of 209 shared features (30.1%), and 377 of 950 LM-specific candidates (39.7%) (Appendix I).

### 4.4 Brain-LM crosscoders learn more interpretable features than a baseline

Because ICA determines each source only up to its sign, we keep the fitted signs and use the positive part $\max(z_{t,i}, 0)$ of each ShIndICA source as its activation. Crosscoder features reach mean detection accuracies of 0.61–0.62 on the test stories, whereas ShIndICA sources stay at 0.49–0.51, close to chance (Figure 3b), and no ShIndICA explanation passes the binomial test after FDR correction (Table 3).

### 4.5 Case study: hypotheses from brain-specific and LM-specific candidates

For the Brain-LM crosscoder trained on UTS03 and Llama-3.1-70B-Instruct, Figure 4 shows the feature with the highest detection accuracy in each category, and Appendix K lists the explanations of all 128 evaluated features. The shared feature (Feature 450, $\Delta^{\mathrm{EV}} = +0.89$) activates on detailed sequential descriptions of physical movement through specific locations. The brain-specific candidate (Feature 224, $\Delta^{\mathrm{EV}} = +0.96$) activates on moments of gaining empathy for others or understanding their suffering. The LM-specific candidate (Feature 230, $\Delta^{\mathrm{EV}} = -1.00$) activates on vivid descriptions of sudden violent impacts or collisions.

Each interpretable candidate states a hypothesis about how the two systems differ. Feature 224 suggests that brain responses encode empathic understanding of others along a dimension that explains little variance in the residual stream of the selected LM layer. This hypothesis agrees with the finding that LMs fail to capture social and emotional aspects of brain responses (Zhou et al., 2024) and specifies the content more narrowly. Feature 230 suggests the converse for sudden violent impacts: the LM represents them along a dimension whose time course brain responses do not follow.

### 4.6 Analysis: breaking temporal alignment reduces sharedness

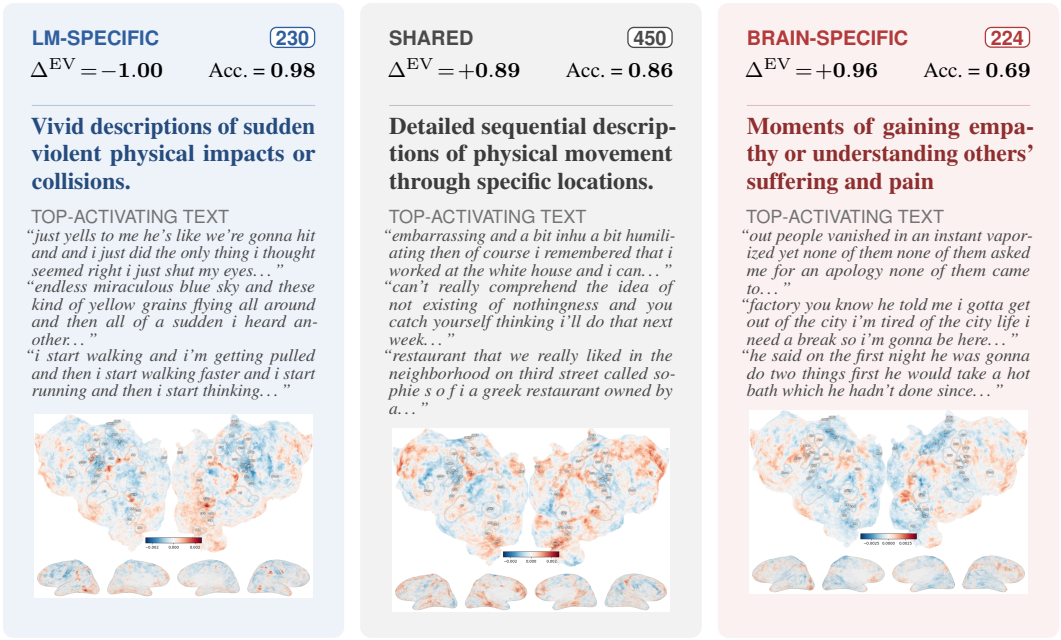



Figure 4: **The feature with the highest detection accuracy in each category, from the Brain-LM crosscoder trained on subject UTS03 with Llama-3.1-70B-Instruct.** $\Delta^{EV}$ is the Brain-LM predominance: positive values indicate that the feature explains more variance in brain responses than in LM representations under leave-one-out cross validation, and negative values indicate the reverse. "Acc." is the detection accuracy of the automatically generated explanation. The text examples per feature are the top-activating transcript snippets. Cortical maps show the brain decoder weights of each feature, the corresponding column of $\mathbf{W}_{\text{dec}}^{\text{brain}}$, on flattened and inflated views of the cortex. Positive weights (red) raise and negative weights (blue) lower the reconstructed response of a voxel when the feature activates.

Shared features should depend on the temporal alignment of Section 3.1 if they reflect information that brain responses and LM representations carry about the same part of a story. To test this, we ablate this alignment by circularly shifting the brain responses of each story by 30 TRs (60 s) relative to the LM representations and retraining all 12 crosscoders. We then compare the two conditions by *sharedness*, the mean of $1 - |\Delta_i^{EV}|$ over the predictive features of a crosscoder. This is 1 if every feature explains the same fraction of variance in brain responses and LM representations and 0 if every feature explains variance in only one of the two. The shift lowers sharedness for all 12 pairs of subject and LM (Figure 5), so shared features depend on the temporal alignment. Story content that persists for more than 60 s may account for the sharedness that remains after the shift.

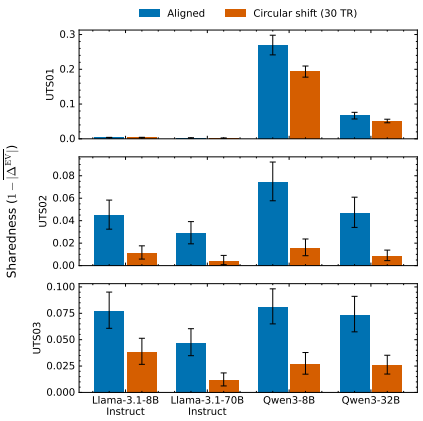



Figure 5: **Sharedness of crosscoders trained with aligned and circularly shifted brain responses.** Sharedness is the mean of $1 - |\Delta_i^{EV}|$ over the features of a crosscoder. The circular shift pairs each LM representation with the brain response recorded 30 TRs (60 s) earlier in the same story and wraps around at the start of the story. Error bars show 95% confidence intervals over features.

