Learning Sparse Semantic-Cortical Atoms for Multisubject Naturalistic fMRI Encoding

Anonymous Author(s)
Affiliation
Address
email

Abstract

Naturalistic language encoding models based on large language model features can predict cortical responses well, but their learned representations remain difficult to interpret. Three challenges are central: semantic selectivity is fundamentally non-identifiable in high-dimensional correlated feature spaces, cortical organization remains implicit in voxelwise weight maps, and cross-subject structure is usually addressed only after model fitting. We introduce a multisubject coupled sparse autoencoder that jointly reconstructs language-model features and predicts subject-specific brain responses, learning a sparse basis of semantic-cortical atoms that makes semantic structure, cortical organization, and cross-subject correspondence explicit within a single model. In naturalistic story-listening fMRI, the model retains 88% of matched dense-feature ridge brain prediction performance using 50 atoms and k=5 active atoms per TR. Relative to matched two-stage baselines, joint training yields the strongest overall tradeoff between prediction and interpretability. The learned atoms are semantically coherent and interpretable, recover more reproducible cross-subject brain maps than standard post hoc encoding-model analyses, align with large-scale cortical structure from Yeo 7 and Neurosynth, and transfer better than ridge to held-out subjects in the low-data regime. This framework provides a more interpretable basis for cognitive neuroscience analyses of semantic-to-brain encoding by linking semantic features to corresponding cortical response patterns across subjects.

1 Introduction

Functional mapping in cognitive neuroscience has evolved from localizing coarse cortical selectivity for controlled stimuli to modeling distributed cortical representations evoked by naturalistic stimuli. Compared to traditional fMRI studies built around experimenter-defined contrasts and group-averaged analyses, naturalistic functional mapping requires more complex inferences about semantic structure, cortical organization, and shared structure across subjects. Recently, activations from large language models (LLMs) have been used as features in voxelwise linear encoding models to predict responses to natural-story listening in densely sampled individual subjects [Huth et al., 2016, Toneva and Wehbe, 2019, Caucheteux and King, 2022, Goldstein et al., 2022, Antonello et al., 2023]. Although these models achieve strong predictive performance and recover rich patterns of semantic organization in cortex, they remain difficult to interpret. The following three difficulties are especially important.

(1) **Semantic selectivity is non-identifiable.** Encoding models routinely use LLM activations to represent natural language stimuli. However, LLMs represent more semantic features than they have embedding dimensions, a phenomenon known as superposition [Elhage et al., 2022, Bricken et al., 2023]. Because different combinations of correlated semantic directions can produce identical fitted weights, voxel selectivity cannot be uniquely determined [Zeng and Gallant, 2025].

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

(2) **Cortical organization remains implicit.** Encoding models typically fit an independent readout for each voxel, a small spatial unit of fMRI measurement. However, semantic information in cortex is represented across distributed, overlapping networks rather than isolated voxelwise units [Haxby et al., 2001, Huth et al., 2016, Zhang et al., 2020]. Because the unit of estimation does not match the structure of the underlying representation, larger-scale cortical organization is not learned directly and must instead be inferred after model fitting.

(3) **Cross-subject structure is not learned directly.** Most approaches derive subject-level cortical structure first and address cross-subject alignment afterward through anatomical registration or functional alignment. As a result, cross-subject correspondence is imposed post hoc rather than learned within the model.

Existing approaches have addressed one or two of these issues, but not all three at once [Fyshe et al., 2014, Bilenko and Gallant, 2016, Van Uden et al., 2018]. The missing link is a factorization that is jointly semantic, cortical, and multisubject within a single training regime.

We address this gap with a **multisubject coupled sparse autoencoder** that learns semantic-cortical atoms from LLM activations while jointly reconstructing LLM features and predicting subject-specific brain responses. We use a top-k sparse autoencoder following recent work on LLM interpretability [Elhage et al., 2022, Bricken et al., 2023, Gao et al., 2024]. The encoder maps stimulus features to a sparse latent code (Fig. 1); a semantic decoder reconstructs the original LLM features, and subject-specific brain decoders map the same code to cortical response patterns. This framework requires no hand-labeled categories, no predefined ROIs, and no post hoc functional alignment. More broadly, this approach supports more interpretable cognitive neuroscience analyses of semantic-to-brain encoding by relating semantic features to corresponding cortical response patterns across subjects.

Our contributions are as follows:

- A **multisubject coupled SAE** that jointly learns semantic structure and cortical organization for naturalistic functional mapping in a single training regime. The learned representation retains 88% of matched dense-ridge brain prediction with a 50-atom dictionary and k=5 active atoms per TR.
- **Sparse semantic-cortical atoms** as the unit of analysis. These atoms recover cortical maps that are more reproducible across subjects than standard post hoc analyses of ridge encoding models. Those maps are externally validated by alignment with established large-scale cortical structure from the Yeo 7 resting-state network parcellation [Yeo et al., 2011] and Neurosynth meta-analytic term maps [Yarkoni et al., 2011].
- **Data-efficient adaptation to held-out subjects** using a frozen encoder and only a linear brain-decoder refit. In the low-data regime, this adaptation strategy outperforms dense-feature ridge regression.
- **A clear advantage of joint coupling**: matched two-stage variants recover either semantic structure or cortical structure, but the jointly trained model gives the strongest overall balance of prediction and interpretability.

2 Related work

Encoding models and interpretability. Encoding models predict neural responses from quantitative stimulus features and are a standard framework for studying brain representations. Under controlled experimental paradigms, hand-labeled categories or contrast-based features often allow for a relatively direct interpretation. For naturalistic language, however, those features have largely been replaced by dense representations from large language models, which enable stronger, broader predictions across cortex [Huth et al., 2016, Toneva and Wehbe, 2019, Caucheteux and King, 2022, Goldstein et al., 2022, Antonello et al., 2023].

This gain in predictive accuracy comes with a loss of interpretability. Dense LLM embeddings exhibit superposition: they encode more semantic features than they have dimensions, entangling distinct concepts along correlated directions [Elhage et al., 2022, Bricken et al., 2023]. Because different combinations of these directions can produce identical fitted weights, regression coefficients are non-identifiable and voxel selectivity cannot be uniquely recovered [Zeng and Gallant, 2025].

Existing analyses typically infer voxel selectivity by projecting fitted weights onto semantic directions of interest or by computing cosine similarity with concept vectors. These heuristics, however, inherit the non-identifiability of the underlying dense feature space. Zeng and Gallant [2025] formalize this problem and propose sparse concept atoms as a principled representational basis for neuroscientific interpretation. They develop this idea in lower-dimensional word-embedding spaces. Extending it to contextual LLM features introduces an additional scaling challenge: Gemma Scope, for example, consists of sparse autoencoders for Gemma 2 2B with widths ranging from about 16K to 1M latents [Lieberum et al., 2024]. At that scale, many learned atoms may be dead, polysemantic, redundant, or simply irrelevant for brain prediction, creating an additional feature-selection problem. Our work addresses this gap by learning sparse atoms jointly with brain prediction, so the retained atoms are shaped directly by brain relevance.

Distributed cortical organization. Large-scale cortical organization is often studied either through resting-state functional connectivity or through sparse and dictionary-learning methods applied directly to blood-oxygen-level-dependent (BOLD) signals. Both can recover distributed structure, but share two important limitations. They do not directly link recovered structure to semantic function, and they can be contaminated by spurious correlations driven by shared noise, such as head motion, physiological fluctuations, or scanner drift [Lee et al., 2011, Xie et al., 2017, Thirion et al., 2024].

Post hoc analyses of encoding-model results improve on this by operating on fitted encoding-model weights rather than raw activity alone. In practice, these analyses typically infer cortical structure from fitted encoding-model weights by clustering voxels with similar weight profiles or by aggregating those weights along semantic directions of interest [Huth et al., 2016, Wehbe et al., 2018, Zhang et al., 2020, Meschke et al., 2023]. However, these analyses still treat cortical organization as a post hoc inference problem, and the resulting interpretations remain constrained by the semantic non-identifiability of the underlying dense feature space.

A further limitation, shared by both resting-state-style network recovery and post hoc analyses of encoding-model results, is the hard assignment of one voxel to one network. This creates a strong bias toward one-to-one mappings, even though cortical representations are widely understood to be overlapping and many-to-many: each site can participate in multiple functions, and each function can recruit multiple sites [Haxby et al., 2001, Rigotti et al., 2013, Huth et al., 2016, Zhang et al., 2020].

Cross-subject alignment. Shared structure across subjects is usually addressed only after subject-level analysis. The simplest approach is anatomical normalization, but anatomy alone does not guarantee functional correspondence. Functional alignment methods such as Hyperalignment, the Shared Response Model, and optimal transport estimate shared spaces from subject-level data [Haxby et al., 2011, Chen et al., 2015, Guntupalli et al., 2016, Bazeille et al., 2019, Thual et al., 2022, 2023, Visconti di Oleggio Castello et al., 2025]. These methods improve cross-subject comparability but leave semantic content and cortical organization as separate inference problems, rather than jointly learned structure.

Sparse autoencoders and interpretability. Sparse autoencoders and related dictionary-learning methods are a standard way to decompose dense neural-network activations into sparse latent features that are often more interpretable than individual embedding dimensions [Elhage et al., 2022, Bricken et al., 2023]. Our architecture builds most directly on TopK sparse autoencoders [Gao et al., 2024], and our interpretability evaluation draws on recent LLM-judge protocols for testing whether latent features admit predictive natural-language descriptions on held-out contexts [Bills et al., 2023, Paulo et al., 2024]. Unlike prior SAE work, our coupled training objective requires sparse factors that support not only reconstruction of language-model activations but also multisubject brain prediction.

