Semantic-Cortical Components: A Multisubject Coupled Sparse Autoencoder for Interpretable Language Encoding in fMRI

Alicia Zeng
University of California, Berkeley
Erised AI
alicia@erised-ai.com

Abstract

Encoding models built on language-model activations predict cortical responses to natural language well but are hard to interpret: semantic selectivity is not identifiable in a dense, correlated activation space; cortical organization stays implicit in voxelwise weights; and correspondence across subjects is usually established only after fitting. To address all three, we introduce a multisubject coupled sparse autoencoder (SAE), trained jointly to reconstruct language-model activations and to predict the brain responses of six subjects from one sparse code. Each of the 50 code dimensions, which we call semantic-cortical components, pairs a semantic direction with a cortical map in every subject. On story-listening fMRI, the code retains about 82% of the brain-prediction accuracy of a ridge regression fit on all 2,304 activation dimensions, with only five components active per time point. We evaluate the components on three properties. First, nameability: an LLM judge's one-sentence description of a component predicts which held-out text activates it. Second, cross-subject consistency: their cortical maps agree across subjects. Many maps also align with resting-state networks and meta-analytic maps of published studies. Third, cross-seed stability: many components recur when the model is retrained, though not all. Only the coupled SAE does well on all three properties; matched models trained on the two objectives in sequence do not. The code also transfers to new subjects: for two subjects unseen in training, a linear readout from the frozen code, fit on 22 stories each, recovers 87% and 102% of the brain-prediction accuracy of ridge with about 2% of its parameters. Together, the components give cognitive neuroscience an interpretable unit of analysis for semantic-to-brain encoding: each links a semantic feature to its cortical response pattern in every subject.

1 Introduction

Functional mapping in cognitive neuroscience has moved from localizing coarse selectivity for controlled stimuli to modeling distributed representations of naturalistic stimuli in individual subjects, each scanned for hours [Huth et al., 2016]. In this setting, linear encoding models now predict brain responses to language from the activations of large language models (LLMs) [Toneva and Wehbe, 2019, Caucheteux and King, 2022, Goldstein et al., 2022, Antonello et al., 2023]. These models predict brain responses across much of cortex and recover its semantic organization. Three neuroscience questions, however, need a model that can be read directly: which semantic features a region responds to, how cortex organizes them, and whether subjects share that organization. Current encoding models remain difficult to interpret for three reasons.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

(1) **Semantic selectivity is non-identifiable.** LLMs are thought to encode more semantic features than they have dimensions by assigning them overlapping, non-orthogonal directions, a phenomenon known as *superposition* [Elhage et al., 2022, Bricken et al., 2023]. As a result, a voxel's fitted weights cannot be uniquely attributed to semantic features [Zeng and Gallant, 2025].

(2) **Cortical organization remains implicit.** Encoding models typically fit an independent readout for each voxel, a small spatial unit of fMRI measurement. Cortex, however, represents semantic information in distributed, overlapping networks that span many voxels [Haxby et al., 2001, Huth et al., 2016, Zhang et al., 2020]. Because the unit of estimation does not match the structure of the representation, larger-scale cortical organization is not learned directly but must be inferred after fitting.

(3) **Cross-subject structure is not learned directly.** Most approaches first derive cortical structure in each subject and then align subjects by anatomical registration or functional alignment. Cross-subject correspondence is therefore imposed after fitting rather than learned by the model.

Existing methods address one or two of these issues, but to our knowledge none addresses all three [Fyshe et al., 2014, Bilenko and Gallant, 2016, Van Uden et al., 2018, Skrill and Norman-Haignere, 2024]. The missing link is a factorization that is jointly semantic, cortical, and multisubject within a single training regime. We address all three with a **multisubject coupled sparse autoencoder** (SAE), trained jointly on LLM activations and the brain responses of six subjects. An SAE re-expresses each dense activation vector as a sparse combination of learned directions [Elhage et al., 2022, Bricken et al., 2023]. We use the TopK variant from LLM interpretability [Gao et al., 2025], in which only the k largest entries of the code stay active. An encoder maps the activations to one sparse code. From that code, a semantic decoder reconstructs the activations and subject-specific brain decoders predict each subject's responses (Fig. 1). Each code dimension therefore pairs a direction in activation space with a cortical map in every subject. We call this pair a *semantic-cortical component*. For example, an LLM judge describes one component as *food and drink items and meals*, and that component has a cortical map in each subject (Fig. 1b; more in Fig. 2). The model requires no hand-labeled categories, no predefined regions of interest, and no post hoc functional alignment.

The components are meant to serve as a unit of analysis for cognitive neuroscience: a researcher selects a component by its description, reads its cortical map in every subject, and compares that map with reference maps of cortical networks. This use requires three properties, which we measure throughout. *Nameability*: an LLM judge's description of the component predicts which held-out text activates it. *Cross-subject consistency*: its cortical maps agree across subjects, although the model is trained without anatomical alignment. *Cross-seed stability*: it recurs when the model is retrained. A sparse code does not make voxel selectivity identifiable, but it replaces arbitrary directions with components that can each be tested on these three properties.

Our contributions are:

- A **multisubject coupled SAE** that jointly learns semantic structure and cortical organization in a single training regime. With 50 components and k=5 active per time point, it retains about 82% of the brain prediction of ridge regression on the same activations.
- **The costs and benefits of coupling.** Compared with matched two-stage models that train the encoder on one objective and then fit a decoder for the other (§4.1 and §4.5), only the coupled SAE does well on both objectives and all three properties. A brain-only basis predicts brain responses slightly better but is worse on semantic reconstruction, nameability and cross-seed stability. A semantics-only basis reconstructs the activations slightly better but is 19% worse on brain prediction and less consistent across subjects.
- **Learned sparse semantic-cortical components** as a unit of analysis. Their cortical maps are more consistent across subjects than maps obtained by projecting ridge encoding weights onto the same directions. Many of these maps align with the Yeo 7 resting-state networks [Yeo et al., 2011] or match a Neurosynth meta-analytic map [Yarkoni et al., 2011].
- **Parameter-efficient adaptation to new subjects**, using the frozen encoder and only a new linear readout. For two new subjects, this readout, fit on 22 stories each, recovers 87% and 102% of the brain prediction of ridge regression on the same activations, with about 2% of its parameters.

2 Related work

Encoding models and non-identifiability. Encoding models predict neural responses from stimulus features. In controlled paradigms, features such as hand-labeled categories or contrasts are directly interpretable; for naturalistic language, they 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 their cosine similarity with concept vectors. Both heuristics inherit this non-identifiability. In lower-dimensional word-embedding spaces, Zeng and Gallant [2025] formalize this problem and propose sparse concept components as a principled basis for neuroscientific interpretation. Extending sparse concept components to contextual LLM activations raises a problem of scale. Gemma Scope, a suite of SAEs for Gemma-2-2b, trains 16,000 to 1 million features (latents) per SAE [Lieberum et al., 2024]. At that scale, many latents may be dead, polysemantic, redundant, or irrelevant to brain prediction, and a brain study must first select among them. The coupled SAE avoids this selection step: it learns its components jointly with brain prediction, so brain relevance shapes them directly.

Recovering distributed cortical organization. Large-scale cortical organization is often studied through resting-state functional connectivity or through sparse and dictionary-learning methods applied directly to blood-oxygen-level-dependent (BOLD) signals [Lee et al., 2011, Xie et al., 2017, Thirion et al., 2024, Hari et al., 2026]. Both can recover distributed structure, but they share two limitations. First, they do not link that structure directly to semantic function. Second, they are sensitive to shared noise (head motion, physiological fluctuations, scanner drift), which can introduce spurious correlations.

Post hoc analyses instead use fitted encoding-model weights, clustering voxels with similar weight profiles or aggregating weights along semantic directions of interest [Huth et al., 2016, Wehbe et al., 2018, Zhang et al., 2020, Meschke et al., 2023, Hendrikx et al., 2026]. These analyses still infer cortical organization after fitting. Resting-state network recovery and post hoc encoding-model analyses also tend to hard-assign each voxel to one network. This biases the result toward one-to-one mappings between sites and functions. Cortical representations, however, are widely understood to be overlapping and many-to-many: each site can serve 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 after fitting. Shared structure across subjects is usually addressed 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]. These methods improve cross-subject comparability but treat semantic content and cortical organization as separate inference problems.

Sparse autoencoders and interpretability evaluation. Our architecture builds on TopK SAEs [Gao et al., 2025], which decompose dense activations into sparse features that are often more interpretable than single dimensions [Elhage et al., 2022, Bricken et al., 2023]. Our nameability evaluation follows LLM-judge protocols that test whether a feature's description predicts its activations on held-out text [Bills et al., 2023, Paulo et al., 2025]. Unlike prior SAEs, ours is also trained to predict the brain responses of multiple subjects.