## 5 Discussion

### 5.1 Differences in features and differences in concepts

It remains unclear whether a difference in features between brain responses and LM representations implies a difference in the concepts that humans and LMs hold.

Jiralerspong & Bricken (2026) draw the same distinction for crosscoders between LMs: a feature exclusive to one model shows that the model has a distinct direction for a concept with no direct linear analogue in the activation space of the other model, and it leaves open whether the other model knows the concept. The distinction carries more weight when one of the two systems is a brain, because fMRI measures neural activity only indirectly. For Feature 230, the LM-specific label states that a linear map from the activation of the feature explains no held out variance of the measured BOLD response, and the brain may still hold a concept of violent collisions. BOLD responses are slow and noisy and are sampled every 2 seconds, whereas LM representations are deterministic and have thousands of dimensions, so the brain may represent distinctions that fMRI does not resolve. The analysis adds its own constraints: the crosscoder maps both systems into one linear latent space, the EV relies on a linear projection from a single feature activation, and each LM contributes one layer. A concept that a system encodes nonlinearly, across several features, or in another layer can therefore produce a specific candidate. The same reasoning applies to the brain-specific candidate Feature 224, since the LM may represent empathy in a form that the selected layer does not expose linearly.

### 5.2 Significance of discovering differences in features

Nevertheless, our method is likely to be important because it reveals features that many prior studies of the relationship between LM representations and brain responses have ignored. These studies fit a linear map from LM representations to brain responses and report how well the map predicts held out data (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Antonello et al., 2023). The map assigns little weight to LM features that are absent from brain responses, so features that only one system represents can exist even when the prediction score is high. Indeed, in the crosscoder trained on UTS03 and Llama-3.1-70B-Instruct, 101 of the 128 evaluated features are LM-specific candidates (Appendix K). Their explanations include vivid descriptions of physical trauma or injury (Feature 115, detection accuracy 0.93) and events that receive media coverage (Feature 163, detection accuracy 0.95). Brain-LM crosscoders recover such features and give each a natural language explanation, which turns the feature into a known unknown. This enables intervention experiments that test whether each candidate reflects a difference in concepts, in how the concepts are represented, or in what fMRI can measure.

### 5.3 Significance for AI safety

Identifying such features also matters for aligning AI systems with humans. Crosscoders between LMs already reveal behaviors relevant to safety that distinguish one model from another (Jiralerspong & Bricken, 2026). Concept alignment has been proposed as a prerequisite for value alignment, because what people value depends on the concepts through which they understand a situation (Rane et al., 2023), and aligning the internal representations of LLMs with the categories in human moral judgments improves their robustness to adversarial attacks more than aligning their outputs does (Li et al., 2026). These findings motivate direct comparisons between the representations of AI systems and those of humans, and Brain-LM crosscoders make such comparisons one feature at a time. Among the LM-specific candidates for UTS03 and Llama-3.1-70B-Instruct, Feature 434 activates when a narrator makes ethically ambiguous or deceptive decisions with consequences for their family (detection accuracy 0.90). Following Section 5.1, this candidate marks a starting point for a test of concept alignment, which would then determine whether the LM represents deception differently from humans. In this way, Brain-LM crosscoders could help detect concepts in which AI systems diverge from humans.

## 6 Conclusions

We proposed Brain-LM crosscoders, which decompose brain responses and LM representations into one dictionary of features and label each feature as shared, a brain-specific candidate, or an LM-specific candidate by its explained variance in each of the two. Across three subjects and four LMs, explanations of 41% of the evaluated features predict held out feature activations better than chance, whereas those of ShIndICA sources stay at chance level. Although a difference in features does not by itself establish a difference in concepts, each interpretable feature states a hypothesis and turns an *unknown unknown* about how brains and LMs differ into a *known unknown*.

## AI USE STATEMENT

In this work, we used generative AI tools to identify relevant literature, summarize or analyze existing literature, propose new hypotheses, create or edit software code, support qualitative and thematic data analysis, draft parts of a research paper, edit a research paper to improve readability, and translate a research paper into a different language. We did not use generative AI tools to generate synthetic data sets or help develop theoretical models or conceptual frameworks.

We have reviewed and verified all AI-assisted work. LLM-generated code was tested for correctness by the first author. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

Our code is available at https://anonymous.4open.science/r/brain-lm-crosscoder-dev-67E8. Experimental setup and hyperparameters are provided in Section 4.1.

## REFERENCES

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fmri with total-variation constrained dictionary learning. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pp. 607–615, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40763-5.

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=zdOGBRQEbz.

M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. doi: 10.1109/TSP.2006.881199.

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

Vinamra Benara, Chandan Singh, John Xavier Morris, Richard Antonello, Ion Stoica, Alexander Huth, and Jianfeng Gao. Crafting interpretable embeddings for language neuroscience by asking LLMs questions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mxMvWwyBWe.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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

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

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

Nilearn contributors. nilearn. URL https://github.com/nilearn/nilearn.

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

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

K.J. Friston, P. Fletcher, O. Josephs, A. Holmes, M.D. Rugg, and R. Turner. Event-related fmri: Characterizing differential responses. *NeuroImage*, 7(1):30–40, 1998. ISSN 1053-8119. doi: https://doi.org/10.1006/nimg.1997.0306. URL https://www.sciencedirect.com/science/article/pii/S1053811997903062.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

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

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

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

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

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

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023.

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

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

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

Lingyu Li, Yan Teng, Yingchun Wang, and Xia Hu. Representational alignment yields generalizable safety in language models. *arXiv preprint arXiv:2609.04022*, 2026.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse Crosscoders for Cross-Layer Features and Model Diffing. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

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

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

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

Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, and Bertrand Thirion. Total variation regularization for fmri-based prediction of behavior. *IEEE Transactions on Medical Imaging*, 30(7):1328–1340, 2011. doi: 10.1109/TMI.2011.2113378.

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

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008. doi: 10.1126/science.1152876. URL https://www.science.org/doi/abs/10.1126/science.1152876.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4R0pugRyN5.

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

Teodora Pandeva and Patrick Forré. Multi-view independent component analysis with shared and individual sources. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1639–1650. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/pandeva23a.html.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EemtbhJOXc.