3 Methods

3.1 Multisubject coupled sparse autoencoder

We model naturalistic semantic encoding across multiple subjects listening to the same narrative stimulus. At each fMRI sampling time point (TR), the model takes a dense language-model feature vector $x_t \in \mathbb{R}^{d_{\text{sem}}}$ and maps it to a sparse latent code $z_t \in \mathbb{R}^K$ with exactly k active atoms. The encoder applies a linear map, ReLU, and TopK selection to produce this code. Two pathways read out



Figure 1: Four example semantic-cortical atoms. Each row shows brain decoder maps for three example subjects together with the strongest activating TRs and surrounding context. The summary panels show the mean fsaverage map, AutoInterp AUC, semantic coherence z, the best-matching Neurosynth term and brain-map correlation, and significant Yeo7 network affiliations. *Food and meals* (a18) strongly activates posterior temporal cortex and lateral prefrontal cortex; *family members* (a23) and *religion and identity* (a10) activate medial prefrontal cortex and temporoparietal regions; and *numerical measurements* (a8) strongly activates right intraparietal sulcus and lateral prefrontal cortex, consistent with prior cognitive neuroscience findings. Because brain decoders are fit independently for each subject, any cross-subject consistency must arise from the common encoder. Yeo7 networks: DMN = Default Mode, FPN = Frontoparietal, DAN = Dorsal Attention, SMN = Somatomotor.

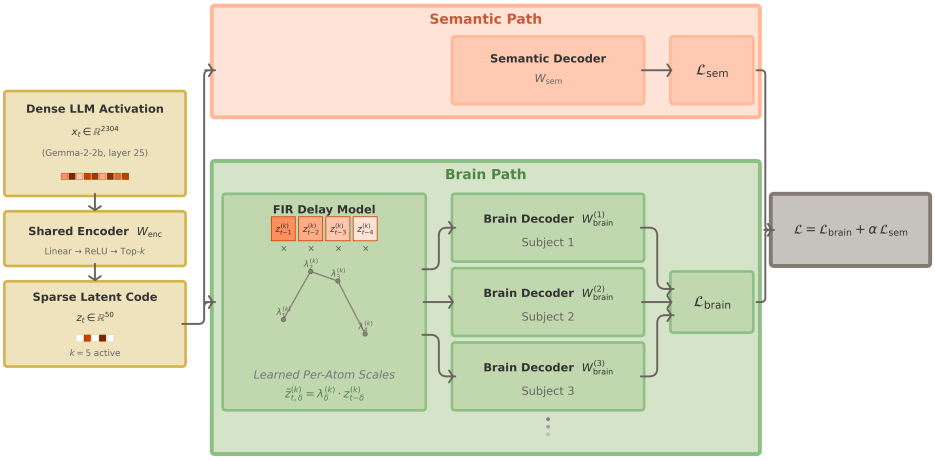


Figure 2: Multisubject coupled sparse autoencoder. A shared encoder maps dense Gemma-2-2b activations $x_t \in \mathbb{R}^{2304}$ to a sparse latent code $z_t \in \mathbb{R}^{50}$ with k=5 active atoms. A semantic decoder reconstructs the original features, while subject-specific brain decoders predict cortical responses from atomwise scaled delayed copies of the same code over 1–4 TR lags, implementing a learned FIR model. The joint objective $\mathcal{L} = \mathcal{L}_{\text{brain}} + \alpha\,\mathcal{L}_{\text{sem}}$ trains the shared encoder for both feature reconstruction and multisubject brain prediction.

from z_t: a semantic decoder reconstructs the original language-model features, and subject-specific brain decoders predict each subject's brain response $\hat{y}_t^{(s)}$ (Fig. 2).

Because neural responses to language are delayed and temporally smoothed by hemodynamics, the brain decoders operate on a finite-impulse-response (FIR) representation of the sparse code rather than on z_t directly. For each delay $\delta \in \{1, 2, 3, 4\}$ TRs, we form a lagged copy of the sparse code and rescale it elementwise with a learned delay-specific weight vector, $\tilde{z}_{t,\delta} = \lambda_\delta \odot z_{t-\delta}$, where $\lambda_\delta \in \mathbb{R}^K$. This allows different atoms to express different temporal profiles. The four delayed codes are then concatenated and passed to a subject-specific linear brain decoder. The matched dense-feature ridge baseline uses the same delay set, making the two models directly comparable:

$$z_t = \mathrm{TopK}\big(\mathrm{ReLU}(W_{\text{enc}}\, x_t + b_{\text{enc}})\big), \tag{1}$$

$$\hat{x}_t = W_{\text{sem}}\, z_t, \tag{2}$$

$$\tilde{z}_{t,\delta} = \lambda_\delta \odot z_{t-\delta}, \quad \delta = 1, \ldots, 4, \tag{3}$$

$$\hat{y}_t^{(s)} = W_{\text{brain}}^{(s)} \big[\tilde{z}_{t,1};\ \tilde{z}_{t,2};\ \tilde{z}_{t,3};\ \tilde{z}_{t,4}\big]. \tag{4}$$

We refer to each latent dimension, together with its semantic decoder direction and per-subject brain-decoder maps, as a *sparse semantic-cortical atom* (see Fig. 1 for examples; additional atoms in Appendix A).

Crucially, the brain decoders receive no anatomical location information. Voxels are treated as an unordered set and the model has no access to spatial coordinates or inter-subject registration. Any cross-subject consistency in the learned cortical maps therefore reflects shared semantic-to-brain structure recovered from the stimulus alone.

The reported model uses the encoder-norm TopK variant, selected from an architecture ablation against matching-pursuit and matryoshka SAEs (Appendix C).

3.2 Joint objective and training

We use 50 atoms with k=5 and set $\alpha = 1.0$, selected from a sweep over atom count, sparsity, and coupling weight (Appendix D). We optimize $\mathcal{L} = \mathcal{L}_{\text{brain}} + \alpha\mathcal{L}_{\text{sem}}$, where both terms are mean-squared error. Each minibatch computes $\mathcal{L}_{\text{brain}}$ only for subjects with data available in that batch. We cycle subjects round-robin across minibatches, a schedule that outperforms computing brain losses for all subjects in every batch (Appendix B).

Training uses AdamW with early stopping on a fixed 4-story validation holdout used for model selection, and models converge in roughly 60 epochs. All reported numbers are 3-seed mean $\pm$ std. Full implementation details are given in Appendix B.

3.3 Data and evaluation

We use all eight subjects in the natural-listening fMRI dataset of LeBel et al. [2023]: UTS01–UTS06 for model training and main evaluation, and UTS07–UTS08 for held-out transfer experiments. Training data are highly uneven across subjects: after excluding the held-out test story, UTS01–UTS03 each contribute 83 stories, UTS04 contributes 25, and UTS05–UTS06 contribute 26 each. Brain responses are z-scored per voxel. To focus modeling on language-responsive voxels, per subject we retain the 30,000 voxels with highest mean leave-one-story-out CV correlation from a Gemma-2-2b ridge regression fit on 10 training stories. Voxel selection uses training data only and is fixed before fitting all compared models. Semantic inputs are Gemma-2-2b activations from layer 25 (2304 dimensions), selected from a layer sweep (Appendix B). These features are then downsampled to the fMRI sampling rate of one TR every 2 s and RMS-normalized to unit scale.

On the held-out test story, we report two primary metrics. The brain correlation score is the mean voxelwise Pearson correlation between predicted and observed responses, averaged over the 30,000 selected voxels and then over subjects. The semantic correlation score is the mean featurewise Pearson correlation between reconstructed and original Gemma-2-2b activations.

4 Results

4.1 Joint coupling preserves brain prediction in a sparse, low-dimensional latent space

To assess whether the learned latent space preserves the information needed for both brain prediction and semantic reconstruction, we evaluate the coupled SAE on the held-out test story shared by all six training subjects (UTS01–UTS06). We compare against two reference models: a matched dense ridge baseline and a random encoder control. Relative to matched ridge, the coupled SAE retains 88% of brain prediction performance (0.162 ± 0.005 compared with 0.185) while reducing the representation from 2304 Gemma-2-2b features to 50 atoms with only $k{=}5$ active per TR. The random encoder yields near-zero brain and semantic correlation scores, indicating that sparsity and dimensionality reduction alone are not sufficient. A per-subject breakdown is given in Appendix L.

To assess the value of joint training, we compare the coupled SAE with two matched two-stage ablations. In Sem$\rightarrow$Brain, the encoder is trained first for semantic reconstruction and the brain decoder is then fit on the frozen latent code. In Brain$\rightarrow$Sem, the encoder is trained first for brain prediction and the semantic decoder is then fit on the frozen latent code. All three models share the same architecture and differ only in whether the encoder is trained jointly or sequentially.

The coupled model gives the best overall balance across brain prediction and interpretability metrics (Table 1). For brain prediction, it outperforms both two-stage baselines (0.162 ± 0.005 compared with 0.131 ± 0.006 for Sem$\rightarrow$Brain and 0.060 ± 0.005 for Brain$\rightarrow$Sem). The especially large gap relative to Brain$\rightarrow$Sem suggests that brain data alone do not sufficiently constrain sparse learning in this setting, and that the semantic objective helps regularize the learned representation. We also compare the models on semantic coherence and AutoInterp AUC, which are defined in the next subsection. Sem$\rightarrow$Brain is slightly stronger on semantic correlation score and semantic coherence (0.539 ± 0.003 compared with 0.523 ± 0.006; coherence 6.1 ± 0.9 compared with 4.9 ± 1.0), consistent with optimizing more directly for the source embedding space. However, the coupled model attains the highest AutoInterp AUC (0.85 ± 0.05, compared with 0.81 ± 0.05 for Sem$\rightarrow$Brain and 0.69 ± 0.04 for Brain$\rightarrow$Sem; Table 6). One possibility is that brain supervision acts as a regularizer, shifting atoms toward representations that are more interpretable to an LLM judge.