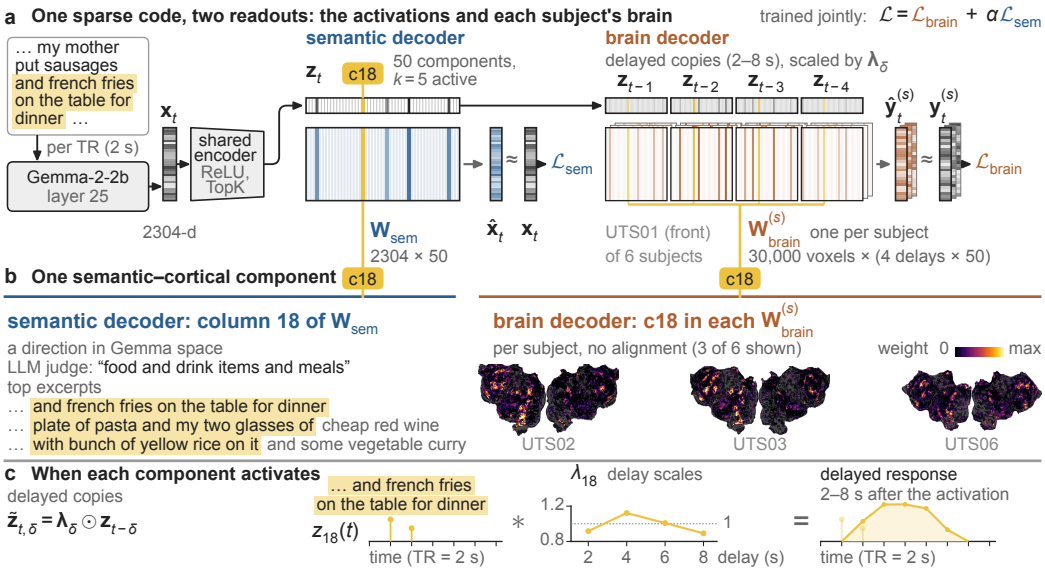


Figure 1: **The coupled SAE reconstructs the activations and predicts every subject's responses from one sparse code, whose dimensions are semantic-cortical components.** Blue: semantic side; orange: brain side; gold: component c18; grey: observed x_t and $y_t^{(s)}$. (a) Gemma-2-2b activations x_t (layer 25, 2304-d, one per TR) are encoded into a sparse code z_t with k=5 of 50 components active (dark cells). The semantic decoder W_{sem} reconstructs $\hat{x}_t$ from the active components' columns (bias b_{pre} omitted). The brain decoder takes copies of the code delayed by 1–4 TRs (2–8 s) and rescaled by learned per-component delay scales λ_δ. Each subject's non-negative $W^{(s)}_{\text{brain}}$ maps these delayed codes to predicted responses $\hat{y}_t^{(s)}$ (one card per subject; one block of 50 columns per delay). All parts are trained jointly on $\mathcal{L} = \mathcal{L}_{\text{brain}} + \alpha\, \mathcal{L}_{\text{sem}}$. (b) Component c18 (seed 0). Left: its column of W_{sem}, summarized by the LLM judge's description (AutoInterp, §4.2) and top-activating excerpts (highlight: the TR where c18 is active). Right: its cortical map (brighter = larger), computed from its four columns of $W^{(s)}_{\text{brain}}$ (Appendix F.3), in three of the six subjects, each on its own flattened cortex. (c) The delay model for c18: a schematic activation trace $z_{18}(t)$ (the TRs of the highlighted words) convolved with its learned delay scales (real values, all near 1; dotted line) gives a response 2–8 s later (faded stems: the activation again).

3 Methods

3.1 Architecture: one sparse code, a semantic decoder and per-subject brain decoders

We train one sparse code jointly on language-model activations and on the brain responses of multiple subjects listening to spoken narrative stories. The stories come from one shared set, but not every subject heard every story (§3.3). At each fMRI sampling time point (TR), the model takes a dense Gemma-2-2b activation vector $x_t \in \mathbb{R}^{d_{\text{sem}}}$. It maps this vector to a sparse latent code $z_t \in \mathbb{R}^K$ over K components, at most k of which are active. The encoder subtracts a learned pre-encoder bias b_{pre}, then applies a linear map, a ReLU and TopK selection. Two kinds of decoder read out from z_t: a semantic decoder reconstructs the original Gemma-2-2b activations (adding b_{pre} back), and subject-specific brain decoders predict each subject's brain response $\hat{y}_t^{(s)}$ (Fig. 1). We use *encoder* and *decoder* in the autoencoder sense throughout. In neuroimaging terms, a brain decoder thus runs in the encoding direction, from stimulus to brain.

Because brain 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. For each delay $\delta \in \{1, 2, 3, 4\}$ TRs, we form a delayed copy of the sparse code and rescale it elementwise by a learned vector of per-component *delay scales* $\lambda_\delta \in \mathbb{R}^K$: $\tilde{z}_{t,\delta} = \lambda_\delta \odot z_{t-\delta}$. Each component thus has its own temporal profile. The four delayed codes are concatenated and passed to each subject's linear brain decoder. The decoder has no bias term, and its columns are unit-norm and non-negative.

The sign of a component's cortical effect is therefore carried by its delay scales (Appendix A). The full model is:

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

$$\hat{x}_t = W_{\mathrm{sem}}\, z_t + b_{\mathrm{pre}}, \tag{2}$$

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

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

A component's *cortical map* in subject s is the absolute value of the delay-weighted sum of its four columns of $W_{\mathrm{brain}}^{(s)}$: for component j, $\big|\sum_\delta [\lambda_\delta]_j\, W_{\mathrm{brain}}^{(s)}[\,\cdot\,, \delta, j]\big|$ (Appendix F.3; example components in Fig. 2 and Appendix F.8). The brain decoders receive no anatomical information: voxels are an unordered set, with no coordinates or inter-subject registration. Any cross-subject consistency in the learned maps therefore cannot come from anatomical constraints in the model. The coupled SAE is a TopK SAE with unit-norm encoder rows (Appendix A); Appendix B compares matching-pursuit [Costa et al., 2025] and matryoshka [Bussmann et al., 2025] variants.

3.2 Objective, training and reporting conventions

We minimize $\mathcal{L} = \mathcal{L}_{\mathrm{brain}} + \alpha \mathcal{L}_{\mathrm{sem}}$, where both terms are mean-squared errors and α is the coupling weight. The reported model has $K = 50$ components, k=5 active per time point and $\alpha = 1$. Appendix C sweeps all three and gives the reasons for these choices. The sparsity k bounds the capacity of the code; it is not a claim about how many concepts the brain represents at once. Each minibatch computes $\mathcal{L}_{\mathrm{brain}}$ for one subject, on the time points of the stories that subject heard. Successive minibatches cycle round-robin across subjects (Appendix A). Training uses AdamW with early stopping on 4 fixed validation stories (patience 10). Scores use the checkpoint at the stopping epoch, not the best validation epoch (checkpoint convention in Appendix A). Unless stated otherwise, reported numbers are the mean $\pm$ population standard deviation over 3 seeds (seeds 0–2). Components are numbered from 0 as in Table 15, and cj denotes component j of the seed-0 model.

3.3 Data, voxel selection, matched ridge baseline and metrics

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 as new subjects for 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. Only UTS01–UTS03 heard the 4 validation stories. The coupled SAE therefore trains on 79 of their 83 stories. Brain responses are z-scored per voxel. The semantic inputs are 2,304-d activations from layer 25, the final layer of the open-weight Gemma-2-2b [Gemma Team, 2024] and the best of layers 15, 20 and 25 (Appendix A). The activations are downsampled to the fMRI sampling rate of one TR every 2 s and normalized by their root-mean-square (RMS) value. To focus on language-responsive voxels, we fit a separate voxel-selection ridge regression on these activations using 10 training stories. For each subject, we keep the 30,000 voxels with the highest mean leave-one-story-out cross-validated (CV) correlation. The count is fixed so that scores are comparable across subjects, and selection uses training data only, before any compared model is fit. Throughout, the *matched ridge baseline* is a ridge regression from the same Gemma-2-2b activations at the same four delays ($4 \times 2{,}304 = 9{,}216$ features per voxel) to the same voxels. Its regularization is chosen on the four validation stories. It is then refit on all of each subject's training stories: 83 for UTS01–UTS03, against 79 for the coupled SAE (Appendix D). On the test story, we report two primary metrics. The *brain score* is the Pearson correlation between predicted and observed responses, computed per voxel and averaged over the 30,000 selected voxels and then over subjects. The *semantic score* is the Pearson correlation between reconstructed and original Gemma-2-2b activations, computed per dimension and averaged over dimensions.

Table 1: **Only the coupled SAE does well on both objectives and all three properties.** Brain and semantic scores on the test story are mean ± std over 3 seeds; AutoInterp AUC is the median over the 50 components, then mean ± std over seeds; cross-subject r is the mean cortical-map correlation between subjects (§4.3); cross-seed stability is the Hungarian-matched cortical-map cosine between seeds (§4.5; Appendix F.6); ridge is deterministic. Ridge: the matched baseline on the dense activations (§3.3). Random encoder: frozen random encoder, decoders trained. Sem→Brain and Brain→Sem: encoder trained on the first-named objective alone, then a decoder for the other fit on the frozen code (§4.1), as is the random encoder's semantic decoder. Dashes: not measured or not applicable. Bold: best sparse model in each column. AutoInterp AUC (per-component nameability, Claude Fable 5.1 judge; §4.2) saturates for two models (Appendix F.1). †Early stopping uses a validation brain loss restricted to each subject's own stories (Appendix A).

Method	Brain score	Semantic score	AutoInterp AUC	Cross-subject r	Cross-seed stability
Ridge (9,216-d)	0.185	—	—	—	—
Coupled SAE	0.162 ± 0.005	0.523 ± 0.006	$\mathbf{0.995 \pm 0.004}$	$\mathbf{0.205 \pm 0.008}$	**0.65**
Random encoder	0.064 ± 0.005	0.407 ± 0.005	—	—	0.32
Sem→Brain	0.131 ± 0.006	$\mathbf{0.537 \pm 0.004}$	0.992 ± 0.003	0.177 ± 0.002	**0.65**
Brain→Sem†	$\mathbf{0.174 \pm 0.003}$	0.394 ± 0.013	0.922 ± 0.012	0.196 ± 0.006	0.54

4 Results

4.1 Fifty sparse components retain most of ridge's brain prediction

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 test story shared by all six training subjects (UTS01–UTS06), against the matched ridge baseline and a random encoder control. The coupled SAE reaches a brain score of 0.151 ± 0.008 over 20 seeds, 82% of ridge (0.185; Appendix A). The three seeds reported throughout (seeds 0–2) are among the higher of the 20 and score 0.162 ± 0.005 (88%). The random encoder, an architecture-matched control whose encoder is frozen at random initialization and whose decoders alone are trained, scores 0.064. Fifty sparse components thus retain most of the brain prediction of the full activations. The gap to the random encoder shows that the retained prediction comes from the learned code, not from sparsity or dimensionality reduction alone. Appendix D gives per-subject scores.

To isolate the effect of joint training, we compare the coupled SAE with two matched two-stage models, named by their training order. Sem→Brain trains the encoder for semantic reconstruction and then fits the brain decoder on the frozen code. Brain→Sem trains the encoder for brain prediction alone ($\alpha = 0$) and then fits the semantic decoder on the frozen code. All three models share the same architecture (recipe details in Appendix A). Each two-stage model falls short on the objective its encoder was not trained for (Table 1). Brain→Sem predicts brain responses slightly better than the coupled SAE (0.174 ± 0.003 against 0.162 ± 0.005). Its code, however, reconstructs the Gemma-2-2b activations at only 0.394 ± 0.013, no better than the random encoder (0.407 ± 0.005) and well below the coupled SAE (0.523 ± 0.006). Sem→Brain reconstructs the activations slightly better than the coupled SAE (0.537 ± 0.004) but predicts brain responses 19% worse (0.131 ± 0.006). The coupled SAE is thus close to the best model on both objectives: compared with Brain→Sem, coupling costs a little brain prediction and buys semantic reconstruction.