Kenny Peng, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. Position: Use sparse autoencoders to discover unknowns. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://openreview.net/forum?id=x1Px74tbvs.

Sunayana Rane, Mark Ho, Ilia Sucholutsky, and Thomas L Griffiths. Concept alignment as a prerequisite for value alignment. *arXiv preprint arXiv:2310.20059*, 2023.

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

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105646118.

Thomas Serre and Ellie Pavlick. From prediction to understanding: Will ai foundation models transform brain science? *Neuron*, 113(21):3504–3508, 2025.

Chandan Singh, Richard J. Antonello, Sihang Guo, Gavin Mischler, Jianfeng Gao, Nima Mesgarani, and Alexander G. Huth. Evaluating scientific theories as predictive models in language neuroscience. *bioRxiv*, 2025. doi: 10.1101/2025.08.12.669958. URL https://www.biorxiv.org/content/early/2025/08/12/2025.08.12.669958.

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

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

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UoaxRN88oR.

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

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757, 2022.

Gael Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In Gábor Székely and Horst K. Hahn (eds.), *Information Processing in Medical Imaging*, pp. 562–573, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-22092-0.

Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, 9(9):1871–1886, 2025.

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

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

# A NOTATION

Table 1: Notation used in this paper.

| Symbol | Description |
|---|---|
| $t$ | time point on the repetition-time grid |
| $n$ | number of training TRs |
| $d_{\text{brain}}$ | dimensionality of the brain response, equal to the number of fMRI voxels |
| $d_{\text{LM}}$ | dimensionality of the LM representation |
| $d_{\text{latent}}$ | dimensionality of the sparse latent space, equal to the dictionary size |
| $k$ | target number of active features in BatchTopK sparsification |
| $\mathbf{x}_t^{\text{brain}} \in \mathbb{R}^{d_{\text{brain}}}$ | brain response after preprocessing at time $t$ |
| $\mathbf{x}_t^{\text{LM}} \in \mathbb{R}^{d_{\text{LM}}}$ | LM representation after preprocessing 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 $i$th 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 |
| $\mathbf{W}_{\text{dec},i}^{\text{brain}} \in \mathbb{R}^{d_{\text{brain}}}$ | $i$th column of $\mathbf{W}_{\text{dec}}^{\text{brain}}$, the brain decoder vector for feature $i$ |
| $\mathbf{W}_{\text{dec},i}^{\text{LM}} \in \mathbb{R}^{d_{\text{LM}}}$ | $i$th 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 $i$ |
| $\text{EV}_i^{\text{LM}}$ | LM explained variance of feature $i$ |
| $\sigma^2_{\text{noise}}$ | brain noise variance estimate |
| $\Delta_i^{\text{EV}}$ | relative explained-variance difference of feature $i$ |

# B PREPROCESSING DETAILS

## B.1 TEMPORAL ALIGNMENT

Word onset and offset times and character positions within each word define the estimated speech interval of each subtoken. We hold its LM activation constant over that interval and convolve the resulting stimulus time series with a fixed SPM canonical HRF. The HRF follows the function and kernel normalization of Nilearn's SPM implementation (contributors), spans 32 seconds, and includes the negative undershoot. We sample the result at the fMRI observation times, beginning 11 seconds after story onset and continuing every 2 seconds thereafter. This temporal transformation supplies LM representations for layer selection, training, reconstruction, and evaluation.

## B.2 ACOUSTIC AND PHONOLOGICAL FEATURES

We construct a common feature matrix from the stimulus audio and alignments. Acoustic features include the log-Mel spectrum in 80 frequency bands, the amplitude envelope, and acoustic onset strength (peakRate). PeakRate events mark positive local maxima of the envelope derivative, with event amplitudes equal to those maxima. Pitch features describe the logarithm of absolute fundamental frequency (F0), relative F0 normalized within each story and encoded in 10 bins, and separate rises and falls in F0 during voiced intervals. Phonological and timing features encode speech, word, and phoneme onsets, phoneme and diphone identities, and primary and secondary lexical stress. Phoneme identities use one-hot vectors over their durations, with stress encoded separately; diphone events occur at the second phoneme's onset, and stress events at the stressed vowel's onset. We apply the same HRF and fMRI sampling times to these features as to the LM activations.

After HRF convolution and sampling, we center all feature columns and scale log-Mel power, envelope amplitude, peakRate magnitude, absolute log-F0, and F0 changes to unit variance. We estimate the means and standard deviations from the training data and apply them to validation and test data.

### B.3 Residualization

We fit separate ridge regressions to predict each brain voxel and each LM dimension from the acoustic and phonological features. We split the training stories into five folds and select each output dimension's regularization coefficient from $\{1, 10, 100, 1000, 10000, 100000\}$ by its mean squared prediction error across the held out folds. Within each fold, the training portion supplies all fitted normalization statistics. We then refit the regressions, including their intercepts, on all training stories and apply the fixed models to the training, validation, and test sets. Let $\mathbf{q}_t$ denote the vector of preprocessed acoustic and phonological features, and let $y^m_{t,j}$ denote dimension $j$ of the aligned vector for modality $m$ before residualization. The residual is

$$y^{m,\mathrm{res}}_{t,j} = y^m_{t,j} - \left(\hat{b}^m_j + \mathbf{q}_t^\top \hat{\mathbf{w}}^m_j\right), \qquad m \in \{\mathrm{brain}, \mathrm{LM}\}. \tag{5}$$

We center and normalize these residuals using training statistics (Section 3.1); layer selection uses brain responses and LM representations before residualization (Appendix E).