4.2 Learned atoms are semantically coherent and interpretable

Prediction and reconstruction performance show that the coupled atoms capture information relevant to both brain responses and semantic content, but they do not by themselves establish that the atoms correspond to coherent, nameable concepts. We therefore evaluate semantic interpretability in two complementary ways.

Table 1: Joint coupling gives the best overall balance of brain prediction and interpretability (3-seed mean ± std; ridge is deterministic, no seed variance). Ridge uses matched dense Gemma-2-2b features. Random is an architecture-matched control with a frozen random encoder. Sem→Brain and Brain→Sem are matched two-stage ablations in which the encoder is trained for one objective before fitting the other decoder.

Method	Brain corr. score	Sem. corr. score	AutoInterp AUC	Coh. z
Ridge (9216-d)	**0.185**	—	—	—
Coupled SAE	$\mathbf{0.162 \pm 0.005}$	0.523 ± 0.006	$\mathbf{0.85 \pm 0.05}$	4.9 ± 1.0
Random encoder	0.064 ± 0.005	0.069 ± 0.002	—	—
Sem→Brain	0.131 ± 0.006	$\mathbf{0.539 \pm 0.003}$	0.81 ± 0.05	$\mathbf{6.1 \pm 0.9}$
Brain→Sem	0.060 ± 0.005	0.367 ± 0.004	0.69 ± 0.04	4.0 ± 0.1

Semantic coherence measures how tightly an atom's top-activating text segments cluster in the source LLM embedding space relative to a random-segment null. Across 50 atoms, semantic coherence is $z = 4.9 \pm 1.0$, indicating greater semantic homogeneity than expected by chance ($z = 0$).

AutoInterp [Bills et al., 2023] evaluates whether an atom admits a predictive natural-language description. An LLM is given an atom's top-activating excerpts and asked to produce a one-sentence description, label the atom as `mono`, `poly`, or `dead`, and score held-out excerpts for how well they match the description. AUC then measures how well those scores distinguish activating from non-activating examples. The LLM also receives structural metadata such as story position, syllable rate, and vocabulary statistics, allowing AutoInterp to also capture atoms driven by non-semantic textual or temporal regularities.

Averaged across all 50 atoms, AutoInterp AUC is 0.85 ± 0.05 (median per atom, mean ± std across 3 seeds), with $87\% \pm 8\%$ of atoms exceeding AUC 0.7 and no dead atoms (0/50 across all 3 seeds). The recovered descriptions span lexical domains such as *"family members"* (a23) and *"geographic landscape"* (a45), as well as non-semantic regularities such as *"story-opening narration"* (a5) and *"silence-firing"* atoms (a30). The full 50-atom catalog with all metrics is given in Table 5 (Appendix E).

4.3 Joint learning yields reproducible brain maps and structured cortical organization

We ask whether each atom recovers consistent cortical structure across subjects, despite the brain decoders being fit independently per subject with no explicit anatomical correspondence constraint. We quantify cross-subject consistency by correlating per-atom brain maps after projection to fsaverage (Appendix J). Averaged across subject pairs and atoms, the mean correlation is $r = 0.195 \pm 0.010$ ($p < 0.0001$, permutation null $r = 0.004 \pm 0.004$). These results indicate that the learned atoms recover reproducible cortical maps for a given semantic feature.

We next compare this result with a common post hoc approach to encoding-model interpretation: projecting ridge encoding-model weights onto semantic directions of interest [Huth et al., 2016, Wehbe et al., 2018, Zeng and Gallant, 2025]. SAE atom maps are significantly more consistent across subjects than this baseline (SAE $r = 0.195 \pm 0.010$, ridge $r = 0.174 \pm 0.006$; $p = 0.0005$). This difference also appears at the level of individual atoms under the same shuffled-atom permutation null: 43/50 SAE atoms survive BH-FDR correction, compared with 19/50 ridge-projection maps. Because both methods use the same semantic directions, this difference is not attributable to semantic geometry and instead reflects the advantage of joint brain-semantic learning.

We next ask whether this shared cortical structure aligns with established descriptions of large-scale brain organization. We evaluate the learned atoms against two complementary external references. Because these references come from different paradigms and analysis pipelines, we do not expect exact alignment with our naturalistic story-listening setting, but we do expect them to provide informative external validation.

Yeo7 [Yeo et al., 2011] is a seven-network parcellation derived from resting-state functional connectivity. For each atom, we compute the cosine similarity between its mean fsaverage decoder map and each network's binary indicator vector, and assess significance with a within-mask spatial spin null (BH-FDR $q < 0.05$ across all 7 network × 50 atom pairs). Significant affiliations are concentrated in the Default Mode Network (23/50 atoms) and Frontoparietal Network (18/50 atoms). Using

AutoInterp descriptions as shorthand labels, Default-associated atoms include *"family members and relatives"* (a23), *"peer-community identity"* (a28), *"confrontational dialogue/threats"* (a46), and the *"silence-firing"* atom (a30). Frontoparietal-associated atoms include *"numerical measurements"* (a8), *"letters and written notes"* (a20), and *"geographic landscape descriptions"* (a45). Dorsal Attention includes *"physical fighting"* (a24), and Somatomotor includes *"story-opening narration"* (a5) and *"long elaborate quoted speech"* (a48). Visual and Limbic show no significant affiliations, as expected in a language-listening task. All atoms mentioned here are shown in Appendix A; per-atom affiliations are listed in Table 5, and methodological details are given in Appendix K.

Neurosynth [Yarkoni et al., 2011] aggregates approximately 14,000 published fMRI studies and provides automated meta-analytic term maps. From its vocabulary, we curated 152 cognitive, perceptual, emotional, and behavioral terms (Appendix H). For each atom, we use sentence-transformer embeddings of its AutoInterp description to select the most semantically similar Neurosynth term, then compare the atom's brain map with that term's meta-analytic map. Because the atom labels come from naturalistic narrative stimuli whereas the Neurosynth maps aggregate heterogeneous task paradigms, we treat these matches as external validation rather than as exact anatomical correspondences. We test significance with a within-mask spatial spin null; 19 of 50 atoms survive BH-FDR correction ($q < 0.05$ across 50 atoms). Significant matches include *"story-opening narration"* (a5) with "speech" ($r = 0.65$), *"family members and relatives"* (a23) with "personal" ($r = 0.36$), *"peer-community identity"* (a28) with "social" ($r = 0.28$), and *"geographic landscape descriptions"* (a45) with "spatial" ($r = 0.25$). Per-atom results are listed in Table 5, and methodological details are given in Appendix H.

Together, these results indicate that the learned dictionary has anatomically structured cortical organization at both the network level and the level of individual atoms.

4.4 Learned semantic-cortical representations support data-efficient adaptation to new subjects

One practical promise of a multisubject encoding model is data-efficient adaptation to new individuals. This is especially important for translational applications, where the multi-hour naturalistic listening protocols common in research settings are often unrealistic [Gordon et al., 2017, Gratton et al., 2020]. The coupled SAE is well suited to this setting because it learns a compact cross-subject representation, so adaptation to a new subject requires fitting only a subject-specific linear decoder.

We evaluate this setting in Appendix I using two held-out subjects (UTS07, UTS08). For each subject, we fit only the linear brain decoder using between 0.5 and 22 stories and test on a held-out story. We compare against matched dense-feature ridge regression on full Gemma-2 embeddings and a random encoder control.

The coupled SAE outperforms both baselines in the low-data regime, with the clearest advantage below 2 stories (approximately 24 minutes of scan time; one-sided paired t-test, $*p < 0.05$ at $n = 0.5$, $\dagger p < 0.1$ at $n = 1$). The random encoder remains substantially worse than both the coupled SAE and ridge at all values of n, indicating that dimensionality reduction and sparsity alone do not explain the gain. At larger sample sizes ($n \geq 8$ stories), dense-feature ridge overtakes the SAE, suggesting that the full high-dimensional feature space becomes advantageous once more subject-specific data is available. At $n = 22$ stories, the coupled SAE reaches $r = 0.100 \pm 0.008$ on UTS07 and $r = 0.086 \pm 0.004$ on UTS08, recovering 75% and 92% of matched-ridge performance, respectively.

These results suggest that the learned representation supports efficient adaptation to new subjects. If this advantage scales with pre-training cohort size, larger training cohorts could further reduce the scan time required for subject-specific adaptation and improve the practical feasibility of personalized encoding models in translational settings.

5 Discussion

This work introduces a sparse latent basis for semantic-to-brain encoding that makes semantic-cortical structure explicit and analyzable across subjects. Rather than improving predictive benchmarks, the primary goal is to provide a more principled unit of analysis for cognitive neuroscience. Recovering

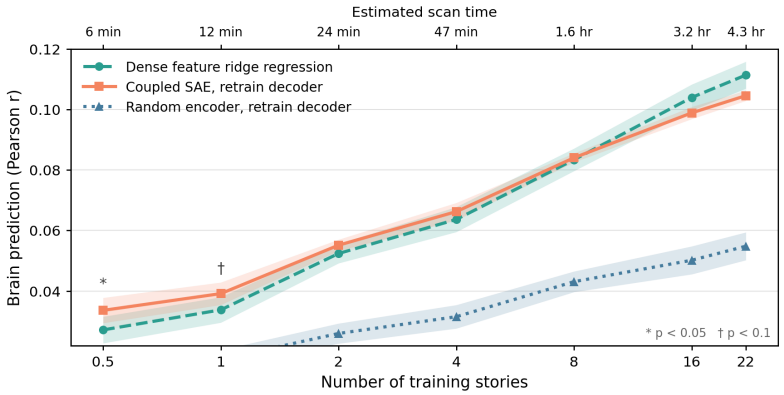


Figure 3: Learned semantic-cortical representations support data-efficient adaptation to new subjects. Brain correlation score is plotted against the number of training stories available for held-out subjects UTS07 and UTS08 (mean $\pm$ s.e. over six random story draws per subject, averaged across UTS07 and UTS08). The coupled SAE (green) provides the strongest inductive prior in the low-data regime: at $n = 0.5$–2 stories ($\sim$6–24 minutes of scan time) it significantly outperforms dense-feature ridge regression on full Gemma-2-2b embeddings ($*p < 0.05$, $\dagger p < 0.1$, one-sided paired t-test). Ridge overtakes the SAE at $n \geq 8$ stories, where the full high-dimensional feature space becomes advantageous. A random encoder control (gray) remains near floor, indicating that the gain comes from the learned representation rather than from dimensionality reduction alone.

semantic-cortical atoms directly within the encoding model may help clarify aspects of distributed cortical coding that are difficult to isolate with voxelwise or post hoc analysis approaches.