4.2 Components are nameable: an LLM judge's description predicts held-out activations

Prediction and reconstruction scores show that the components carry information about both brain responses and semantic content, but not that each component corresponds to one nameable concept. We test this in two ways, by semantic coherence and by AutoInterp. For semantic coherence, we embed a component's top-activating text segments with a sentence-transformer and express their mean pairwise cosine similarity as a z-score against random text segments. The median z is 4.4 ± 0.3 for the coupled SAE and lies between 4 and 5 for each of the three trained models of Table 1 (Appendix F.2). The text that activates a component is thus far more similar in meaning than random text.

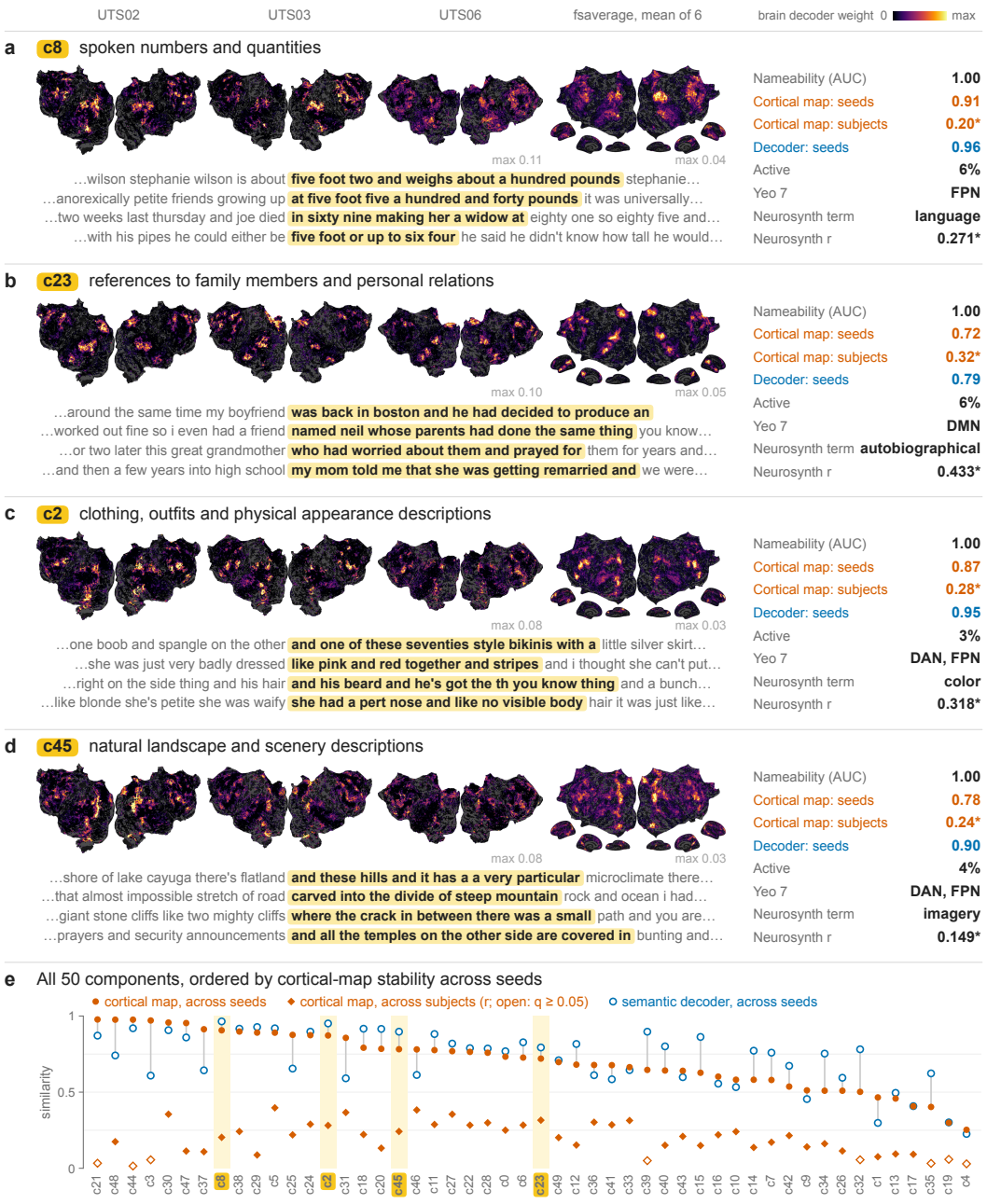


Figure 2: **Each component pairs a nameable description with a cortical map in every subject.** (a) c8, (b) c23, (c) c2 and (d) c45 of the seed-0 model. Each row shows the AutoInterp judge's description, the component's cortical map in three subjects and averaged on fsaverage (a common cortical surface), and top-activating excerpts with the firing TR highlighted. Each map has its own colour scale, with its maximum given below it, because the average map is weaker than any single-subject map. Side column: nameability (AutoInterp AUC); cross-seed stability of the cortical map; cross-subject r of the cortical maps; cross-seed stability of the semantic decoder; active % of time points; significant Yeo 7 networks (DMN = default mode, FPN = frontoparietal, DAN = dorsal attention); the best Neurosynth term and its map correlation r (*: $q < 0.05$). (e) All 50 components, ordered by cross-seed stability of the cortical map. Filled orange markers show the cross-seed stability of the cortical map and hollow blue markers that of the semantic decoder, each over 20 seeds (best-match cosine). Orange diamonds show the cross-subject consistency r of the cortical map, open if $q \geq 0.05$. Components a–d are highlighted (Appendices F.6 and F.3).

AutoInterp [Bills et al., 2023] evaluates whether a component admits a predictive natural-language description. An LLM judge is given a component's top-activating excerpts and asked to produce a one-sentence description. It also labels the component as `mono` (one concept), `poly` (several unrelated concepts) or `dead`, and scores held-out excerpts for how well they match the description. The area under the ROC curve (AUC) then measures how well those scores distinguish activating from non-activating excerpts. Because the judge also receives the story position and syllable count of each excerpt, a high AUC can also reflect non-semantic textual or temporal regularities. With Claude Fable 5.1 as judge, the median per-component AutoInterp AUC is 0.995 ± 0.004 across 3 seeds. Over these seeds, 98% of components exceed an AUC of 0.7, 48–50 of 50 are judged monosemantic, and none is dead. Nearly every component thus admits a one-sentence description that predicts its held-out activations. All reported values use the same judge, prompts and answer keys (Appendix F.1). The recovered descriptions span lexical domains such as *family members and personal relations* (c23), *food and drink items and meals* (c18) and *natural landscape and scenery* (c45), as well as non-semantic regularities such as story openings (c5). Five *speech-timing components* (c21, c30, c47, c48, c49) fire most strongly on pauses or rapid speech and are the most frequently active (34–53% of time points; Appendix F.7). These five components thus track the timing of speech rather than its meaning. Table 15 (Appendix F.9) and an interactive viewer[1] list all 50 components.

4.3 Cortical maps agree across subjects and align with resting-state and meta-analytic maps

We next ask whether each component recovers consistent cortical structure across subjects, even though every subject has its own brain decoder and no anatomical constraint applies. We measure cross-subject consistency by correlating each component's cortical maps between subjects after projection to fsaverage, a common cortical surface template (Appendix F.3). Averaged across subject pairs and components, the correlation is $r = 0.205 \pm 0.008$. On the seed-0 model, none of 1,000 permutations that shuffle the component labels independently within each subject reaches this value ($p < 10^{-3}$; null $r = 0.006 \pm 0.004$). The learned components thus recover reproducible cortical maps for a given semantic direction.

We next compare this result with the common post hoc approach to encoding-model interpretation: projecting ridge weights onto semantic directions of interest [Huth et al., 2016, Wehbe et al., 2018, Zeng and Gallant, 2025]. For each component we project every subject's ridge weights onto that component's direction, taken either as its encoder row or as its semantic-decoder column. This gives one ridge map per component and subject (Appendix F.3). Both projections give lower consistency than the coupled SAE: ridge $r = 0.174 \pm 0.002$ with encoder directions and 0.177 ± 0.004 with decoder directions, against 0.205 ± 0.008 (one-sided paired t-tests over the 50 components; $p < 10^{-5}$ for both). At the level of single components, 42/50 coupled-SAE maps survive Benjamini–Hochberg false-discovery-rate (BH-FDR) correction under the shuffled-component null, against 19/50 for the encoder-direction ridge projection and 34/50 for the decoder-direction projection (Appendix F.3). Because the ridge projections use the components' own directions, the gain comes from estimating each map within a shared sparse model, not from the choice of semantic directions.

Consistency across subjects shows that the maps are shared, not that they correspond to known cortical organization, so we next compare them with two external references, the Yeo 7 resting-state networks and Neurosynth meta-analytic maps. Both come from paradigms other than naturalistic story listening, so we expect partial rather than exact alignment. Yeo 7 [Yeo et al., 2011] divides cortex into seven networks of correlated spontaneous activity in resting-state fMRI. For each component, we compute in each subject the cosine similarity between its fsaverage cortical map and each network's binary indicator vector, and average over subjects. Significance comes from a spatial spin null, which rotates the map on the cortical sphere and so preserves its spatial smoothness (BH-FDR $q < 0.05$ across all 7×50 network–component pairs; Appendix F.4).

Significant affiliations are concentrated in the default mode network (DMN; 23/50 components) and the frontoparietal network (FPN; 18/50 components). DMN components include *family members and personal relations* (c23), *plural groups or categories of people* (c28) and *quoted speech addressing 'you'* (c46). FPN components include *spoken numbers and quantities* (c8), *letters, documents and written objects* (c20), and *natural landscape and scenery* (c45). Among the other networks, the dorsal attention network (DAN) includes *forceful physical body actions* (c24). The somatomotor network includes the story-opening (c5) and rapid-speech (c48) components. The ventral attention

[1] `https://alicialitrtwe.github.io/research/jointsae.html`

network includes two other speech-timing components. The visual and limbic networks have no significant affiliations, as expected for a listening task with no visual stimuli (per-component affiliations: Table 15).