## C CrossCoder architecture details

### C.1 BatchTopK activation function

BatchTopK (Bussmann et al., 2024) is a sparse activation function that enforces an approximate $\ell_0$ sparsity constraint without an explicit $\ell_1$ penalty on the activations. Let $\mathbf{a}_t = \mathbf{W}^{\mathrm{brain}}_{\mathrm{enc}}\mathbf{x}^{\mathrm{brain}}_t + \mathbf{W}^{\mathrm{LM}}_{\mathrm{enc}}\mathbf{x}^{\mathrm{LM}}_t + \mathbf{b}_{\mathrm{enc}} \in \mathbb{R}^{d_{\mathrm{latent}}}$ denote the pre-activation vector at time $t$. Each pre-activation $a_{t,i}$ is scaled by the factor $s_i = \|\mathbf{W}^{\mathrm{brain}}_{\mathrm{dec},i}\|_2 + \|\mathbf{W}^{\mathrm{LM}}_{\mathrm{dec},i}\|_2$ before ranking. 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:

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

where $\theta_B$ is the $Bk$-th largest value among the $B \cdot d_{\mathrm{latent}}$ 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 $\theta$, estimated during training as an exponential moving average of $\theta_B$ with decay $\gamma$:

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

The inference-time activation is

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

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

Unlike sparse autoencoders with $\ell_1$ regularization, BatchTopK avoids feature suppression, in which the $\ell_1$ 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 feature $i$, with brain-decoder weights $\mathbf{w}_i \in \mathbb{R}^V$, we compute

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

The TV term in the loss is the mean of $\mathrm{TV}(\mathbf{w}_i)$ over 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 $\mathcal{D} \subseteq \{1, \dots, d_{\text{latent}}\}$ 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 $\mathbf{W}_{\text{dec}}^m$. The auxiliary loss measures how well the dead features reduce the reconstruction residual:

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

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

## D EXPLAINED VARIANCE WITH LEAVE-ONE-OUT CROSS VALIDATION

For the EV in Section 3.3, we fit a regression for each feature on the training stories and evaluate it at each training TR with leave-one-out cross validation (LOOCV). Let $\mathcal{T}_{\text{train}}$ be the set of $n$ training TRs. Let $\mathbf{Y} \in \mathbb{R}^{n \times m}$ be a training target, whose rows $\mathbf{y}_t \in \mathbb{R}^m$ are indexed by $t \in \mathcal{T}_{\text{train}}$. The feature activations $z_{t,i}$ are entries of $\mathbf{Z}$. Let $\bar{z}_i$ and $\bar{\mathbf{y}}$ denote the means of $z_{t,i}$ and $\mathbf{y}_t$ over $\mathcal{T}_{\text{train}}$. The least-squares regression of $\mathbf{y}_t$ on $z_{t,i}$ with an intercept has the slope and intercept

$$\hat{\boldsymbol{\beta}}_i(\mathbf{Y}) = \frac{\sum_{t \in \mathcal{T}_{\text{train}}} (z_{t,i} - \bar{z}_i)(\mathbf{y}_t - \bar{\mathbf{y}})}{\sum_{t \in \mathcal{T}_{\text{train}}} (z_{t,i} - \bar{z}_i)^2} \in \mathbb{R}^m, \qquad \hat{\mathbf{b}}_i(\mathbf{Y}) = \bar{\mathbf{y}} - \bar{z}_i \hat{\boldsymbol{\beta}}_i(\mathbf{Y}) \in \mathbb{R}^m. \tag{12}$$

With the leverage of TR $t$,

$$h_{t,i} = \frac{1}{n} + \frac{(z_{t,i} - \bar{z}_i)^2}{\sum_{s \in \mathcal{T}_{\text{train}}} (z_{s,i} - \bar{z}_i)^2}, \tag{13}$$

the regression refitted without TR $t$ leaves the residual

$$\hat{\boldsymbol{\epsilon}}_{t,i}^{\text{LOO}}(\mathbf{Y}) = \frac{\mathbf{y}_t - \hat{\mathbf{b}}_i(\mathbf{Y}) - z_{t,i} \hat{\boldsymbol{\beta}}_i(\mathbf{Y})}{1 - h_{t,i}} \tag{14}$$

at TR $t$, so the $n$ refits need not be computed separately. The total variance is

$$V(\mathbf{Y}) = \frac{1}{n} \sum_{t \in \mathcal{T}_{\text{train}}} \|\mathbf{y}_t - \bar{\mathbf{y}}\|_2^2, \tag{15}$$

and the LOOCV EV is

$$\text{EV}_i(\mathbf{Y}, \sigma^2) = \frac{V(\mathbf{Y}) - \frac{1}{n} \sum_{t \in \mathcal{T}_{\text{train}}} \|\hat{\boldsymbol{\epsilon}}_{t,i}^{\text{LOO}}(\mathbf{Y})\|_2^2}{V(\mathbf{Y}) - \sigma^2}. \tag{16}$$

The numerator is the variance that the held out predictions recover, and the denominator subtracts the noise variance $\sigma^2$ from the total variance to obtain the predictable variance. The brain and LM explained variances of feature $i$ are

$$\text{EV}_i^{\text{brain}} = \text{EV}_i(\mathbf{X}^{\text{brain}}, \sigma^2_{\text{noise}}), \qquad \text{EV}_i^{\text{LM}} = \text{EV}_i(\mathbf{X}^{\text{LM}}, 0), \tag{17}$$

where $\mathbf{X}^{\text{brain}}$ and $\mathbf{X}^{\text{LM}}$ contain the training brain responses and LM representations. We estimate the brain noise variance $\sigma^2_{\text{noise}}$ from repeated responses to the story `wheretheressmoke` after applying the residualization and normalization used for the crosscoder inputs. 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

Layer selection uses brain responses and LM representations before acoustic and phonological residualization. For each candidate LM layer, kernel ridge regression predicts brain responses from LM representations after HRF convolution and sampling. We center and scale the inputs and targets using training statistics. We select the regularization coefficient from $\{1, 10, 100, 1000, 10000, 100000\}$ by cross validation with five folds of the training data and choose the layer with the highest mean correlation on the validation data. This procedure selects layer 13 for Llama-3.1-8B-Instruct, layer 33 for Llama-3.1-70B-Instruct, layer 16 for Qwen3-8B, and layer 42 for Qwen3-32B. Figure 6 shows the validation mean correlation across normalized layer depth for the language models used in the layer-selection analysis.

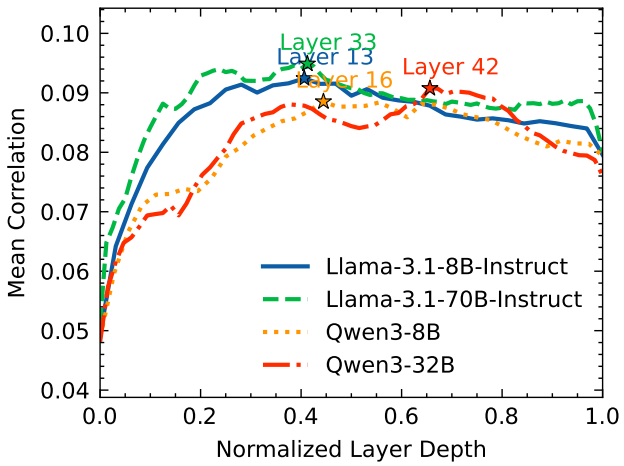



Figure 6: **Layer selection for language model representations used in crosscoder training.** For each language model layer, a kernel ridge regression encoding model predicts brain responses from LM representations after HRF convolution and sampling. The plot shows mean Pearson correlation between predicted and observed brain responses on the validation set as a function of normalized layer depth (layer index divided by total number of layers). Stars mark the selected layer for each model.