The learned atoms should nonetheless be understood as useful descriptive features rather than a canonical or mechanistic decomposition. Across random seeds, the coupled SAE shows moderate atom-level stability after Hungarian matching but high representational similarity by CKA (Centered Kernel Alignment; Kornblith et al. 2019), indicating that global latent structure is more reproducible than any particular set of atoms (Appendix M). Architecture ablations show that this gap varies across model variants, suggesting that atom-level stability is an important design target for future work.

This paper establishes the framework and evaluates the learned representation in a setting where semantic stimuli are relatively well matched across subjects. Other naturalistic neuroimaging datasets are available, but they typically involve additional heterogeneity, including visual stimulus content, uneven subject overlap, different stories, and varying per-subject depth. The present framework is well suited to such datasets, but addressing that heterogeneity is outside the scope of the current paper and a natural direction for future work.

A practical implication of this work is that the learned representation may serve as a reusable prior for new-subject adaptation. This is especially relevant in translational settings where long naturalistic scanning sessions are impractical. More broadly, more accurate and data-efficient semantic-to-brain models raise privacy considerations: better inference of semantic content from neural data could make sensitive information easier to extract from neural recordings, particularly outside carefully governed research or clinical settings.

References

Aaron F. Alexander-Bloch, Haochang Shou, Siyuan Liu, Theodore D. Satterthwaite, David C. Glahn, Russell T. Shinohara, Simon N. Vandekar, and Armin Raznahan. On testing for spatial correspondence between maps of human brain structure and function. *NeuroImage*, 178:540–551, 2018. doi: 10.1016/j.neuroimage.2018.05.070.

Richard J. Antonello, Aditya R. Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fMRI. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Thomas Bazeille, Hugo Richard, Hicham Janati, and Bertrand Thirion. Local optimal transport for functional brain template estimation. In *Information Processing in Medical Imaging*, 2019. doi:

10.1007/978-3-030-20351-1_18.

Natalia Y. Bilenko and Jack L. Gallant. Pyrcca: Regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, 10:49, 2016. doi: 10.3389/fninf.2016.00049.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. OpenAI Blog, 2023. URL `https://openai.com/index/language-models-can-explain-neurons-in-language-models/`.

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. URL `https://transformer-circuits.pub/2023/monosemantic-features/`.

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

Po-Hsuan Cameron Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James V. Haxby, and Peter J. Ramadge. A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/toy_model/index.html`.

Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

Alona Fyshe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. Interpretable semantic vectors from a joint model of brain- and text-based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 489–499, 2014. doi: 10.3115/v1/P14-1046.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

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

Evan M. Gordon, Timothy O. Laumann, Adrian W. Gilmore, Dillan J. Newbold, Deanna J. Greene, Jeffrey J. Berg, Mario Ortega, Catherine Hoyt-Drazen, Caterina Gratton, Haoxin Sun, Jacqueline M. Hampton, Rebecca S. Coalson, Annie L. Nguyen, Kathleen B. McDermott, Joshua S. Shimony, Abraham Z. Snyder, Bradley L. Schlaggar, Steven E. Petersen, Steven M. Nelson, and Nico U. F. Dosenbach. Precision functional mapping of individual human brains. *Neuron*, 95(4):791–807, 2017. doi: 10.1016/j.neuron.2017.07.011.

Caterina Gratton, Brian T. Kraus, Deanna J. Greene, Evan M. Gordon, Timothy O. Laumann, Steven M. Nelson, Nico U. F. Dosenbach, and Steven E. Petersen. Defining individual-specific functional neuroanatomy for precision psychiatry. *Biological Psychiatry*, 88(1):28–39, 2020. doi: 10.1016/j.biopsych.2019.10.026.

J. Swaroop Guntupalli, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 26(6):2919–2934, 2016. doi: 10.1093/cercor/bhw068.

James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. doi: 10.1126/science.1063736.

James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011. doi: 10.1016/j.neuron.2011.08.026.

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: 453–458, 2016. doi: 10.1038/nature17637.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

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*, 10:555, 2023. doi: 10.1038/s41597-023-02437-z.

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.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Emily X. Meschke, Matteo Visconti di Oleggio Castello, Tom Dupré la Tour, and Jack L. Gallant. Model connectivity: leveraging the power of encoding models to overcome the limitations of functional connectivity. *bioRxiv*, 2023. doi: 10.1101/2023.07.17.549356.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):1–12, 2010. doi: 10.2202/1544-6115.1585.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497: 585–590, 2013. doi: 10.1038/nature12160.

Bertrand Thirion, Himanshu Aggarwal, Ana Fernanda Ponce, Ana Luísa Pinho, and Alexis Thual. Should one go for individual- or group-level brain parcellations? A deep-phenotyping benchmark. *Brain Structure and Function*, 229:161–181, 2024. doi: 10.1007/s00429-023-02723-x.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced Gromov-Wasserstein. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alexis Thual, Yohann Benchetrit, Felix Geilert, Jérémy Rapin, Iurii Makarov, Hubert Banville, and Jean-Rémi King. Aligning brain functions boosts the decoding of visual semantics in novel subjects. *arXiv preprint arXiv:2312.06467*, 2023.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Cara E. Van Uden, Samuel A. Nastase, Andrew C. Connolly, Ma Feilong, Isla Hansen, M. Ida Gobbini, and James V. Haxby. Modeling semantic encoding in a common neural representational space. *Frontiers in Neuroscience*, 12:437, 2018. doi: 10.3389/fnins.2018.00437.

Matteo Visconti di Oleggio Castello, Tom Dupré la Tour, and Jack L. Gallant. Individual differences shape conceptual representation in the brain. *bioRxiv*, 2025. doi: 10.1101/2025.08.22.671848.

Leila Wehbe, Alexander G. Huth, Fatma Deniz, James S. Gao, Marie-Luise Kieseler, and Jack L. Gallant. BOLD predictions: automated simulation of fMRI experiments. In *Conference on Cognitive Computational Neuroscience*, 2018.

Jianwen Xie, Pamela K. Douglas, Ying Nian Wu, Arthur L. Brody, and Ariana E. Anderson. Decoding the encoding of functional brain networks: An fMRI classification comparison of non-negative matrix factorization (NMF), independent component analysis (ICA), and sparse coding algorithms. *Journal of Neuroscience Methods*, 282:81–94, 2017. doi: 10.1016/j.jneumeth.2017.03.008.

Tal Yarkoni, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8: 665–670, 2011. doi: 10.1038/nmeth.1635.

B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. doi: 10.1152/jn.00338.2011.

Alicia Zeng and Jack L. Gallant. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, 11:1877, 2020. doi: 10.1038/s41467-020-15804-w.

A Additional atom examples

Fig. 1 in the main text shows four example atoms. Here we show all remaining atoms referenced in the main text, presented in the same format: per-subject brain decoder maps (UTS02, UTS03, UTS06), top-activating text excerpts, and a summary panel with the mean fsaverage map and interpretability metrics. The full 50-atom catalog with all metrics is in Table 5.

Of the 50 learned atoms, 46 have all-positive hemodynamic delay scales, meaning their activation predicts an *increase* in brain response. Three atoms have weakly negative mean delay scales: a44 (reported quoted dialogue, mean $= -0.046$) and a21 (rapid back-and-forth dialogue, mean $= -0.021$), both near zero. One atom is strongly suppressive: **a30** (silence-firing, mean $= -0.980$), whose delay scales are all negative across all four lags (2–8 s), predicting a decrease in language-responsive cortical activity during silent TRs.

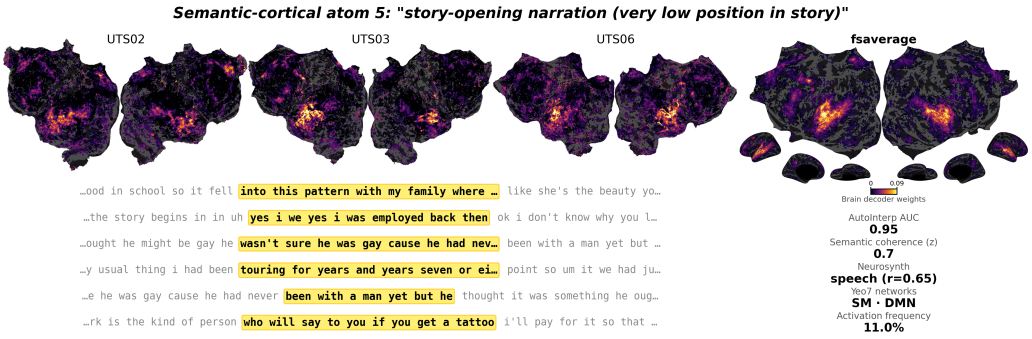


Figure 4: Atom 5: *"story-opening narration"* (a5) (very low position in story).