Neurosynth [Yarkoni et al., 2011] aggregates approximately 14,000 published fMRI studies and provides, for each term in its vocabulary, an automated meta-analytic map of where studies that mention the term report activation. From its vocabulary, we curated 152 cognitive, perceptual, emotional, and behavioral terms (Appendix F.5). For each component, we use sentence-transformer embeddings of its AutoInterp description to select the most semantically similar Neurosynth term. We then compare the component's cortical map with that term's meta-analytic map under the same spin null. Because the Neurosynth maps aggregate heterogeneous task paradigms, we treat these matches as external validation rather than as exact anatomical correspondences. Of the 50 components, 21 match their term significantly (BH-FDR $q < 0.05$ across all component–term pairs). Significant matches include *quoted speech addressing 'you'* (c46) with "speech" ($r = 0.47$), *family members and personal relations* (c23) with "autobiographical" ($r = 0.43$), *forceful physical body actions* (c24) with "action" ($r = 0.37$), and *clothing and physical appearance* (c2) with "color" ($r = 0.32$). For these 21 components, the cortical map thus overlaps significantly with where published studies of the most similar vocabulary term report activation. Taken together, the two references place many components' maps within known large-scale organization, at the level of networks and at the level of individual terms. Table 15 lists per-component results.

4.4 Most components recur across seeds

We next ask whether the components recur when the model is retrained from a different random seed. The code as a whole is highly reproducible across seeds: linear centered kernel alignment (CKA) [Kornblith et al., 2019], a rotation-invariant similarity between the codes of two seeds, is 0.94. Individual components are only partly reproducible. We call a seed-0 component stable if its best-match cosine, averaged over the other 19 seeds, is above 0.5 on both the semantic-decoder column and the cortical map. Of the 50 components, 43 pass, including every component named in §4.2 and §4.3; the 7 that fail are narrow or story-specific (Appendix F.6). Table 15 marks the 7 that fail. Under a stricter criterion (reappearance above cosine 0.7 in at least 70% of the other seeds), about half of the components recur (Appendix F.6). Component stability changes little when the model is trained on different subjects but drops when it is trained on different stories. A matching-pursuit variant is more stable but predicts the brain less well (Appendix F.6).

4.5 Only joint training gives all three properties

To see whether joint training is what gives the components these properties, we compare the three models on nameability, cross-subject consistency and cross-seed stability (Table 1, last three columns). The coupled SAE is the only model that does well on all three. Brain→Sem components are less nameable (AUC 0.922 against 0.995 and 0.992, below the other two models in every seed; 2–9 of 50 judged polysemantic against 0–2; Appendix F.1) and less stable across seeds (Hungarian-matched cortical-map cosine 0.54 against 0.65; Appendix F.6). Sem→Brain maps have the lowest mean cross-subject consistency ($r = 0.177$ against 0.205 and 0.196; Appendix F.3). The stability and cross-subject margins are small and rest on 3 seeds.

4.6 A linear readout from the frozen code transfers to new subjects with 2% of ridge's parameters

Adaptation to new subjects from little data matters for translational applications, where multi-hour protocols are often unrealistic [Gordon et al., 2017, Gratton et al., 2020]. To adapt the coupled SAE to a new subject, we fit only an unconstrained linear readout on the frozen code. This readout has 200 weights per voxel (50 components $\times$ 4 delays), against 9,216 for the matched ridge baseline. We test this adaptation on the new subjects UTS07 and UTS08. For each subject, we fit the readout on n training stories, from 0.5 (the first half of one story) to 22, and evaluate it on the test story. The 4 validation stories are held out and voxels are selected as in training (Appendix E). We compare the readout with the matched ridge baseline, fit on the same data, and with the random encoder control. For each n below 22 we draw the n training stories at random seven times per subject and average over the distinct training sets; at $n = 22$ there is one per subject (Appendix E).

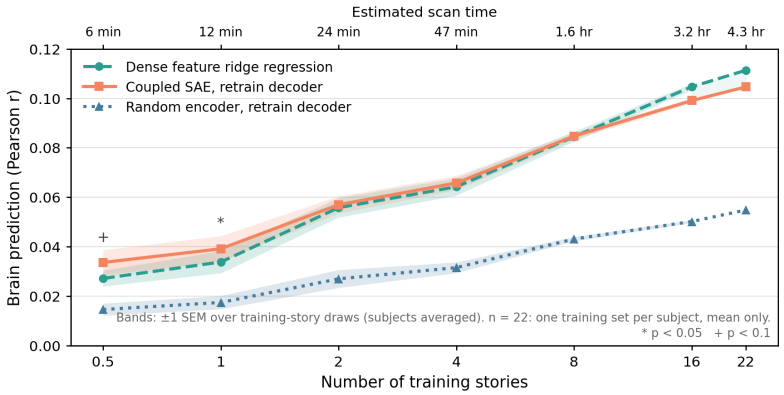


Figure 3: **A linear readout fit on the frozen code transfers to the new subjects UTS07 and UTS08, tracking the matched ridge baseline.** Brain score on the test story against the number of training stories n (log axis; top: estimated scan time) for the coupled SAE (coral), the matched ridge baseline (teal; legend: dense feature ridge regression) and the random encoder control (blue). Lines: mean over distinct training-story draws (Appendix E), subjects averaged within each draw; bands: ± 1 s.e.m. over draws; at $n = 22$ there is one training set per subject. $*$ and $+$: coupled SAE above ridge at $p < 0.05$ and $p < 0.1$ (one-sided paired t-test over draws). The coupled SAE matches or exceeds ridge at $n \leq 1$ (6–12 min), ridge catches up at $n = 8$ and leads at $n \geq 16$; the random encoder stays well below both.

Trained on 22 stories, the readout reaches a brain score of 0.110 on UTS07 and 0.099 on UTS08 (Fig. 3). This is 87% and 102% of ridge (0.126 and 0.097): the frozen code falls 13% short of ridge in UTS07 and exceeds it by 2% in UTS08.

The coupled SAE is above ridge at small n and below it at large n (Fig. 3). This crossover is what a bias–variance trade-off predicts. For a new subject, the frozen readout fits about 200 weights per voxel on a basis learned from six other subjects, whereas ridge fits 9,216 and must estimate the whole mapping from that subject's data alone. With little data the smaller model should generalize better, and with more data the flexibility of ridge should pay off. The data follow this pattern. Pooled over the two subjects, the frozen code matches or exceeds ridge below two stories of data, about 6 and 12 minutes of scanning, with the margin coming mostly from UTS08 (one-sided paired t-test over the 6 distinct training sets, the two subjects averaged within each: $p = 0.065$ at half a story and 0.041 at one story); ridge matches it at 8 stories and leads from 16. Because this pattern rests on two new subjects, we state its extension as a hypothesis: the low-data advantage should grow with the number of subjects the shared basis is trained on, which a larger cohort can test.

5 Discussion, limitations and ethical concerns

We show that one sparse latent basis can be shared across subjects: each component has a natural-language description and a cortical map in every subject. The maps agree across subjects although the model is trained without alignment. A component can therefore be chosen by its description and compared across subjects. This unit of analysis, not improved benchmark scores, is the goal of the work (§1). Because components are only partly stable across seeds (§4.4), we treat them as descriptive units, not a canonical or mechanistic decomposition. Language-model SAEs behave the same way: individual features depend on the seed [Paulo and Belrose, 2026], while the subspace they span is reproducible [Gerasimov et al., 2026].

The study has limits: all results come from one dataset, one language model and layer, and one test story; and the margins between the three models rest on 3 seeds. Other naturalistic datasets, with visual stimuli, different stories or a different pattern of subject overlap, are a natural next test; input-adaptive sparsity such as JumpReLU [Rajamanoharan et al., 2024] could let the number of active components vary with the input. More accurate or cheaper semantic-to-brain models also raise privacy concerns: better inference of semantic content could make sensitive information easier to extract from neural recordings outside governed research or clinical settings.

Acknowledgments and Disclosure of Funding

This work was carried out in part while A.Z. was a postdoctoral researcher at UC Berkeley, supported by the National Eye Institute (NIH grant R01 EY031455), and in part while A.Z. was employed by Erised AI, of which A.Z. is a co-founder and equity holder.

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*, pages 237–248, 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/`.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pages 6077–6101, 2025.

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.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=Ll5miDx8KB`.

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

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. In *International Conference on Learning Representations*, 2025.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders. *arXiv preprint arXiv:2606.12138*, 2026.

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.

Kamya Hari, Taha Binhuraib, Jin Li, Cory Shain, and Anna A. Ivanova. Independent-component-based encoding models of brain activity during story comprehension. arXiv preprint arXiv:2604.24942, 2026. Accepted to CCN 2026 (Proceedings Track).

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.

Evi Hendrikx, Alicia Zeng, Yashaswini, Matteo Visconti di Oleggio Castello, and Jack L. Gallant. A distributed cortical network integrates semantic representations of number, space, and time. In *9th Annual Conference on Cognitive Computational Neuroscience (CCN)*, 2026. Poster.

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. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations*, 2026.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pages 48393–48421, 2025.

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):Article 39, 2010. doi: 10.2202/1544-6115.1585.

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

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.

Taylor Salo, Tal Yarkoni, Thomas E. Nichols, Jean-Baptiste Poline, Murat Bilgel, Katherine L. Bottenhorn, et al. NiMARE: Neuroimaging meta-analysis research environment. *Aperture Neuro*, 3, 2023.

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

David Skrill and Sam V. Norman-Haignere. Component encoding: Interpretable and predictive models of neural computation. In *Conference on Cognitive Computational Neuroscience (CCN)*, 2024.

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.

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 Implementation details

A.1 Architectural design choices

TopK over L_1. We use hard TopK selection [Gao et al., 2025] instead of L_1-penalized ReLU sparsity [Bricken et al., 2023]. TopK keeps at most k components active per TR, so the sparsity level is set directly and is easy to control.

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

Untied semantic decoder. The semantic decoder is initialized as the transpose of the encoder and then trained independently. The transposed initialization follows Gao et al. [2025].

Brain-decoder unit-norm with gradient projection. Before each optimizer step, we remove the radial component of the brain-decoder gradient [Bricken et al., 2023]. After the step, brain-decoder columns are projected to unit norm, one column per component and delay, which fixes the scale ambiguity between a column and the delay-scaled code value it multiplies. The gradient projection keeps optimization from moving parameters in directions that the norm projection would then discard. The brain decoders have no bias term, and responses are z-scored per voxel (Appendix A.3). Brain-decoder weights are also constrained to be non-negative, so a component's delay scales carry the sign of its cortical effect.