## F DETECTION ACCURACY BY SUBJECT AND LANGUAGE MODEL

Table 2 lists the mean detection accuracy and the number of significant features for each of the 12 crosscoders.

## G STATISTICAL TESTING DETAILS OF AUTOMATIC INTERPRETATION

We test detection accuracy against chance with a one sided binomial test and a null probability of 0.5. For each feature, the detector classifies up to 50 positive and 50 negative transcript snippets from the test set; a feature that activates at fewer than 50 test time points contributes all of its activating snippets. Detection accuracy is the balanced accuracy, the mean of the fractions of positive and negative snippets that the detector classifies correctly. Under the null hypothesis, the numbers of correctly classified positive and negative snippets follow independent binomial distributions with probability 0.5, and the p-value is the probability under these distributions that the detection accuracy is at least the observed value. We apply Benjamini–Hochberg FDR correction at $q = 0.05$ within each configuration of subject and model.

Table 2: **Detection accuracy of automatically generated feature explanations for each of the 12 subject–language model pairs.** Each crosscoder learns 512 features, of which 128 are evaluated. For each evaluated feature, an explainer LLM generates a description from the top-activating transcript snippets, and a detector LLM classifies held-out test snippets using this description. Mean detection accuracy is the average of the detection accuracies of the 128 evaluated features. Significant features is the number of evaluated 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 | Significant features |
|---|---|---|---|
| Llama-3.1-8B-Instruct | UTS01 | 0.617 | 66/128 |
| Llama-3.1-8B-Instruct | UTS02 | 0.617 | 69/128 |
| Llama-3.1-8B-Instruct | UTS03 | 0.610 | 62/128 |
| Llama-3.1-70B-Instruct | UTS01 | 0.634 | 70/128 |
| Llama-3.1-70B-Instruct | UTS02 | 0.633 | 73/128 |
| Llama-3.1-70B-Instruct | UTS03 | 0.646 | 75/128 |
| Qwen3-8B | UTS01 | 0.575 | 30/128 |
| Qwen3-8B | UTS02 | 0.584 | 38/128 |
| Qwen3-8B | UTS03 | 0.570 | 20/128 |
| Qwen3-32B | UTS01 | 0.596 | 56/128 |
| Qwen3-32B | UTS02 | 0.571 | 36/128 |
| Qwen3-32B | UTS03 | 0.581 | 39/128 |

## H ShIndICA baseline

**Model.** ShIndICA (Pandeva & Forré, 2023) models view $m$ as $\mathbf{x}^m = \mathbf{A}^m(\tilde{\mathbf{s}}^m + \boldsymbol{\varepsilon}^m)$, where $\tilde{\mathbf{s}}^m$ stacks $c$ sources that all views share and the sources specific to view $m$, $\mathbf{A}^m$ is an invertible mixing matrix, and $\boldsymbol{\varepsilon}^m$ is Gaussian noise. The two views are the preprocessed brain responses and LM representations of Section 3.1. We reduce each view to its first $r = (d_{\text{latent}} + c)/2$ principal components on the training stories and whiten it. Orthogonal unmixing matrices $\mathbf{W}^m \in \mathbb{R}^{r\times r}$ map the whitened views to $\mathbf{y}_t^m$, whose first $c$ coordinates estimate the shared sources and whose remaining $(d_{\text{latent}} - c)/2$ coordinates estimate the sources specific to view $m \in \{\text{brain}, \text{LM}\}$. The feature vector $\mathbf{z}_t \in \mathbb{R}^{d_{\text{latent}}}$ concatenates the average of the two estimates of the $c$ shared sources with the $d_{\text{latent}} - c$ specific sources of both views, so it counts each shared source once.

**Estimation.** Following Eq. (3) of Pandeva & Forré (2023) for two views with $g(s) = -\log\cosh s$ and $\lambda = 1$, we minimize

$$\frac{1}{n}\sum_{t\in\mathcal{T}_{\text{train}}}\left(\sum_{i=1}^{d_{\text{latent}}}\log\cosh z_{t,i} - \frac{\lambda}{2}\sum_{j=1}^{c} y_{t,j}^{\text{brain}}\, y_{t,j}^{\text{LM}}\right) \tag{18}$$

over the orthogonal matrices. The first term is the negative log density of a super-Gaussian source model up to a constant, and the second term rewards agreement between the two estimates of each shared source.

**Selecting the number of shared sources.** A binary search runs over even values of $c$ from 2 to 510. For each evaluated candidate, we fit ShIndICA on the training stories and compute the NRE of Pandeva & Forré (2023) on the validation stories,

$$\text{NRE}(c) = \frac{1}{2\, n_{\text{val}}\, c}\sum_{t\in\mathcal{T}_{\text{val}}}\left\|\mathbf{y}_{t,1:c}^{\text{brain}} - \mathbf{y}_{t,1:c}^{\text{LM}}\right\|_2^2, \tag{19}$$

where $\mathcal{T}_{\text{val}}$ is the set of $n_{\text{val}}$ validation TRs. Among the 16 candidates evaluated for each subject, we select the one with the smallest NRE and choose the largest $c$ in case of a tie, following Pandeva & Forré (2023). For all three subjects, $c = 2$ yields the smallest NRE (0.48–0.56), and every other evaluated candidate yields at least 0.72.

**Feature evaluation.** We keep the fitted signs of the sources and apply $\max(z_{t,i}, 0)$ only in automatic interpretation, so time points with non-positive values supply the non-activating snippets. All 512

ShIndICA features of each subject take both positive and non-positive values in the training and test stories, so every feature is eligible for evaluation. For each subject, we randomly sample 128 features and apply the procedure of Section 3.4; each sample contains one shared source, 63 sources specific to brain responses, and 64 sources specific to LM representations. For the crosscoder, we use the 128 features of Section 4.3 for Llama-3.1-8B-Instruct, which were sampled at random from the features that activate in both the training and test stories. The error bars in Figure 3b are 95% percentile intervals from 10,000 bootstrap resamples of the evaluated features, and significance follows the binomial test and FDR correction of Appendix G within each pair of subject and method (Table 3).