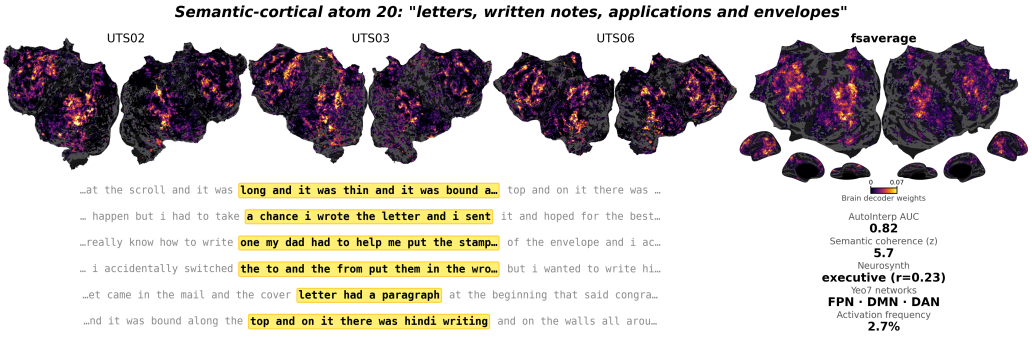


Figure 5: Atom 20: *"letters, written notes, applications and envelopes"* (a20).

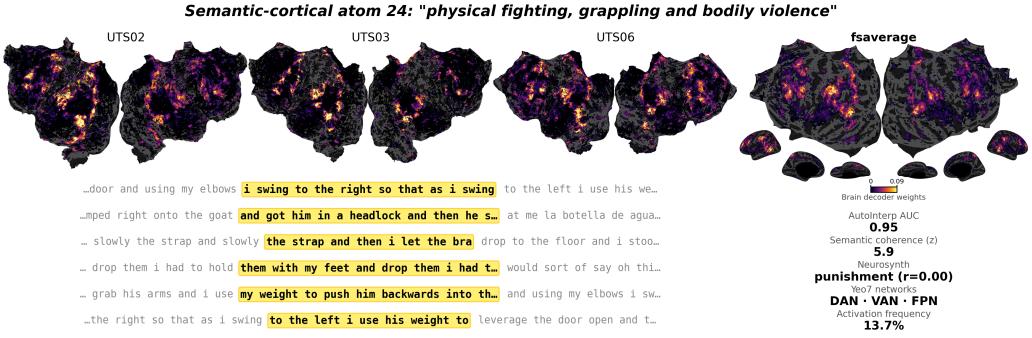


Figure 6: Atom 24: *"physical fighting, grappling and bodily violence"* (a24).



Figure 7: Atom 28: *"groups, friends or peer-community identity"* (a28).

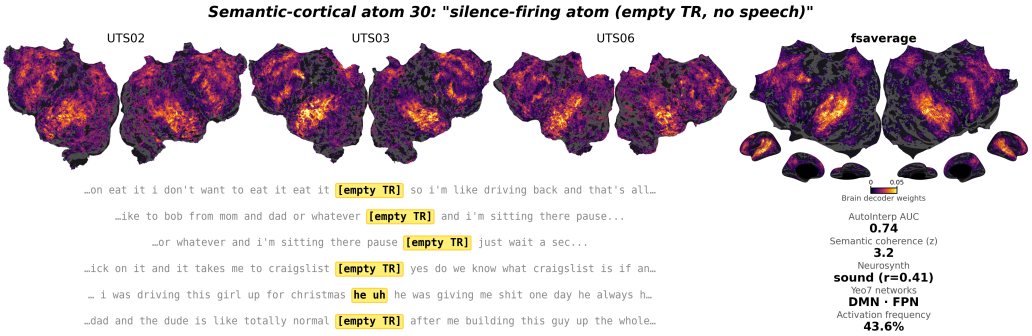


Figure 8: Atom 30: *"silence-firing"* (a30) (empty TR, no speech). This is the only atom with all-negative delay scales (mean $= -0.980$), predicting suppression of language-responsive cortex during silent TRs.

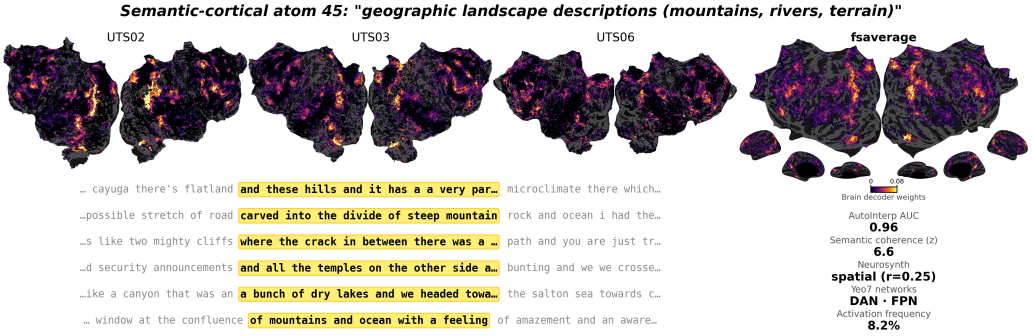


Figure 9: Atom 45: *"geographic landscape descriptions"* (a45) (mountains, rivers, terrain).

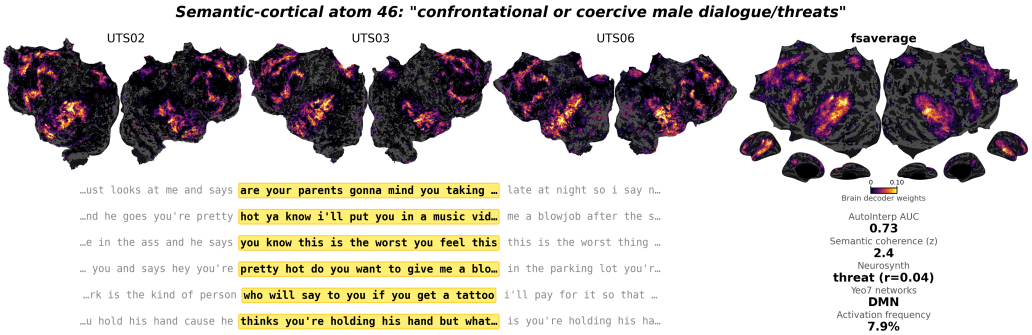


Figure 10: Atom 46: *"confrontational or coercive male dialogue/threats"* (a46).

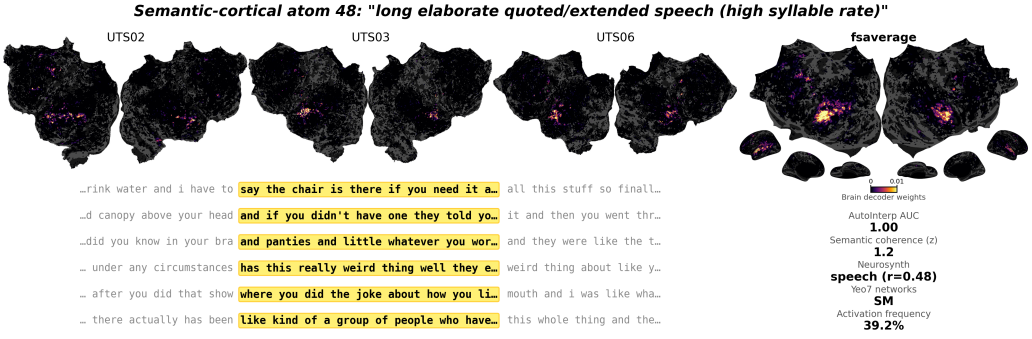


Figure 11: Atom 48: *"long elaborate quoted/extended speech"* (a48) (high syllable rate).

B Implementation details

This appendix provides full implementation details for the methods described in §3, including architectural design rationales, preprocessing specifications, hyperparameter settings, and reproducibility statistics.

B.1 Architectural design choices

TopK over L1. We use hard top-k selection [Gao et al., 2024] rather than L_1-penalized ReLU sparsity [Bricken et al., 2023]. Top-k enforces exactly k active atoms per TR, making sparsity explicit and easier to control.

Pre-encoder bias. The pre-encoder bias b_{pre} is initialized to the geometric median of the training features and then trained as a free parameter. It is subtracted before encoding and added back after decoding, so the sparse code represents deviations from a typical input rather than the full input itself.

Untied semantic decoder. The semantic decoder is initialized as the transpose of the encoder and then trained independently. Following Gao et al. [2024], this allows reconstruction residuals to be absorbed by the decoder without distorting the encoder directions.

Brain-decoder unit-norm with gradient projection. Brain-decoder columns are projected to unit norm after every optimizer step. We also remove the radial gradient component before each step [Bricken et al., 2023], so optimization does not push parameters in directions that the norm projection would immediately discard.

No per-voxel scale in the best model. The architecture supports an optional learnable per-voxel scale on the brain decoder, but the best model omits it. Ablation results (Appendix C) show that removing this term improves mean brain correlation score and reduces variance across seeds.

AuxK dead-feature recovery. Atoms that have not fired in any of the last 50 batches contribute to an auxiliary loss that reconstructs the LLM-feature residual through the dead atoms only, with coefficient $1/32$ [Gao et al., 2024]. In the final 50-atom configuration, no atoms remain dead at convergence.

Hemodynamic response. We model the BOLD response with a finite-impulse response over four delays at 2, 4, 6, and 8 seconds (TR = 2 s), using a learned per-atom scale at each delay. This lets different atoms express different HRF shapes.

B.2 Semantic feature extraction

Backbone. Gemma-2-2b (26 transformer layers, hidden size 2304). For each story we tokenize the transcript and run the model with sliding-window context of up to 8192 tokens, so activations at a given word can attend to long-range narrative context.

Token-to-TR alignment. Subword tokens are average-pooled to word level using the dataset's word boundaries, and word-level activations are Lanczos-interpolated to the TR rate (TR = 2 s). We also demean each story so that story-level offsets do not get absorbed into the encoder bias.