No per-voxel scale in the coupled SAE. The architecture supports an optional learnable per-voxel scale on the brain decoder, but the coupled SAE omits it. In the ablation (Appendix B), adding this scale gives a brain score of 0.156 ± 0.010. This is within the seed spread of the coupled SAE, so we keep the simpler model.

AuxK dead-component recovery. A component counts as dead if it has not fired in any of the last 50 batches. Dead components enter an auxiliary loss, with coefficient $1/32$, that reconstructs the residual of the Gemma-2-2b activations through the dead components only [Gao et al., 2025]. The loss uses up to $4k$ dead components. Their pre-activations are computed without gradient, so as implemented the auxiliary loss updates only the semantic-decoder columns. In the reported configuration, no component remains dead at convergence. The AuxK term was used for the coupled SAE, the sweeps of Appendix C and all ablation variants of Table 3. It was not used for the two-stage and random-encoder models of Table 1.

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). Each component has its own learned delay scales, so different components can express different hemodynamic response shapes.

A.2 Gemma-2-2b activations

Backbone. We use Gemma-2-2b (26 transformer layers, hidden size 2,304). For each story we tokenize the transcript and run the model with a sliding context window of up to 8,192 tokens, so the activations at each word can depend on up to 8,192 tokens of preceding 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 the encoder bias does not absorb story-level offsets.

Layer choice. "Layer 25" denotes the output of the final transformer block of the 26-layer model after its final normalization (`hidden_states[-1]`). We use this final layer throughout. We compared only intermediate-to-late layers, because these are known to predict cortical responses to language better than early layers [Schrimpf et al., 2021, Caucheteux and King, 2022]. Among layers 15, 20 and 25, the final layer gives the highest brain score (Table 6).

A.3 Brain data and preprocessing

Subjects and stimuli. We use eight subjects from the natural-listening fMRI dataset of LeBel et al. [2023]. UTS01–UTS06 are used to train and evaluate the coupled SAE, and UTS07–UTS08 are used only in the transfer experiments of §4.6. The amount of training data differs widely across subjects. After excluding the test story, UTS01–UTS03 each contribute 83 stories, UTS04 contributes 25, and UTS05–UTS08 contribute 26 each. All eight subjects heard the test story, and the released preprocessed data store it as the average across repeated presentations. We take brain responses from the dataset's preprocessed release and z-score them per voxel. For UTS04–UTS06, the training matrix is zero-padded for the stories those subjects did not hear, and the padding is applied before z-scoring. Their real time points therefore have a variance of about 3 (2.9–3.1) and a correspondingly larger weight in the brain loss. Test-story correlations are unaffected.

Voxel selection. For each subject, we fit a voxel-selection ridge regression on Gemma-2-2b activations (`himalaya.KernelRidgeCV`; Dupré La Tour et al., 2022). The regression uses the first 10 of the subject's non-test stories in alphabetical order. We rank voxels by their mean leave-one-story-out CV correlation in this regression and retain the top 30,000, or 27–37% of the 81,126–109,469 voxels in each subject's released data (UTS01–UTS06). The remaining voxels have CV correlation near zero (median 0.003–0.014 per subject). For UTS01–UTS06, none of the ten stories is a validation story; for UTS07–UTS08, see Appendix E. Selection never sees the test story.

Voxel-to-fsaverage mapper. For each subject, we precompute a sparse matrix that maps the full brain-mask voxel space to the fsaverage surface (327,684 vertices) using FreeSurfer [Fischl, 2012] `mri_surf2surf`. The mapper is built once per subject and does not depend on any SAE model. The Yeo 7 (Appendix F.4), Neurosynth (Appendix F.5) and cross-subject consistency (Appendix F.3) analyses use it.

Cross-subject transfer (UTS07, UTS08). Data and protocol for the new subjects are given in Appendix E.

A.4 Training and optimization

Optimizer. We train with AdamW [Loshchilov and Hutter, 2019] with zero weight decay and batch size 1024. The learning rate is 10^{-3} after a 5-epoch linear warmup, and gradients are clipped at $\|\nabla\|_2 \leq 1$. A cosine schedule over a nominal 600 epochs is attached, but the coupled SAE stops after 69–105 epochs, so its rate decays by less than 7%. Before each step, the radial gradient component of the brain decoders is zeroed (gradient projection; §A.1). After each step, encoder rows and brain-decoder columns are projected to unit norm, and brain-decoder weights are clamped to be non-negative.

Coupling weight. We report $\alpha = 1$, the equal-weight default, which was not chosen on validation data. For $\alpha = 0.3$, 1, 3 and 10, the validation brain score (UTS01–UTS03, the only subjects with validation stories) is 0.081 ± 0.001, 0.076 ± 0.004, 0.073 ± 0.003 and 0.065 ± 0.002. The corresponding test brain scores are 0.156, 0.162, 0.142 and 0.125 (Table 4). The validation optimum is at $\alpha = 0.3$, and $\alpha = 1$ is within about one seed standard deviation of it. Appendix C reports the full sweep over component count, sparsity, and coupling weight.

Multisubject scheme. For each minibatch, $\mathcal{L}_{\text{brain}}$ is computed for a single subject on the TRs that subject has data for, and $\mathcal{L}_{\text{sem}}$ on every TR. The coupled SAE cycles through the subjects round-robin across minibatches (UTS01 $\rightarrow$ UTS02 $\rightarrow \cdots \rightarrow$ UTS06 $\rightarrow$ UTS01 $\rightarrow \cdots$). The same subject mask also supports an all-subjects-per-batch schedule, in which the mask handles the TRs that are missing because subjects heard different stories (§A.3). Variant F of Table 3 trains the six-subject model with this schedule.

Validation split. Validation uses 4 fixed validation stories, held out from the training stories at the story level (random seed 0). Because the split is by story and not by TR, no validation TR shares its 4-delay input window with any training TR. The four validation stories are drawn from the union of all subjects' stories, and only UTS01–UTS03 heard them. After excluding the test story, the effective training sets are therefore 79 stories for UTS01–UTS03 and all 25, 26 and 26 stories for UTS04,

UTS05 and UTS06. UTS07–UTS08, used in the transfer experiment (§4.6), have 22 stories available after setting aside their own 4 validation stories.

Early stopping. Early stopping monitors the total validation loss with patience 10. For the coupled SAE, Sem→Brain and all sweep and ablation runs, the validation brain loss is unrestricted. It is averaged over all six subjects, with UTS04–UTS06 scored against zero-padded targets for the validation stories they did not hear. For the Brain→Sem models, the validation brain loss is restricted to each subject's own stories, because without a semantic term the unrestricted loss stops brain-only training after about six epochs. When the semantic term is present, training runs for 69–105 epochs for the coupled SAE (68–146 for the ablation and robustness runs). The criterion is still the sum of the two terms, so the unrestricted brain term, including its zero-padded part, also influences the stopping epoch. For the 3 seeds of the coupled SAE the best validation epoch is 59, 62 and 95 and training stops at epochs 69, 72 and 105.

Robustness check: own-story validation and own-time-point z-scoring. We retrained the coupled SAE (AuxK on, 3 seeds, final epoch) with both choices corrected: the validation brain loss is restricted to each subject's own stories, which leaves validation on UTS01–UTS03, and responses are z-scored over each subject's own time points, so the real time points of UTS04–UTS06 have unit variance. Training then stops at epochs 114–146. Together, the two corrections change the brain score by -0.007 (seeds 0.141–0.164), less than the seed std of the corrected runs, and the semantic score by $+0.010$ (Table 2). UTS04 changes by -0.003 and stays above ridge; the largest per-subject changes are for UTS05 and UTS06.

Table 2: **Correcting both the validation criterion and the z-scoring changes the coupled SAE's brain score by -0.007.** Test-story brain and semantic scores (3-seed mean $\pm$ std) and per-subject brain scores (3-seed means) at the final epoch, for the reported coupled SAE and for the model retrained with validation on each subject's own stories and z-scoring over each subject's own time points.

Model	Brain score	Semantic score	UTS01	UTS02	UTS03	UTS04	UTS05	UTS06
Coupled SAE (reported)	0.162 ± 0.005	0.523 ± 0.006	0.152	0.198	0.232	0.130	0.111	0.150
Both corrections	0.155 ± 0.010	0.533 ± 0.001	0.147	0.199	0.223	0.127	0.099	0.133

Checkpoint convention and TR trimming. Every reported score and the component analyses of §4.2–§4.5 use the final-epoch checkpoint. Appendix C and Table 3 follow the same convention. Best-epoch and final-epoch brain scores of the coupled SAE differ by 0.003 on the 3-seed mean and by up to 0.009 for a single seed. When activations are saved, the first 10 and last 5 TRs of each story are dropped. No further TRs are trimmed at training time, so the first four TRs of each story receive delayed copies of the code of an all-zero input.

Reproducibility. Unless stated otherwise, main-text numbers are 3-seed mean $\pm$ std. Aggregate scores vary little across seeds (Table 1). Across 20 seeds of the reported recipe, the final-epoch brain score is 0.151 ± 0.008, or 82% of ridge; this is the retained fraction quoted in the main text. The three reported seeds (0–2) are among the higher ones. Cross-seed stability of individual components is lower and depends on which part of a component is compared. Across 20 seeds at the final epoch, the Hungarian-matched cosine similarity is 0.69 for semantic-decoder columns, 0.66 for cortical maps and 0.45 for encoder rows (null 0.04). Appendix F.6 explains why other settings give somewhat different values. Similar global solutions can therefore be realized by different component-level decompositions.

A.5 Software and compute

The model and analyses are implemented in PyTorch [Paszke et al., 2019]. In our environment, training a single 50-component k=5 model on 6 subjects to early stopping takes 8–15 minutes per run (8.1, 10.6 and 14.2 min for the 3 seeds) on a single RTX A6000-class GPU (48 GB memory). The two-stage models fit a 4 GB GPU when the dataset is held in CPU memory and one batch at a time is moved to the GPU. 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 sends one prompt per model to the LLM judge (Appendix F.1).

Code, configuration files, and scripts to reproduce the main experiments are released at `https://github.com/alicialitrtwe/semantic-cortical-components`, with instructions for data access, preprocessing, training, evaluation, and figure generation. An interactive viewer of all 50 components and their details is at `https://alicialitrtwe.github.io/research/jointsae.html`.

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. The Neurosynth database is distributed under the Open Database License (ODbL 1.0), NiMARE under the MIT license, and the `all-mpnet-base-v2` sentence encoder under Apache-2.0.