Table 3: **Detection accuracy of Brain-LM crosscoder features and ShIndICA sources for Llama-3.1-8B-Instruct.** Evaluated counts the sampled features with a valid detection result, and excluded counts the sampled features whose detection failed. Brackets show 95% confidence intervals. Significant counts the evaluated features whose detection accuracy exceeds chance after FDR correction.

| Subject | Method | $c$ | Evaluated | Excluded | Mean detection accuracy | Significant |
|---|---|---|---|---|---|---|
| UTS01 | Brain-LM crosscoder | – | 128 | 0 | 0.617 [0.598, 0.635] | 66 |
| UTS01 | ShIndICA | 2 | 128 | 0 | 0.494 [0.485, 0.502] | 0 |
| UTS02 | Brain-LM crosscoder | – | 128 | 0 | 0.617 [0.598, 0.634] | 69 |
| UTS02 | ShIndICA | 2 | 128 | 0 | 0.506 [0.498, 0.513] | 0 |
| UTS03 | Brain-LM crosscoder | – | 128 | 0 | 0.610 [0.592, 0.628] | 62 |
| UTS03 | ShIndICA | 2 | 127 | 1 | 0.501 [0.493, 0.509] | 0 |

## I Detection accuracy by feature category

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

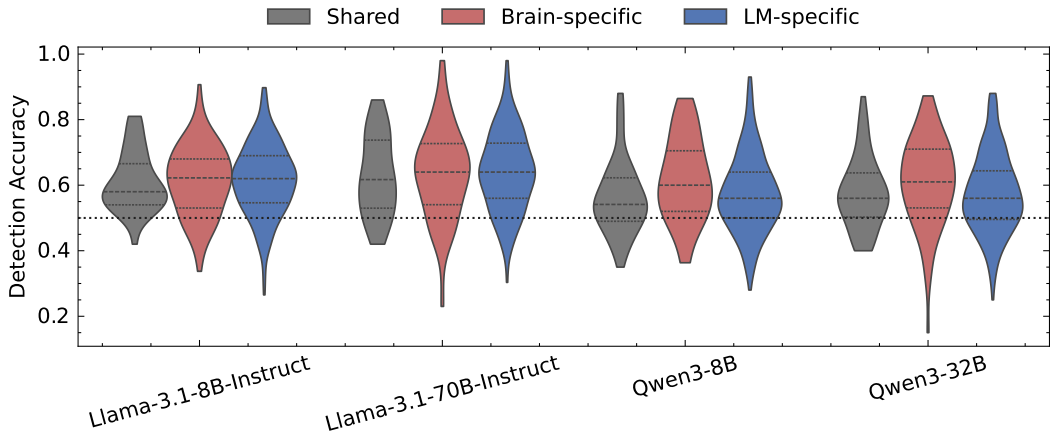



Figure 7: **Detection accuracy of automatically generated feature explanations, grouped by feature category.** Features are classified as shared ($|\Delta^{\mathrm{EV}}| \leq 0.9$), brain-specific candidates ($\Delta^{\mathrm{EV}} > 0.9$), or LM-specific candidates ($\Delta^{\mathrm{EV}} < -0.9$) based on the relative difference between explained variance for brain responses and for LM representations, both estimated with leave-one-out cross validation on the training stories. The dotted line marks chance level (accuracy = 0.5). The one evaluated feature labeled "neither" (both EV values non-positive) is omitted.

## J Automatic interpretation prompts

The explainer prompt presents positive and negative transcript snippets for one feature and asks the LLM to generate a short description that distinguishes the positive snippets from the negative snippets. The detector prompt presents the generated description and a held out transcript snippet, then asks

for a binary prediction of whether the feature should activate. Three fixed example exchanges, each pairing an explanation and five text examples with the expected labels, precede the user message of the detector. 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 10 highest-activation transcript snippets, with at most one snippet per story. Negative examples are 10 randomly selected zero-activation snippets. For detection, we use the test split. Positive examples are up to 50 activating snippets sampled across 50 activation quantiles, and negative examples are 50 randomly selected zero-activation snippets. The detector receives shuffled batches of five snippets. Both the explainer and the 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 $q = 0.05$.

### J.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
```

### J.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}}
```

## K EVALUATED FEATURES FOR UTS03 AND LLAMA-3.1-70B-INSTRUCT

Table 4 lists the evaluated features from the crosscoder trained on UTS03 and Llama-3.1-70B-Instruct.

Table 4: 128 evaluated features from the Brain-LM crosscoder trained on subject UTS03, Llama-3.1-70B-Instruct, sorted by Brain-LM predominance $\Delta^{\mathrm{EV}}$. $\Delta^{\mathrm{EV}}$ is the relative difference between explained variance in brain responses and in LM representations, both estimated with leave-one-out cross validation on the training stories; 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 snippets.

| Feature | $\Delta^{\mathrm{EV}}$ | Acc. | Explanation |
|---|---|---|---|
| 285 | 1.00 | 0.45 | Narrator's self-awareness or personal character growth from experience. |
| 457 | 1.00 | 0.47 | Narrator experiencing raw emotional vulnerability amid vivid sensory detail. |
| 444 | 0.98 | 0.58 | Interwoven sensory details and reflective personal commentary about meaning. |
| 156 | 0.98 | 0.51 | Internal monologue revealing conflicting emotions and self-doubt. |
| 73 | 0.96 | 0.47 | Intimate everyday social moments reflecting personal growth or relationships. |
| 224 | 0.96 | 0.69 | Moments of gaining empathy or understanding others' suffering and pain |
| 157 | 0.95 | 0.44 | Observational revelation of individual variation and hidden differences within categories. |
| 345 | 0.94 | 0.64 | Opening sentences that establish specific timeframes or life milestones. |
| 456 | 0.93 | 0.61 | Conversational interactions where the speaker observes multiple people simultaneously engaging in different activities. |
| 121 | 0.89 | 0.53 | Emotional conflict or uncertainty about romantic relationships and personal life choices. |
| 450 | 0.89 | 0.86 | Detailed sequential descriptions of physical movement through specific locations. |
| 19 | 0.87 | 0.61 | Introspective moments revealing personal identity or existential awareness. |
| 287 | 0.78 | 0.61 | Personal connection to famous historical figures or movements. |
| 60 | 0.73 | 0.77 | Sudden unexpected physical sensory disruption or interruption. |
| 486 | 0.62 | 0.54 | Moments featuring acute physical or social vulnerability with visceral reactions. |
| 379 | 0.61 | 0.74 | Specific geographical locations establishing scene before narrative action. |
| 477 | 0.60 | 0.77 | Vivid, specific physical descriptions of individual people observed by narrator. |
| 189 | 0.56 | 0.73 | Active information-seeking behavior through direct contact or research. |
| 441 | 0.47 | 0.62 | Specific physical objects given, received, or discovered within the narrative. |