Layer choice. Early hyperparameter exploration compared layers 15 ($\sim$58% depth), 20 ($\sim$77%), and 25 (last). We selected layer 25 because it gave the highest validation brain correlation score in our setup (Table 4), consistent with prior reports for similarly-sized LMs [Caucheteux and King, 2022].

B.3 Brain data and preprocessing

Subjects and stimuli. Eight subjects from the natural-listening fMRI dataset of LeBel et al. [2023]. UTS01–UTS06 are used for training and evaluation in the main model, while UTS07–UTS08 are used only in the transfer experiments of §4.4. Training-data availability is highly uneven across subjects: after excluding the held-out test story, UTS01–UTS03 each contribute 83 stories, UTS04

contributes 25, and UTS05–UTS08 contribute 26 each. The held-out test story was listened to by all eight subjects and is stored in the released preprocessed data as the average across 10 repeated presentations. Brain responses are taken from the dataset's preprocessed release and z-scored per voxel in our pipeline.

Voxel selection. For each subject we rank voxels by their mean leave-one-story-out CV correlation in a Gemma-2-2b ridge regression fit on 10 training stories [Dupré La Tour et al., 2022], and retain the top 30,000. The held-out test story is not seen at any point during selection.

Voxel-to-fsaverage mapper. For each subject, we precompute a sparse matrix mapping the full brain-mask voxel space to the fsaverage surface (327,684 vertices) using FreeSurfer [Fischl, 2012] `mri_surf2surf`. This mapper is used by the Yeo 7 (Appendix K), Neurosynth (Appendix H), and cross-subject consistency (Appendix J) analyses. The mapper is independent of the SAE model and is built once per subject.

Cross-subject transfer (UTS07, UTS08). For the cross-subject experiments (Fig. 3), the encoder trained on UTS01–UTS06 is frozen and only the linear brain decoder is re-trained for each held-out subject. UTS07 and UTS08 each contribute 27 stories total (26 excluding test), yielding 22 effective training stories after the same 4-story validation holdout used for model selection.

B.4 Training and optimization

Optimizer. AdamW [Loshchilov and Hutter, 2019] with learning rate 10^{-3}, cosine-annealing schedule, 5-epoch linear warmup, gradient clipping at $\|\nabla\|_2 \leq 1$, and batch size 1024. Encoder rows are projected to unit norm at the end of each epoch; brain-decoder columns are projected to unit norm after every step, and the radial gradient component is zeroed before each step (gradient projection; §B.1).

Coupling weight. $\alpha = 1.0$ was selected over $\{0.3, 1, 3, 10\}$ by validation brain correlation score. Appendix D reports the full sweep over atom count, sparsity, and coupling weight.

Multi-subject scheme. For each minibatch, $\mathcal{L}_{\text{brain}}$ is computed for a single subject on the TRs that subject has data for. The subject mask generalises to the all-subjects-per-batch setting as well, handling missing-TR cases caused by the heterogeneous story sets in §B.3. The best model uses round-robin cycling (UTS01 $\rightarrow$ UTS02 $\rightarrow \cdots \rightarrow$ UTS06 $\rightarrow$ UTS01 $\rightarrow \cdots$) across minibatches, which outperforms the all-subjects-per-batch alternative (Table 2). $\mathcal{L}_{\text{sem}}$ is computed on every TR of every minibatch.

Validation and early stopping. Validation uses a fixed story-level holdout of 4 training stories (random seed 0) for model selection. Story-level, rather than TR-level, splitting ensures that no validation TR shares its 4-lag input window with any training TR. In the per-subject effective-count summaries reported in the main text, this yields 79 training stories for UTS01–UTS03, 21 for UTS04, and 22 for UTS05–UTS06 after excluding the held-out test story. UTS07–UTS08 also have 22 stories available, used in the transfer experiment (§4.4). Early stopping triggers on validation total loss with patience 10, and models converge at epochs 57–67.

Reproducibility. All main-text numbers are reported as 3-seed mean $\pm$ std. Aggregate performance is stable across seeds: for the best model, brain correlation score = 0.162 ± 0.005 and the semantic correlation score = 0.523 ± 0.006. Atom-level stability is weaker: after Hungarian matching, cross-seed encoder-row cosine similarity is only moderate (~ 0.4), indicating that similar global solutions can be realized by different atom-level decompositions (Table 8).

B.5 Software and compute

The model and analyses are implemented in PyTorch. In our environment, training a single 50-atom $k = 5$ model on 6 subjects to convergence takes approximately 20 minutes per run on a single RTX A6000-class GPU (48 GB memory). Voxel selection (`himalaya.KernelRidgeCV` [Dupré La Tour et al., 2022]) and most downstream interpretability analyses (Yeo 7 mapping, semantic coherence,

AutoInterp prompt preparation) run on CPU; AutoInterp itself dispatches a single Sonnet judge per model in ≈60 seconds (Appendix F).

We will release anonymized code, configuration files, and scripts to reproduce the main experiments together with the submission. These materials will include instructions for data access, preprocessing, training, evaluation, and figure generation.

The natural-listening fMRI dataset of LeBel et al. [2023] is accessed through OpenNeuro and distributed under CC0. Gemma-2-2b is used under the Gemma terms of use. FreeSurfer requires a free license key under its license terms, and `himalaya` is distributed under the BSD-3-Clause license.

C Architecture ablation

We compare six architectural variants, all trained with the same hyperparameters (50 atoms, $k = 5$, $\alpha = 1.0$, 6 subjects, Gemma-2-2b layer 25). Each variant is trained with 3 random seeds; we report mean $\pm$ std on the held-out test story.

Table 2: Architecture ablation (3-seed mean $\pm$ std). Encoder-row unit-norm TopK gives the strongest brain correlation score with the lowest seed variance among the tested variants.

Variant	Brain corr. score	Sem. corr. score
A. Baseline (no encoder norm)	0.145 ± 0.012	0.519 ± 0.001
B. + Encoder norm (best)	$\mathbf{0.162 \pm 0.005}$	0.523 ± 0.006
C. + Matryoshka $[2, 5]$	0.158 ± 0.004	0.530 ± 0.002
D. Matching-pursuit SAE	0.137 ± 0.004	$\mathbf{0.535 \pm 0.000}$
E. + Per-voxel scale	0.156 ± 0.002	0.516 ± 0.002
F. All-subjects per batch (vs. round-robin)	0.127 ± 0.002	0.493 ± 0.005

Encoder-row unit-norm projection (variant B) gives the strongest brain correlation score with the lowest seed variance. Matching-pursuit SAE (variant D) achieves the highest semantic correlation score, but at a substantial cost in brain prediction. Removing the round-robin subject schedule also degrades both metrics, supporting the use of joint multisubject training with alternating subject updates.

D Atom count, sparsity, and coupling-weight sweep

We sweep atom count $K \in \{25, 50, 100\}$, top-k sparsity $k \in \{2, 5\}$, and coupling weight $\alpha \in \{0.3, 1, 3, 10\}$, holding all other settings at the best architecture. Each configuration is a single seed; the 50-atom k=5 α=1 entry (†) is the 3-seed production model.

Table 3: Hyperparameter sweep over atom count, sparsity, and coupling weight, reported as brain correlation score / semantic correlation score on the held-out test story.

Config	$\alpha = 0.3$	$\alpha = 1.0$	$\alpha = 3.0$	$\alpha = 10$
25 atoms, k=5	.159 / .489	.144 / .502	.108 / .503	.130 / .513
50 atoms, k=5	.149 / .504	**.166**† / .518†	.139 / .533	.120 / .537
50 atoms, k=2	.119 / .451	.116 / .461	.105 / .438	.089 / .427
100 atoms, k=5	.140 / .521	.163 / .543	.145 / .554	.118 / **.557**

The selected configuration (50 atoms, k=5, α=1) provides a good balance between brain correlation score and semantic correlation score. In the single k=2 sweep we ran (50 atoms), performance is worse on both metrics than the corresponding k=5 model, suggesting that two active atoms per TR provide insufficient capacity in this setting. The preferred coupling weight also depends on dictionary size: 25 atoms favor α=0.3, whereas 50 and 100 atoms favor α=1.0.

Semantic correlation score increases monotonically with layer depth, while brain correlation score is highest at the last layer. The last layer (L25) outperforms L15 by $+0.016$ brain correlation score and $+0.143$ semantic correlation score, consistent with prior reports that later layers of language models are more predictive of cortical responses to naturalistic speech [Caucheteux and King, 2022].

Table 4: Gemma-2-2b layer comparison (50 atoms, k=5, α=1.0, round-robin; 3-seed mean $\pm$ std). Later layers improve semantic reconstruction, and the last layer gives the strongest brain correlation score.

Layer	Brain corr. score	Sem. corr. score
Layer 15 ($\sim$58% depth)	0.146 ± 0.003	0.380 ± 0.004
Layer 20 ($\sim$77% depth)	0.143 ± 0.004	0.424 ± 0.002
Layer 25 (last; best)	$\mathbf{0.162 \pm 0.005}$	$\mathbf{0.523 \pm 0.006}$

E Full 50-atom catalog

Table 5 lists all 50 atoms from the best model (encoder norm, no per-voxel scale, seed 0). Columns: AutoInterp description, AUC, semantic coherence z-score, significant Yeo7 networks (BH-FDR $q < 0.05$), best-matching Neurosynth term, and Pearson r and BH-FDR q for the brain-map match. Median AUC = 0.90, median coherence z = 3.5. Of 50 atoms, 45 support coherent descriptions, 5 are polysemantic, and 0 are dead. Significant Neurosynth matches ($q < 0.05$) are marked with $*$.

Table 5: Full 50-atom catalog for the seed 0 coupled model. NS = matched Neurosynth term; D = Default, FP = Frontoparietal, DA = Dorsal Attention, VA = Ventral Attention, SM = Somatomotor.