B Architecture ablation: three variants are within seed spread on brain score

We compare five architectural variants and one training-schedule variant of the 6-subject model (50 components, k=5, $\alpha = 1$, Gemma-2-2b layer 25). Each variant is trained with 3 random seeds, and we report mean $\pm$ std on the test story. Matryoshka SAEs [Bussmann et al., 2025] are designed to learn nested dictionaries usable at several dictionary sizes. Matching-pursuit SAEs [Costa et al., 2025] are designed to select components sequentially from the residual, to recover hierarchical, conditionally orthogonal features. All variants are trained with the AuxK term and scored at the final epoch. The rows are not perfectly matched: the matching-pursuit variant D also has a tied semantic decoder and no encoder bias. Variant F keeps the architecture of B but computes $\mathcal{L}_{\text{brain}}$ for all subjects in every minibatch instead of cycling round-robin (Appendix A); it was trained with the same settings.

Table 3: **Encoder-row unit-norm TopK (B, the coupled SAE), Matryoshka (C) and B with a per-voxel scale (E) are within seed spread on brain score.** Test-story scores of five architectural variants and one training-schedule variant (F), 3-seed mean $\pm$ std at the final epoch; all rows use AuxK. Bold: best in each column. Variant D also has a tied semantic decoder and no encoder bias.

Variant	Brain score	Semantic score
A. Baseline (no encoder norm)	0.149 ± 0.012	0.518 ± 0.001
B. + Encoder norm (coupled SAE)	$\mathbf{0.162 \pm 0.005}$	0.523 ± 0.006
C. + Matryoshka $[2, 5]$	0.159 ± 0.004	0.528 ± 0.005
D. Matching-pursuit SAE	0.135 ± 0.004	$\mathbf{0.537 \pm 0.003}$
E. B + per-voxel scale	0.156 ± 0.010	0.524 ± 0.005
F. B, all subjects per batch	0.150 ± 0.005	0.527 ± 0.005

Variants B, C and E are within seed spread of each other on brain score. A is lower on average but varies widely across seeds (0.133–0.163), and D is lower. Adding a per-voxel scale (E) does not raise the brain score. The matching-pursuit SAE (variant D) has the highest semantic score, but its brain score is substantially lower. Computing the brain loss for all subjects in every minibatch (F) instead of round-robin (B) gives a lower brain score (0.150 ± 0.005 against 0.162 ± 0.005; the difference is more than twice the seed std of either row, although B's three seeds are among the higher of 20 seeds of the same recipe, whose mean is 0.151) and a semantic score within seed spread (0.527 ± 0.005 against 0.523 ± 0.006).

C Component count, sparsity and coupling weight

We sweep dictionary size $K \in \{25, 50, 100\}$, TopK sparsity $k \in \{2, 5\}$, and coupling weight $\alpha \in \{0.3, 1, 3, 10\}$, holding all other settings at those of the coupled SAE. The Brain→Sem model of Table 1 corresponds to $\alpha = 0$. Every entry of Table 4 is the final-epoch test score of early-stopped training, as in Table 1. The 50-component k=5 row at every α and the α=1 cells at 25 and 100 components use 3 seeds and report mean $\pm$ std. The remaining cells are single-seed. Table 5 extends the sparsity axis to $k \in \{1, 3, 8, 10\}$ at 3 seeds each. The reported model keeps $\alpha = 1$, the equal-weight default (Appendix A); 50 components rather than 100, because they are simpler to

catalog and their cortical maps are more stable across seeds; and k=5 rather than k=10, because k=10 predicts the brain slightly better but its semantic decoder is less stable across seeds. The three points below give the numbers.

Table 4: **At k=2, both scores are lower than at k=5 for every coupling weight.** Entries are brain score / semantic score on the test story (final epoch) over component count, sparsity and coupling weight α. Cells with $\pm$ are 3-seed mean $\pm$ std; the others are single-seed. Bold: largest entry of each score.

Config	$\alpha = 0.3$	$\alpha = 1$	$\alpha = 3$	$\alpha = 10$
25 components, k=5	0.159 / 0.489	0.141±0.008 / 0.504±0.001	0.108 / 0.503	0.130 / 0.513
50 components, k=5	0.156±0.006 / 0.506±0.002	**0.162±0.005** / 0.523±0.006	0.142±0.004 / 0.536±0.002	0.125±0.004 / 0.536±0.002
50 components, k=2	0.119 / 0.451	0.116 / 0.461	0.105 / 0.438	0.089 / 0.427
100 components, k=5	0.140 / 0.521	0.157±0.008 / 0.541±0.003	0.145 / 0.554	0.118 / **0.557**

Table 5: **Brain score rises with k over the tested range and is highest at k=10.** Test-story brain and semantic scores (rows) for each sparsity k (50 components, α=1, final epoch), 3-seed mean $\pm$ std (k=2: single seed).

k	1	2	3	5	8	10
Brain	0.072 ± 0.001	0.116	0.143 ± 0.005	0.162 ± 0.005	0.167 ± 0.004	0.174 ± 0.011
Sem.	0.396 ± 0.007	0.461	0.499 ± 0.003	0.523 ± 0.006	0.554 ± 0.001	0.562 ± 0.001

(1) Dictionary size. At α=1, 100 components match the 50-component model on brain prediction (0.157 ± 0.008 against 0.162 ± 0.005, within seed spread) and gain $+0.018$ semantic score. 50 components are simpler to catalog, and their cortical maps have somewhat higher cross-seed stability: in Appendix F.6, the Hungarian-matched cortical-map cosine is 0.65 against 0.56 at 100 components, and the semantic-decoder cosine is the same (0.64). (2) Sparsity. Brain prediction rises with k over the tested range and is highest at k=10 (Table 5), but beyond k=5 the gain is only about one seed standard deviation (0.167 ± 0.004 at k=8 and 0.174 ± 0.011 at k=10, against 0.162 ± 0.005), whereas semantic-decoder stability is 0.62 and 0.55, against 0.64 (Table 12); the difference at k=8 may be within what 3 seeds can resolve (Appendix F.6). Thus k=5 is not the brain-prediction maximum. At k=2, both metrics are worse at every α. (3) Layer. Table 6 compares layers. The grid is coarse because each configuration is a full multisubject training run.

Table 6: **The final Gemma-2-2b layer (25) gives the highest brain score.** Test-story scores of the reported recipe trained on each layer (final epoch, 3-seed mean $\pm$ std) and the epochs at which the three seeds stopped. All rows use the same 30,000 voxels per subject, selected with layer 25. Semantic scores reconstruct each layer's own activations and are not comparable across rows.

Layer	Brain score	Semantic score	Stop epochs
15	0.145 ± 0.005	0.410 ± 0.001	102, 89, 95
20	0.145 ± 0.002	0.457 ± 0.004	85, 80, 82
25 (reported)	$\mathbf{0.162 \pm 0.005}$	0.523 ± 0.006	69, 72, 105

D Per-subject brain scores: 0.78–1.11× ridge

Per subject, the coupled SAE reaches 0.78–1.11× the matched ridge baseline (Table 7). Table 1 averages these six subjects. The ratio of the 3-seed mean brain scores in Table 7 is 0.88× (the mean of the six per-subject ratios is 0.91×); across 20 seeds the mean brain score is 82% of ridge (Appendix A).

All $\pm$ values, in the mean row and per subject, are std across seeds (as in Table 1). The coupled SAE exceeds ridge only on UTS04, the subject with the fewest training stories. This persists when the extra weight of UTS04–UTS06 in the brain loss is removed (0.127 against 0.117 for ridge; Appendix A.4). The two models do not see exactly the same stories. The matched ridge baseline selects its regularization on four validation stories and is then refit on all of the subject's training stories. For UTS01–UTS03, it therefore also trains on the four stories that the coupled SAE holds out for validation (83 against 79). For UTS04–UTS06, both models use the same stories.

Table 7: **The coupled SAE exceeds ridge only on UTS04, the subject with the fewest training stories.** Per-subject brain score on the test story; n_{train} = training stories of each model. Coupled-SAE values are 3-seed mean ± std. Ridge values are the matched ridge baseline (single fit; no seed variance). Coupled SAE / ridge ratios use the 3-seed coupled-SAE mean; the Mean row gives the ratio of the mean scores.

Subject	n_{train} (coupled SAE / ridge)	Coupled SAE	Ridge	Coupled SAE / ridge
UTS01	79 / 83	0.152 ± 0.013	0.181	$0.84\times$
UTS02	79 / 83	0.198 ± 0.011	0.234	$0.85\times$
UTS03	79 / 83	0.232 ± 0.013	0.295	$0.78\times$
UTS04	25 / 25	0.130 ± 0.011	0.117	$1.11\times$
UTS05	26 / 26	0.111 ± 0.005	0.120	$0.93\times$
UTS06	26 / 26	0.150 ± 0.007	0.161	$0.93\times$
Mean		$\mathbf{0.162 \pm 0.005}$	**0.185**	**0.88**$\times$

E New-subject transfer details

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 retrained for each new subject. This reduces the transfer problem to a linear regression on the fixed 200-dimensional input (50 components $\times$ 4 delays). A full encoding model would instead be fit on 9,216-dimensional dense activations ($2{,}304 \times 4$ delays). Inputs to the frozen encoder are scaled by the RMS of all 83 training stories (11.86) rather than the RMS used in training (11.32); rescaling with 11.32 at half a story changed scores by -0.0003 on average (at most 0.002). Each new subject's four validation stories are the same four used for the main model.

New subjects. UTS07 and UTS08 each contribute 27 stories in total (26 excluding the test story), which leaves 22 training stories after the 4 validation stories are set aside. The encoder never saw these subjects during 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 a linear map without a bias term from the 200 delay-scaled code dimensions to the voxels. Unlike the decoders of the six training subjects, it is not constrained to be non-negative or to have unit-norm columns. It is trained with AdamW (no weight decay), early-stopped on the brain score of the 4 validation stories (patience 10), and scored at the best validation epoch. The delay scales are those of the six-subject model and stay frozen. The four validation stories are never part of a training draw and are not counted in n. Voxel selection (Appendix A.3) uses the first ten of the subject's 26 non-test stories in alphabetical order. One of them is a validation story and the other nine lie in the 22-story training pool, so a training draw can include stories that voxel selection also saw. These ten stories are not counted in n either. Even the $n = 0.5$ fit therefore relies on ten of the subject's stories for voxel selection and four for validation. For each n, we repeat the procedure with seven random seeds per subject. Because the pool holds only 22 stories, different seeds sometimes draw the same training set. We keep each distinct training set once. This leaves 12, 12, 10, 14, 14 and 14 sets pooled over the two subjects at $n = 0.5$–16, and a single set per subject at $n = 22$. The matched ridge baseline uses the same story draws and validation protocol; the random encoder control uses the same protocol with 4–6 distinct draws per n.