Table 4: 128 evaluated features for UTS03, Llama-3.1-70B-Instruct (continued).

| Feature | $\Delta^{\text{EV}}$ | Acc. | Explanation |
|---|---|---|---|
| 508 | 0.34 | 0.48 | Contrasting opposing concepts or states within personal narratives. |
| 267 | 0.33 | 0.56 | Processing emotional aftermath of difficult personal revelations or experiences. |
| 32 | 0.14 | 0.52 | Explicit emotional reflection or cathartic emotional revelation by narrator. |
| 112 | -0.22 | 0.82 | Explicit acknowledgment of mental health conditions or psychological struggles. |
| 118 | -0.25 | 0.76 | Specific ages and historical time periods marking life transitions. |
| 206 | -0.29 | 0.49 | Descriptions of intense internal emotional pain or psychological distress. |
| 302 | -0.68 | 0.65 | Encountering or navigating through large crowds or masses of people. |
| 364 | -0.78 | 0.63 | Narrative moments of absolution, forgiveness, or transcendent emotional resolution. |
| 4 | -1.00 | 0.74 | Specific visual details describing people, objects, or domestic settings. |
| 5 | -1.00 | 0.62 | Vivid descriptions of physical injury or bodily harm occurring to people. |
| 10 | -1.00 | 0.64 | References to cultural traditions, spiritual beliefs, or philosophical frameworks. |
| 13 | -1.00 | 0.51 | Active pursuit of knowledge through formal education or structured learning. |
| 28 | -1.00 | 0.63 | Witnessing unexpected moments that trigger emotional or perspective shifts. |
| 30 | -1.00 | 0.76 | Vivid descriptions of physical journeys through unexpected landscapes and environments. |
| 35 | -1.00 | 0.62 | Vivid sensory details marking unforgettable moments of personal trauma or crisis. |
| 40 | -1.00 | 0.81 | Unexpected surreal physical events with absurdist sensory detail. |
| 52 | -1.00 | 0.66 | Entering a new group or institutional setting with peers. |
| 55 | -1.00 | 0.60 | Reflections on how relationships revealed new truths about oneself. |
| 57 | -1.00 | 0.58 | Direct quotations or verbatim spoken words embedded in the narrative. |
| 58 | -1.00 | 0.60 | Narrative of reaching out to reconnect with estranged people. |
| 61 | -1.00 | 0.72 | Introspection about ordinary life with specific consumer product references. |
| 71 | -1.00 | 0.53 | Narrator struggling to communicate or lacking control in interpersonal exchanges. |
| 74 | -1.00 | 0.56 | Humorous observations about human folly and social absurdities. |
| 83 | -1.00 | 0.55 | Elaborately constructed imaginary worlds or detailed creative scenarios. |
| 85 | -1.00 | 0.66 | Creating false identities or personas to deceive others online. |
| 88 | -1.00 | 0.65 | Reflection on irreversible loss or profound personal tragedy |
| 91 | -1.00 | 0.77 | Detailed visual descriptions of specific objects and physical scenes. |
| 92 | -1.00 | 0.54 | References to significant historical events or public figures. |
| 99 | -1.00 | 0.75 | Receiving devastating medical diagnoses or health-related news from doctors. |
| 103 | -1.00 | 0.63 | Specific details about animals encountered during storytelling. |
| 108 | -1.00 | 0.64 | Characters contemplating competing life paths or identity realignments. |
| 111 | -1.00 | 0.50 | Explicit graphic descriptions of bodily damage or bodily functions. |
| 115 | -1.00 | 0.93 | Vivid descriptions of physical trauma, injury, or life-threatening accidents. |
| 123 | -1.00 | 0.64 | Urgent need for transportation met by unexpected stranger assistance. |
| 124 | -1.00 | 0.79 | Personal struggles with mental health or medical crises. |
| 125 | -1.00 | 0.50 | Narrator being taught or mentored with moments of realization. |
| 127 | -1.00 | 0.82 | Institutional settings with specific procedural calls and roll call sequences |
| 136 | -1.00 | 0.44 | Concrete details about specific people directly involved in events. |
| 140 | -1.00 | 0.62 | Internal emotional states, particularly loneliness, longing, and melancholy. |
| 142 | -1.00 | 0.65 | Explicit reflection on personal psychological or emotional processing. |
| 143 | -1.00 | 0.52 | Casual, sensory-detailed recollections of everyday social moments |
| 148 | -1.00 | 0.78 | Narrator reaching for and discovering specific physical objects. |
| 153 | -1.00 | 0.50 | Dramatic metaphors describing emotional or physical overwhelm intensely |
| 155 | -1.00 | 0.66 | Comparisons between personal experiences and fictional movies or shows. |
| 161 | -1.00 | 0.79 | Internal conflict about romantic relationships and personal identity choices. |
| 163 | -1.00 | 0.95 | Events receiving media coverage or widespread public attention and visibility. |
| 165 | -1.00 | 0.64 | Specific named objects with embedded personal or emotional significance. |
| 170 | -1.00 | 0.57 | Personal realization or self-awareness emerging from unexpected or transgressive experiences. |
| 174 | -1.00 | 0.64 | Extended anticipation or waiting followed by a decisive moment or outcome. |
| 188 | -1.00 | 0.54 | Immediate physical sensations and visceral internal emotional reactions. |
| 195 | -1.00 | 0.61 | Unexpected plot twists and specific surprising narrative revelations. |
| 197 | -1.00 | 0.56 | Specific vivid sensory details vividly describing objects and surroundings. |
| 199 | -1.00 | 0.68 | Narrator acknowledging unspoken emotional truths within family moments. |
| 202 | -1.00 | 0.60 | Detailed descriptions of specific objects and visual sensory details. |