#	Description	AUC	z	Yeo7	NS term	r	q
0	eye contact, looking at each other	.99	4.0	D	face	+.085	.277
1	child's school/teacher concerns	.95	7.2	D	reappraisal	+.091	.179
2	clothing, outfits, hair descriptions	.97	7.4	DA, FP	color	+.318	.006*
3	mixed scenes (poly)	.71	2.5	VA	imagery	+.019	.579
4	dogs, puppies, baby animals	.99	9.8	FP, D	object	−.036	.922
5	story-opening narration	.95	0.7	SM	speech	+.646	.006*
6	named cities/towns/places	.94	3.9	D	place	+.166	.006*
7	loud crowd action: drumming, dancing	.99	5.3	D	sound	+.099	.101
8	numerical measurements	.98	1.3	FP	numerical	+.286	.006*
9	cameras, film, photography	.82	8.9	FP	imagery	+.021	.755
10	religion, faith, cultural identity	.53	3.1	D	identity	+.129	.128
11	male character physical actions	1.00	5.1	D	movement	−.140	1.00
12	reflective emotional summing-up	.78	3.6	FP	emotional	+.017	.603
13	tasks, training, duties	.82	2.4	D	conditioning	−.137	1.00
14	emotionally intense moments	.81	2.2	—	emotional	+.006	.704
15	close relationships, bonding	.96	7.6	DA, FP	communic.	−.069	1.00
16	searching, wandering	.83	3.6	FP	planning	+.103	.101
17	therapy, counseling settings	.95	8.7	—	communic.	+.305	.006*
18	food and meals	1.00	6.4	DA, FP	food	+.094	.158
19	Oregon Trail / wagon-train journey	.68	5.5	D	navigation	−.120	1.00
20	letters, written notes, envelopes	.82	5.7	DA, FP	executive	+.235	.006*
21	rapid back-and-forth dialogue	.91	4.4	D	speech	+.123	.006*
22	traveling: driving, taxis	.99	2.2	FP, D	navigation	+.172	.065
23	family members and relatives	.87	2.4	D	personal	+.358	.006*
24	physical fighting, violence	.95	5.9	DA, VA	punishment	+.004	.706
25	intense emotional reaction	.94	3.2	D	emotional	+.074	.329
26	parachute survival training	1.00	20.3	DA, FP	planning	+.092	.187
27	rooms, architectural spaces	1.00	2.8	DA	orienting	+.113	.103
28	peer-community identity (poly)	.85	2.8	D	social	+.277	.006*
29	mixed instructions/objects (poly)	.55	0.0	SM	syntactic	+.136	.065
30	silence-firing (empty TR)	.74	3.2	D	sound	+.409	.006*
31	"I was like" reported quotation	.77	3.3	D	autobio.	+.166	.028*
32	medical procedures, surgery	.96	6.1	FP, D	orthographic	+.116	.179
33	mixed proper-noun phrases (poly)	.60	0.7	D	nouns	+.374	.006*
34	career, life-stage transitions	1.00	7.4	—	motivation	−.003	.782
35	rituals, ceremonies, religious practices	1.00	10.5	FP, D	consciousness	+.060	.233
36	narrative time-marker phrases	.76	0.7	FP	episodic	+.161	.016*
37	urgent fast physical action	.92	5.8	VA, FP	action	+.022	.476

Continued on next page

#	Description	AUC	z	Yeo7	NS term	r	q
38	foreign words, brand names (poly)	.75	2.6	—	nouns	+.378	.006*
39	personality/character descriptions	.87	3.0	—	autobio.	+.058	.312
40	depression, distress, illness	.94	3.4	FP	stress	−.012	.795
41	introspective self-reflection	.82	3.3	DA, FP	autobio.	−.130	1.00
42	sequenced "you'd"/"I'd" actions	.89	3.5	D	action	−.041	.943
43	father-son dialogue, football	.82	8.6	D	grasping	−.103	1.00
44	reported quoted dialogue	.89	3.8	D	speech	+.305	.006*
45	geographic landscape, terrain	.96	6.6	DA, FP	spatial	+.253	.006*
46	confrontational male dialogue	.73	2.4	D	threat	+.035	.531
47	silence-firing (empty TR)	.85	2.0	VA	sound	+.191	.025*
48	long elaborate quoted speech	1.00	1.2	SM	speech	+.476	.006*
49	silence-firing (empty TR)	.85	1.9	VA	sound	+.256	.016*

F AutoInterp implementation

Protocol. For each atom we select the 10 highest-activating TRs as train exemplars and 20 held-out items (10 positives = TRs ranked 11–20 by activation; 10 negatives = randomly sampled non-firing TRs). Each exemplar is presented as the text segment aligned to that TR, with the peak-activation word wrapped in bold (or `[silence]` for empty TRs). We additionally provide per-exemplar metadata: `pos` (story position, 0–1) and `spr` (syllable count in the firing TR), with atom-level mean $\pm$ std so the judge can assess whether positional or rate cues are diagnostic.

All 50 atoms are scored in a single Claude Sonnet prompt per model ($\sim$50K input tokens). The judge is asked to: (a) produce a one-sentence description of each atom, (b) label each atom as `mono` (single coherent concept), `poly` (multiple unrelated triggers), or `dead` (no pattern), and (c) score each of the 20 held-out items 1–5 for match to the description (without seeing ground-truth labels). AUC is computed per atom from the judge's scores using the held-out answer key.

Results across coupling ablations. We run the protocol on all three coupling variants (Table 1) at 3 seeds each (9 total runs). Table 6 reports 3-seed mean $\pm$ std.

Table 6: AutoInterp metrics (3-seed mean $\pm$ std).

Model	Median AUC	Frac ≥ 0.7	Frac ≥ 0.9	Mono / Poly / Dead
Coupled	$\mathbf{0.847 \pm 0.048}$	$\mathbf{87\% \pm 8\%}$	$\mathbf{33\% \pm 15\%}$	45–49 / 1–5 / 0
Sem→Brain	0.808 ± 0.052	$75\% \pm 11\%$	$27\% \pm 7\%$	44–48 / 2–6 / 0
Brain→Sem	0.690 ± 0.036	$45\% \pm 11\%$	$9\% \pm 5\%$	50 / 0 / 0

The ordering Coupled $>$ Sem→Brain $>$ Brain→Sem holds across all 3 seeds on median AUC.

G Semantic coherence methodology

Semantic coherence measures how tightly the text segments that strongly activate an atom cluster in the source LLM's representation space. For each atom k, we take the Gemma-2-2b embeddings (layer 25, TR-aligned) of its top-n activating TRs and compute the mean pairwise cosine similarity among them. We then compare this to a null distribution: for each of 1,000 permutations, we draw n random TRs (matched for story and position) and compute the same mean pairwise cosine. The coherence score is the z-score of the observed mean cosine relative to this null distribution.

We use $n = 20$ (top-20 activating TRs per atom). The metric is computed on training data only (not the held-out test story). Across the 50 coupled-SAE atoms, median coherence $z = 3.4$ (3-seed mean $z = 4.9 \pm 1.0$), indicating that atom activations cluster in semantic feature space well above chance. Low-coherence atoms ($z < 2$) tend to be polysemantic or encode positional/prosodic rather than lexical features (e.g., silence-firing atoms, story-opening atoms).

H Neurosynth term matching

Term selection. The Neurosynth v7 TF-IDF vocabulary contains 3,228 terms. To obtain a broad yet interpretable reference set, we curated single-word terms denoting cognitive, perceptual, emotional, or behavioral constructs, excluding anatomical labels, neuroimaging methodology, and generic English words. We then required each term to appear in at least 50 Neurosynth studies, yielding a final set of 152 terms.

Meta-analytic maps. For each of the 152 terms, we generate a meta-analytic association map using the Neurosynth v7 database (14,371 studies) via the NiMARE MKDA Density estimator. Each term map is a z-statistic volume in MNI space. We project it to the fsaverage surface using FreeSurfer `mri_vol2surf` (mid-cortical depth, white surface), obtaining a 327,684-vertex map per term.

Atom–term matching. Atom descriptions produced by AutoInterp (Appendix F) are embedded with a sentence-transformer (`all-mpnet-base-v2`; Reimers and Gurevych 2019). Term names are embedded identically. For each atom k we identify its best-matching term by cosine similarity of their embeddings. Matching is non-unique: multiple atoms may match the same term.

Brain map correlation. For each atom–term pair $(k, t^*(k))$ we compute cosine similarity and Pearson r between the atom's per-subject brain map and the term's z-map (positive z only), then average the metric across subjects. This per-subject-then-average scheme matches the Yeo 7 analysis (Appendix K). The analysis mask per subject is the intersection of the subject's selected-voxel coverage (which fsaverage vertices receive any projection from the subject's selected voxels, independent of atom weights) and the Neurosynth union coverage (any vertex where at least one term has $z > 0$). This mask is the same for all atoms and all terms within a subject, avoiding atom-weight-dependent masking.

Permutation test (spatial spin). We test whether each atom's brain map is more correlated with its matched Neurosynth term map than expected under a null that preserves spatial autocorrelation. We use the within-mask spin test [Alexander-Bloch et al., 2018]: for each of 1,000 permutations, a uniform random rotation from $SO(3)$ is applied to both hemispheres of the shared coverage mask, reassigning each rotated vertex to its nearest in-mask neighbor. The same rotation is applied to all 152 term maps simultaneously, preserving the joint null structure. Per-atom p-values are the fraction of permuted Pearson r values that equal or exceed the observed value, with the add-one correction of Phipson and Smyth [2010] to prevent $p = 0$. We apply Benjamini–Hochberg FDR correction across $K = 50$ atoms and report q-values at FDR $= 0.05$.

I Cross-subject transfer details

This appendix provides additional details on the cross-subject transfer experiment described in §4.4.