Matched ridge baseline. The matched ridge baseline uses the same Gemma-2-2b layer-25 activations with 4 hemodynamic delays, fit with `himalaya.KernelRidgeCV` [Dupré La Tour et al., 2022] using the same validation stories and story draws as the transfer of the coupled SAE. Both methods therefore use the same training data, validation split and 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 for the coupled SAE. Its delay scales are those learned with the random encoder on UTS01–UTS06. This control isolates the contribution of dimensionality reduction and the TopK bottleneck from the learned representation.

F Per-component analyses and catalog

Appendices F.1–F.5 give the per-component analyses of the coupled SAE (naming, semantic coherence, cross-subject consistency of the cortical maps, Yeo 7 network affiliations and Neurosynth term matches), Appendix F.6 their cross-seed stability, Appendix F.7 the five speech-timing components, Appendix F.8 example components and Appendix F.9 the catalog (Table 15). Network affiliations and term matches are computed on the seed-0 model, the model catalogued in Table 15 and shown in the figures, because components are not aligned across seeds: each seed learns its own set, which matches that of another seed only approximately (Appendix F.6).

F.1 Naming (AutoInterp): Brain→Sem components are less nameable

Protocol. For each component we select the 10 highest-activating TRs as training exemplars. We also select 20 held-out items: 10 positives (the TRs ranked 11–20 by activation) and 10 negatives (randomly sampled non-firing TRs). Each exemplar is the text of the firing TR in bold (or `[silence]` for empty TRs) with up to eight words of context on either side. Each exemplar also carries two metadata fields, `pos` (story position, 0–1) and `spr` (syllable count in the firing TR). We also give the component-level mean $\pm$ std of both fields, computed over the component's 20 strongest TRs, which include the ten held-out positives. These values let the judge assess whether positional or rate cues are diagnostic.

All 50 components are scored in a single prompt per model (about 50,000 input tokens). The judge is Claude Fable 5.1 (`claude-fable-5-1`). It runs once per model as a Claude Code subagent with the interface's default sampling settings. All models receive identical prompts and answer keys. The judge (a) writes a one-sentence description of each component, (b) labels each component as `mono` (single coherent concept), `poly` (multiple unrelated triggers) or `dead` (no pattern), and (c) scores each of the 20 held-out items 1–5 for match to the description, without seeing ground-truth labels. We then compute AUC per component from the judge's scores and the held-out answer key.

Brain→Sem components are the least nameable. We run the protocol on all three models of Table 1 at 3 seeds each (9 runs in total). Table 8 reports 3-seed mean $\pm$ std.

Table 8: **Brain→Sem components are the least nameable.** AutoInterp metrics from the Claude Fable 5.1 judge for the three models of Table 1 (3 seeds each). Median AUC over the 50 components; Frac > 0.7 and > 0.9: fraction of components with AUC above the threshold; Mono / Poly / Dead: number of components the judge labels monosemantic, polysemantic or dead. The first three columns are mean $\pm$ std over seeds, and the last is the range over seeds.

Model	Median AUC	Frac > 0.7	Frac > 0.9	Mono / Poly / Dead
Coupled SAE	0.995 ± 0.004	$98\% \pm 0\%$	$89\% \pm 2\%$	48–50 / 0–2 / 0
Sem→Brain	0.992 ± 0.003	$97\% \pm 1\%$	$88\% \pm 2\%$	48–50 / 0–2 / 0
Brain→Sem	0.922 ± 0.012	$93\% \pm 1\%$	$58\% \pm 6\%$	41–48 / 2–9 / 0

AUC is saturated for the coupled SAE and Sem→Brain, so it does not separate these two models. Brain→Sem has a lower median AUC than both in every seed (0.905–0.935 against 0.988–1.000 for the six runs of the other two models). It also has more polysemantic components (2–9 against 0–2). In this comparison, Brain→Sem components are less nameable than those of the two models trained with a semantic objective, although their brain prediction is at least as good. This comparison rests on one judge run per model.

F.2 Coherence: median $z = 4.4$ against a 200-draw null

Semantic coherence measures how tightly the text segments that strongly activate a component cluster in a sentence-embedding space. We compute the metric on the training and validation stories and exclude the test story. For each component j, we take its top-m ($m = 20$) activating TRs and extract the transcript window of ± 5 TRs (about 22 s of speech) around each. We embed each window with the sentence-transformer `all-mpnet-base-v2` and compute the mean pairwise cosine similarity among the m embeddings. For the null distribution, we draw m random TRs from the training and validation stories, build the same windows and compute the same mean pairwise cosine. The

coherence score is the z-score of the observed mean cosine against the mean and standard deviation of 200 such draws (null mean 0.154 ± 0.019 for the seed-0 model). Table 9 compares the three models of Table 1 under the same null procedure. The components of all three models cluster far above chance.

Table 9: **All three models are far above chance on coherence.** Median semantic-coherence z of the 50 components (top-activating text windows against a 200-draw null of random windows), mean $\pm$ std over 3 seeds.

Model	Median coherence z
Coupled SAE	4.4 ± 0.3
Sem→Brain	5.0 ± 0.2
Brain→Sem	4.7 ± 0.3

Across the 50 components of the seed-0 model, median coherence is $z = 4.4$ (4.4 ± 0.3 over 3 seeds). The top-activating windows therefore cluster in sentence-embedding space well above chance. The top windows of a component can overlap in time. Excluding overlapping pairs lowers the seed-0 median z from 4.4 to 3.7, with 40 of 50 components above 2. Without that exclusion, six components of the seed-0 model have $z < 2$. They are the story-opening component (c5), the speech-timing component c48, spoken numbers (c8) and three lexically heterogeneous components (c29, c33, c36).

F.3 Cross-subject consistency of cortical maps: $r = 0.205$ against 0.174–0.177 for ridge

Cortical maps. For each subject s and component j, we form the component's *cortical map* by summing the brain-decoder weights over hemodynamic delays, weighted by the learned delay scales: $d_{s,j}[v] = \left|\sum_\delta [\lambda_\delta]_j \, W^{(s)}_{\text{brain}}[v, \delta, j]\right|$, where v indexes the subject's selected voxels. We place $d_{s,j}$ back into the subject's full voxel space and project it to the fsaverage surface (327,684 vertices) with the subject's voxel-to-fsaverage mapper (Appendix A). We clip to zero the negative values that interpolation introduces.

Cross-subject Pearson r. For each subject pair (s_1, s_2) we use a per-pair intersection mask, which contains the fsaverage vertices that receive a projection from the selected voxels of both subjects. The mask avoids correlating real map values in one subject against structural zeros in another. We then compute the Pearson correlation between $d_{s_1,j}$ and $d_{s_2,j}$ across all intersection vertices for each of the $\binom{6}{2} = 15$ pairs among UTS01–UTS06. Per-component cross-subject consistency is the mean Pearson r across all 15 pairs. We report all values as the mean $\pm$ std over 3 seeds, each an independent training run.

Permutation null. For each of 1,000 permutations, component indices are shuffled independently within each subject's cortical-map matrix. At each component index, the shuffle places one map from each subject, in general from a different component in each subject. We recompute the mean Pearson r over the 15 pairs between these maps. The permutation mean is the average of these values over the 50 indices. The shuffle breaks the component-label correspondence and preserves each subject's cortical-map structure. It therefore gives the distribution expected if component assignments were random. The global p-value compares the observed mean over components with the 1,000 permutation means. The per-component p-value of component j compares its observed mean r with the 1,000 permuted values at index j. Each permuted value is a mean over the 15 pairs for one random pairing of components, not a mean over components. Both p-values are right-tail fractions with the add-one correction of Phipson and Smyth [2010], so the smallest attainable p is $1/1{,}001$. No permutation mean reaches the observed global mean. Per-component p-values are BH-FDR corrected across the 50 components, and components with $q < 0.05$ are counted as significant. The null and the per-component counts are computed on the seed-0 model.

Ridge-projection baseline. For each component j, we project each subject's delay-summed ridge weights onto a direction in activation space that represents the component. This projection gives a direction-matched map on the same fsaverage mask and subject pairs. We use two directions. The first is the component's encoder row, and the second is its unit-normalized semantic-decoder column, which is the direction the component reconstructs. The two directions differ (median cosine

0.55 in the seed-0 model). Cross-subject consistency of these maps is computed exactly as above. We compare the coupled SAE with each baseline by a one-sided paired t-test on 50 values. Each pair is formed within a seed and joins a component's coupled-SAE map with the ridge projection onto that same component's direction. Each of the 50 values is averaged over the 3 seeds at that component index, so components need not be matched across seeds. The test gives $t = 5.2$ with encoder directions and $t = 5.0$ with decoder directions. Under the same BH-FDR test, 19/50 encoder-direction and 34/50 decoder-direction projections are significant, against 42/50 coupled-SAE maps. Under the shuffled-component null the ridge-projection maps have a nonzero null mean ($r = 0.10$ and 0.11), because projections onto correlated directions give correlated maps. All values use the final-epoch checkpoints.

Brain-trained encoders give more consistent maps. We apply the same computation to the cortical maps of the two-stage models. Brain→Sem gives 0.196 ± 0.006 (seed-0 null 0.005 ± 0.004; 45/50 components at $q < 0.05$). Sem→Brain gives 0.177 ± 0.002 (null 0.005 ± 0.003; 49/50), and the coupled SAE gives 0.205 ± 0.008 (42/50). The coupled SAE and Brain→Sem are within about one standard deviation of each other. Both are above Sem→Brain.

F.4 Networks (Yeo 7): 23 DMN and 18 FPN components

Network cosines. To assign each component to the Yeo 7 parcellation [Yeo et al., 2011], we compute in each subject the cosine similarity between the component's fsaverage cortical map (Appendix F.3) and the binary indicator vector of each network. We then average the per-subject cosine profiles across subjects to give the mean profile.