Table 4: 128 evaluated features for UTS03, Llama-3.1-70B-Instruct (continued).

| Feature | $\Delta^{\text{EV}}$ | Acc. | Explanation |
|---|---|---|---|
| 211 | -1.00 | 0.71 | Immediate life-threatening or high-stakes medical emergencies experienced firsthand. |
| 212 | -1.00 | 0.54 | Vivid sensory and physical descriptions creating embodied emotional moments |
| 214 | -1.00 | 0.73 | People using humor or laughter to cope with difficult situations. |
| 216 | -1.00 | 0.51 | Describing deliberate preparation and incremental step-by-step processes. |
| 218 | -1.00 | 0.82 | Narrator experiencing acute physical danger or intense bodily sensation. |
| 219 | -1.00 | 0.84 | Vivid physiological responses to life-threatening or catastrophic moments. |
| 225 | -1.00 | 0.78 | Narrator demonstrating competence or mastery in unfamiliar situations. |
| 230 | -1.00 | 0.98 | Vivid descriptions of sudden violent physical impacts or collisions. |
| 241 | -1.00 | 0.82 | Educational or learning activities introduced by a parent or mentor figure. |
| 245 | -1.00 | 0.58 | Acute physical danger or injury with visceral bodily responses. |
| 246 | -1.00 | 0.49 | Narrator making deliberate choices and taking direct personal action. |
| 249 | -1.00 | 0.70 | Onomatopoeia and phonetic sound words anchoring sensory memories |
| 251 | -1.00 | 0.70 | Active confusion or bewilderment about immediate situation unfolding |
| 254 | -1.00 | 0.61 | Unexpected reframing or revelation that recontextualizes preceding narrative. |
| 259 | -1.00 | 0.84 | Detailed descriptions of physical spaces and their structural layouts. |
| 262 | -1.00 | 0.73 | Relationships that fundamentally altered the narrator's personal values or identity. |
| 270 | -1.00 | 0.67 | Medical diagnoses or conditions resulting in death or severe disability. |
| 273 | -1.00 | 0.66 | Named prestigious venues paired with professional achievement or performance moments. |
| 274 | -1.00 | 0.71 | Narrator in acute crisis, distress, or confrontation with figures of authority. |
| 279 | -1.00 | 0.65 | Personal revelation about shifts in narrator's beliefs or self-identity. |
| 291 | -1.00 | 0.43 | Richly detailed sensory descriptions of specific lived moments. |
| 294 | -1.00 | 0.89 | Immediate physical action or violence in response to a moment. |
| 299 | -1.00 | 0.67 | Real-time physical action with immediate danger or high stakes. |
| 310 | -1.00 | 0.81 | Specific dates or precise temporal markers anchoring key moments. |
| 311 | -1.00 | 0.77 | Specific geographic locations and place names establishing narrative setting. |
| 327 | -1.00 | 0.69 | Narrator describing formative moments that fundamentally altered their life direction. |
| 341 | -1.00 | 0.65 | References to specific geographic locations and cultural artifacts visited or researched. |
| 344 | -1.00 | 0.61 | Immersive sensory and emotional experiences vividly felt in the moment. |
| 363 | -1.00 | 0.36 | Creating imaginary worlds or transforming ordinary spaces imaginatively. |
| 365 | -1.00 | 0.83 | Hyperbolic humor and absurdist comparisons about ordinary situations. |
| 366 | -1.00 | 0.63 | Speaker questioning their own moral judgment or character choices. |
| 371 | -1.00 | 0.48 | Narrating pivotal life decisions rejecting societal conventions or expectations. |
| 372 | -1.00 | 0.62 | Intense physical sensations accompanying strong emotions or moments of beauty. |
| 373 | -1.00 | 0.56 | Detailed descriptions of specific physical injuries or bodily damage. |
| 387 | -1.00 | 0.75 | Personal narratives focused on romantic relationships and dating experiences. |
| 390 | -1.00 | 0.63 | Deep reflection on personal identity, mortality, or psychological self. |
| 396 | -1.00 | 0.66 | Returning to past experiences or people after years of separation to reflect. |
| 398 | -1.00 | 0.67 | Multiple family members present or referenced in narrative moments. |
| 402 | -1.00 | 0.75 | Vivid descriptions of physical panic symptoms and bodily fear responses. |
| 405 | -1.00 | 0.64 | Reconciliation or reconnection with someone from the past. |
| 407 | -1.00 | 0.59 | Distinctive, quirky details from narrator's formative childhood experiences. |
| 423 | -1.00 | 0.60 | Humorous reflection on past naive assumptions or theories proven wrong by reality. |
| 428 | -1.00 | 0.76 | Detailed descriptions of confined physical spaces and tangible objects. |
| 434 | -1.00 | 0.90 | Narrator making ethically ambiguous or deceptive decisions with family consequences. |
| 438 | -1.00 | 0.70 | Immersive discovery or confrontation with unfamiliar, transformative experiences. |
| 459 | -1.00 | 0.42 | Intimate observations about people close to the narrator. |
| 461 | -1.00 | 0.68 | Situations involving physical confinement, institutional authority, or danger. |
| 462 | -1.00 | 0.68 | Specific physical details revealing hidden danger, injustice, or contrast. |
| 464 | -1.00 | 0.48 | Vivid specific childhood memories with identifiable people performing actions. |
| 468 | -1.00 | 0.73 | Explicit references to sexual attraction, development, or experiences. |

Table 4: 128 evaluated features for UTS03, Llama-3.1-70B-Instruct (continued).

| Feature | $\Delta^{\mathrm{EV}}$ | Acc. | Explanation |
|---|---|---|---|
| 472 | -1.00 | 0.53 | Multiple interconnected people introduced through tangential conversational digressions. |
| 482 | -1.00 | 0.65 | Stories centered on human suffering, trauma, or interpersonal conflict. |
| 483 | -1.00 | 0.75 | Describing serious situations with high stakes or systemic consequences. |
| 485 | -1.00 | 0.41 | Vivid, unexpected moments creating immediate psychological contrast or confrontation. |
| 497 | -1.00 | 0.43 | Internal reflection on meaning amid difficult personal circumstances. |