Encoder freezing. The encoder trained jointly on UTS01–UTS06 is frozen (all encoder weights, pre-encoder bias, and semantic decoder weights are fixed). Only the subject-specific linear brain decoder is re-trained for each held-out subject. This reduces the transfer problem to a linear regression on the fixed 200-dimensional input (50 atoms $\times$ 4 delays), rather than fitting a full encoding model on 9,216-dimensional dense features (2304×4 delays).

Held-out subjects. UTS07 and UTS08 each contribute 27 stories total (26 excluding test), yielding 22 effective training stories after the same 4-story validation holdout used for model selection. These subjects were not seen during encoder training.

Learning curve protocol. For each training-story count $n \in \{0.5, 1, 2, 4, 8, 16, 22\}$, we randomly sample n stories from the available training pool (the $n = 0.5$ condition uses the first half of a single randomly selected story). The brain decoder is fit via ridge regression with the regularization parameter selected on the 4-story validation set. For each n, we repeat the procedure with 6 different random story draws and average results across both held-out subjects, yielding 12 data points per condition. The same story draws and validation protocol are used for the matched dense-feature ridge and random encoder baselines.

Ridge baseline. The matched ridge baseline uses the same Gemma-2-2b features (layer 25, RMS-normalized) with 4 hemodynamic delays, fit with ridge regression [Dupré La Tour et al., 2022] using the same validation stories and story draws as the SAE transfer. This ensures a fair comparison: both methods see the same data, the same validation split, and the same test story.

Random encoder control. The random encoder control uses an architecture-identical SAE with Xavier-initialized weights that are frozen (never trained). Only the brain decoder is fit, as in the SAE transfer. This control isolates the contribution of dimensionality reduction and the TopK bottleneck from the learned representation.

J Cross-subject cortical map consistency

Cross-subject Pearson r. For each subject pair (s_1, s_2) we use the per-pair intersection mask: vertices where both subjects have nonzero decoder magnitude for at least one atom. This matches the shared-vertex criterion of the fsaverage generalization analysis (§4.4) and avoids correlating real decoder weights in one subject against structural zeros in another. We then compute the Pearson correlation between $d_{s_1,k}$ and $d_{s_2,k}$ across all intersection vertices for every pair from the $\binom{6}{2} = 15$ pairs among UTS01–UTS06. Per-atom consistency is the mean Pearson r across all 15 pairs. All values are reported as 3-seed mean $\pm$ std across independent training runs.

Permutation null. For each of 1,000 permutations, atom indices are shuffled independently within each subject's decoder map matrix, and the mean Pearson r across all pairs and atoms is recomputed. This breaks the atom-label correspondence while preserving each subject's cortical map structure, giving the distribution expected if atom assignments were random. The global p-value is the fraction of permutation means that equal or exceed the observed mean; per-atom p-values use the same null distribution.

K Yeo 7 network mapping

Brain maps. Atom brain maps are constructed as described in Appendix J and projected to the fsaverage surface (327,684 vertices). We assign each atom to the Yeo–Krienen 7-network parcellation [Yeo et al., 2011] by computing the cosine similarity between each atom's decoder magnitude map and the binary indicator vector of each network, averaged across subjects.

Per-subject cosine profiles are averaged across subjects to give the mean profile, with SEM across subjects. The dominant network for each atom is the network with highest mean cosine similarity.

Permutation test (spatial spin). To assess whether each atom's network enrichment exceeds spatial chance, we use the spin test of Alexander-Bloch et al. (2018). For each permutation, independent uniform random rotations from $SO(3)$ (QR decomposition, $\det = +1$ enforced) are applied to the sphere coordinates of the subjects' union analysis mask, mapping each vertex to its nearest neighbor within the same hemispheric mask. This spatially rotates the atom decoder weight maps relative to fixed network boundaries while preserving their spatial autocorrelation. The mask is restricted to selected-voxel coverage (excluding medial wall), so no rotated vertex lands outside the analysis footprint. One permutation is drawn per iteration and applied to all atoms simultaneously. Per-cell p-values use a one-sided right-tail test with the add-one correction of Phipson & Smyth (2010). FDR correction across all $7 \times 50 = 350$ network–atom cells uses the Benjamini–Hochberg procedure at $q = 0.05$.

L Per-subject brain prediction breakdown

The brain correlation score reported in Table 1 is averaged across all six training subjects. Table 7 gives the per-subject breakdown comparing the coupled SAE against the matched ridge baseline.

The mean row $\pm$ std is across seeds (matching Table 1); per-subject $\pm$ std is also across seeds. The SAE exceeds ridge only on UTS04, the subject with the fewest training stories—consistent with the sparse bottleneck acting as implicit regularization under low data.

Table 7: Per-subject brain correlation score on the held-out test story. SAE values are 3-seed mean $\pm$ std. Ridge values are the matched baseline (single fit; no seed variance). SAE / Ridge ratios use the 3-seed SAE mean.

Subject	n_{train}	SAE brain corr.	Ridge brain corr.	SAE / Ridge
UTS01	79	0.152 ± 0.013	0.181	0.84×
UTS02	79	0.198 ± 0.011	0.234	0.85×
UTS03	79	0.232 ± 0.013	0.295	0.79×
UTS04	21	0.130 ± 0.011	0.117	1.11×
UTS05	22	0.111 ± 0.005	0.120	0.93×
UTS06	22	0.150 ± 0.007	0.161	0.93×
Mean		$\mathbf{0.162 \pm 0.005}$	**0.185**	**0.88**×

M Cross-seed representational stability

We assess whether models with different random seeds converge to equivalent representations using two complementary measures computed on 10 held-out training stories (UTS03): (1) **linear CKA** [Kornblith et al., 2019] between activation matrices $S \in \mathbb{R}^{T \times K}$ across seeds, which captures similarity of the full representation space; and (2) **Hungarian-matched encoder cosine similarity** [Kuhn, 1955], which measures whether individual atoms align across seeds after optimal one-to-one matching.

Table 8: Cross-seed stability of learned representations. Mean Hungarian-matched encoder cosine similarity and linear CKA across seed pairs, evaluated on 10 held-out stories. Architecture variants A–E correspond to the rows of Table 2.

Model	Hungarian cos. sim.	CKA
B. Coupled SAE (enc. norm, best)	**0.41**	**0.944**
A. Baseline (no encoder norm)	0.33	0.890
C. Matryoshka $[2, 5]$	0.51	0.989
D. Matching-pursuit SAE	0.81	0.982
E. + Per-voxel scale	0.45	0.964
Sem→Brain	0.27	0.777
Brain→Sem	0.31	0.976
Random encoder	0.04	0.485

Linear CKA is at or above 0.89 for all non-random coupled variants, with the best model and most architecture variants reaching 0.94 or higher, confirming that the global activation space is highly consistent across seeds even when individual atom identities differ; the random encoder control drops to 0.485, showing that this consistency requires a learned representation. At the atom level, the MP-SAE variant achieves Hungarian cosine similarity of 0.81 alongside CKA of 0.982, suggesting that greedy pursuit imposes a stronger inductive bias toward a canonical atom ordering than TopK training, and understanding this relationship between encoder architecture and atom-level reproducibility is an open question for future work.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction match the empirical scope of the paper: retained brain prediction under compression, interpretability evaluated through semantic and cortical analyses, and transfer to held-out subjects with a frozen encoder and linear decoder refit; see the Abstract, §1, and §4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The Discussion states the main limitations explicitly, including that the learned atoms are descriptive rather than mechanistic, that experiments are restricted to one naturalistic dataset and stimulus setting, and that broader heterogeneous multisubject settings remain future work; see §5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not present theoretical results or formal proofs.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper specifies the model, loss, preprocessing, data splits, optimization settings, ablations, and evaluation procedures in §3 and Appendix A–M, which together provide the information needed to reproduce the reported experiments.

   Guidelines:

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

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide anonymized code, configuration files, and scripts with the submission, together with instructions for data access, preprocessing, training, evaluation, and figure generation; see Appendix B (*Software and compute*).

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The main paper gives the experimental setup needed to interpret the results (§3, §4), and the appendix gives the remaining training, preprocessing, hyperparameter, interpretability, and transfer details needed to understand how each result was produced.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Main results are reported as 3-seed mean $\pm$ std, the transfer curves report variability across repeated story draws, and the cortical analyses report permutation-based significance with multiple-comparison correction where appropriate; see §4 and Appendices E, G, H, I, J, and K.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix B (*Software and compute*) reports the software stack, representative training time per run, the GPU class and memory, and which components run on CPU versus GPU, which is sufficient to estimate the compute needed to reproduce the experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: To the best of our knowledge, the work conforms to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The Discussion notes a practical benefit in translational settings where long scanning sessions are impractical, and a potential negative impact in the form of increased privacy risk from inferring sensitive information from neural data; see §5.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The released assets are analysis code, configurations, and scripts for a public fMRI dataset; the paper does not release a high-risk generative model, scraped dataset, or similarly high-misuse asset.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper cites the external dataset, model, and software assets it uses, and Appendix B (*Software and compute*) states the relevant licenses or terms of use for the OpenNeuro distribution of the LeBel et al. data (CC0), Gemma-2-2b (Gemma terms of use), FreeSurfer (registration-based academic license), and `himalaya` (BSD-3-Clause).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: We release anonymized code, configuration files, and scripts together with documentation and reproduction instructions; see Appendix B (*Software and compute*).

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not report new crowdsourcing experiments or new data collection from human subjects; it analyzes an existing public dataset.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve new human-subject data collection by the authors; it analyzes an existing public dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: LLMs are a central part of the method: Gemma-2-2b provides the semantic input representation, and AutoInterp uses an LLM judge to evaluate whether learned atoms admit predictive natural-language descriptions on held-out contexts; see §3, §4, and Appendix E.

    Guidelines:

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