Permutation test (spatial spin). To assess whether each component's network enrichment exceeds spatial chance, we use the spin test of Alexander-Bloch et al. [2018]. For each of 1,000 permutations, we apply independent uniform random rotations from $SO(3)$ (QR decomposition, $\det = +1$ enforced) to the sphere coordinates of the subjects' union analysis mask. Each rotated vertex is mapped to its nearest neighbor within the same hemispheric mask. The rotation moves the cortical maps relative to fixed network boundaries and preserves their spatial autocorrelation. The mask is restricted to selected-voxel coverage (excluding the medial wall), so no rotated vertex lands outside the analysis footprint. Each permutation is applied to all components simultaneously. Per-cell p-values use a one-sided right-tail test with the add-one correction of Phipson and Smyth [2010]. We then apply Benjamini–Hochberg FDR correction at $q = 0.05$ across all $7 \times 50 = 350$ network–component cells.

F.5 Neurosynth: 21 of 50 cortical maps match their term map

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

Meta-analytic maps. For each of the 152 terms, we generate a meta-analytic association map from the Neurosynth v7 database (14,371 studies) with the MKDA Density estimator of NiMARE [Salo et al., 2023]. Each term map is a z-statistic volume in MNI space. We project each map to the fsaverage surface with FreeSurfer `mri_vol2surf` (mid-cortical depth, white surface), which gives a 327,684-vertex map per term.

Component–term matching. We embed the AutoInterp component descriptions (Appendix F.1) and the term names with a sentence-transformer (`all-mpnet-base-v2`; Reimers and Gurevych 2019). Each component j is matched to the term whose embedding has the highest cosine similarity with the embedding of its description. The matching is not unique, so several components may match the same term.

Map correlation. For each component–term pair $(j, t^*(j))$ we compute Pearson r between the component's cortical map, averaged over the six subjects, and the term's z-map (positive z only). Both maps are restricted to one shared mask. It contains the fsaverage vertices that receive a projection

from the selected voxels and where at least one term has $z > 0$. The mask is the same for all components and all terms, so it does not depend on component weights.

Permutation test (spatial spin). We test whether each component's cortical 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, we apply independent uniform random rotations from $SO(3)$ to the two hemispheres of the shared coverage mask. Each rotated vertex is reassigned to its nearest in-mask neighbor. The same rotation is applied to all 152 term maps at once, which preserves the joint null structure. The p-value of a component–term pair is the fraction of permuted Pearson r values that equal or exceed the observed value, with a fraction of zero set to $1/1001$. We apply Benjamini–Hochberg FDR correction across all 50×152 component–term pairs and report the q-value of each component's matched term at FDR $= 0.05$.

F.6 Cross-seed stability: cortical maps and semantic-decoder columns are more stable than encoder rows

Setup. We trained the coupled SAE with 20 random seeds and compared every pair of seeds (190 pairs). We compare four aspects of each component across seeds. The *encoder row* is the direction that detects the component, and the *semantic-decoder column* is the direction it reconstructs. The *activation* is its time series on five probe stories (the test story and four training stories). The *cortical map* is defined in Appendix F.3 and averaged over subjects. For each aspect and seed pair, we find the best-match (maximum) signed cosine of every component in the other model. We report the mean of these cosines (MMCS) and the fraction above $\theta = 0.7$ [Gerasimov et al., 2026]. The reappearance rate $\hat{p}$ of a component is the fraction of the other 19 seeds in which it has a match above θ. We obtain its confidence interval with a seed-level bootstrap. Tables 12, 14 and 11 use one-to-one Hungarian matching [Kuhn, 1955]. The stability value depends on the number of seeds and on the matching. We therefore report the coupled SAE's cross-seed stability in three settings. Hungarian matching over 3 seeds gives 0.65 for the cortical map and 0.64 for the semantic decoder (§4.5, Tables 12 and 14). Hungarian matching over 20 seeds gives 0.66 and 0.69 (Appendix A). The best match over 20 seeds gives 0.71 and 0.73 (MMCS, Table 10). Each table states its setting. All stability analyses use final-epoch checkpoints, except the perturbation analysis of Table 13 (best validation epoch). Table 10 gives the four aspects for the coupled SAE. Table 11 compares architectures on the encoder aspect (3 seeds each), and Table 14 compares training objectives.

Table 10: **Semantic-decoder columns and cortical maps are far more stable across seeds than encoder rows.** 20 seeds of the coupled SAE (190 pairs), signed cosine, final epoch. MMCS = mean maximum cosine; frac> 0.7 = fraction of best-match cosines above 0.7; $\hat{p} \geq 0.7$ = fraction of components that reappear above 0.7 in at least 70% of reseeded models, with a seed-level bootstrap 95% CI. Shuffled nulls (component-label, TR or vertex shuffles) give MMCS 0.05–0.07 for the encoder row, semantic decoder and activation, and 0.28–0.30 for the cortical map. These nulls break temporal and spatial autocorrelation, so they are lower bounds on chance. Table 14 gives a random-encoder control.

Aspect	MMCS	frac> 0.7	$\hat{p} \geq 0.7$ [95% CI]
Semantic-decoder column	0.73	0.61	0.49 [0.47, 0.54]
Cortical map (delay-weighted)	0.71	0.54	0.46 [0.44, 0.50]
Activation	0.58	0.36	0.27 [0.24, 0.31]
Encoder row	0.48	0.16	0.08 [0.06, 0.13]

Output aspects are more stable than encoder rows. The two output aspects (what a component reconstructs and where it maps) have much higher cross-seed stability than the input aspect (what detects it). Semantic-decoder MMCS is 0.73 and cortical-map MMCS is 0.71, against 0.48 for the encoder row. The semantic-decoder aspect needs a caution. A semantic decoder fit on a frozen *random* encoder is as similar across seeds as that of the coupled SAE (Hungarian-matched cosine 0.61 against 0.64, Table 14). This aspect therefore does not by itself show that a code was learned, and shuffled or rotation nulls are too lenient for it. The cortical map separates trained from untrained encoders (0.65 against 0.32), so we rely on this aspect. The encoder aspect (~ 0.4, Table 11) is the

most conservative view. The representation as a whole is highly reproducible (linear CKA 0.94 against 0.49 for a random encoder; 3 seeds, ten training stories, Table 11).

Table 11: **All trained encoders give reproducible representations (CKA 0.88–0.99, against 0.49 for a random encoder).** Mean Hungarian-matched *encoder-row* cosine similarity and linear CKA across 3 seed pairs per model, evaluated on 10 training stories. The encoder row is the aspect with the lowest cross-seed stability (Table 10). Architecture variants A–F correspond to the rows of Table 3. The Brain→Sem row uses the Brain→Sem encoders of Table 14, for which CKA was not computed. For the matching-pursuit variant the encoder is the tied semantic-decoder dictionary.

Model	Hungarian cos. sim.	CKA
B. Coupled SAE (enc. norm)	0.41	0.944
A. Baseline (no encoder norm)	0.33	0.875
C. Matryoshka $[2, 5]$	0.49	0.991
D. Matching-pursuit SAE (tied dictionary)	0.78	0.988
E. + Per-voxel scale	0.45	0.964
F. All subjects per batch	0.45	0.984
Sem→Brain	0.31	0.976
Brain→Sem	0.47	—
Random encoder	0.04	0.485

Per-component stability. For each component, we average its best-match cosine over the 19 reseeded models. The median of this average is 0.77 on the semantic decoder and 0.72 on the cortical map. Of the 50 components, 43 exceed 0.5 on both. The least stable are narrow or story-specific components (c4 dogs and puppies, c19 the Oregon Trail, c1 a story opening about a child's school, c17 therapy settings). The catalog in Appendix F.9 marks with ‡ the seven components below 0.5 on either. Within the seed-0 model, no two components are near-duplicates: across all 1,225 component pairs, the largest semantic-decoder cosine is 0.45 (c3 and c47), and two components with related descriptions, c14 (evaluative summaries of an experience) and c25 (introspective feeling statements), have cosine 0.09 (final epoch). The only near-opposite pairs are the two speech-timing pairs of Appendix F.7.

Stability rises with α; semantic-decoder stability falls at k=10. We vary one training axis at a time, with 3 seeds per setting (Table 12). Stability rises with the coupling weight over $\alpha = 0.3$–10, and semantic-decoder stability is lower at k=10. On the semantic decoder, the differences between dictionary sizes and among k=1–8 may be within what 3 seeds can resolve. For comparison, seeds 0–2 used here give 0.64, whereas the 20-seed value is 0.69. Table 14 gives the $\alpha = 0$ case separately, using the Brain→Sem encoders. The matching-pursuit variant [Costa et al., 2025] is more stable across seeds at some cost in brain prediction, but the main model keeps TopK. Its matched encoder cosine is 0.78, against 0.41 for the reported model (Table 11), and its brain score 0.135 ± 0.004, against 0.162 ± 0.005 (Table 3; both final epoch). Its encoder and semantic decoder are one tied dictionary, however, so the fairer comparison for its 0.78 is the semantic-decoder value of the TopK model (0.64–0.69, depending on the number of seeds), which narrows the gap.

Table 12: **Cross-seed stability rises with α, and semantic-decoder stability is lowest at k=10.** Mean Hungarian-matched cosine to a reseeded model, with 3 seeds per setting and one training axis varied at a time (final epoch).

Axis	Setting	Sem. decoder	Cortical map
α	0.3 / 1 / 3 / 10	0.57 / 0.64 / 0.74 / 0.78	0.61 / 0.65 / 0.67 / 0.69
Components K	25 / 50 / 100	0.68 / 0.64 / 0.64	0.69 / 0.65 / 0.56
Sparsity k	1 / 3 / 5 / 8 / 10	0.70 / 0.69 / 0.64 / 0.62 / 0.55	0.51 / 0.62 / 0.65 / 0.68 / 0.67

Changing the stories moves the components more than changing the subjects. We also retrained the model with the training subjects, the stories or the amount of data changed, one at a time, with 3 seeds per condition (50 components, k=5, $\alpha = 1$; Table 13). Subjects: UTS01–UTS03 or UTS01–UTS02 against UTS01–UTS06, all on the same 21 training stories that all six heard. Stories: two

disjoint draws of 32 training stories (UTS01–UTS03). Data: nested subsets of 32, 16 and 8 training stories against 64 (UTS01–UTS03). Validation stories are fixed within each axis. Every model of one condition is matched one-to-one (Hungarian) with every model of the condition it is compared with, on the semantic-decoder column and on the cortical map (averaged over the subjects both models have). The reference is the same statistic between seeds within the two conditions, so the question is whether a perturbation moves the components more than reseeding does. Changing the subjects gives cosines within 0.03 of the reference; changing the stories at matched data size lowers the cosine (0.51 against a reference of 0.67 on the semantic decoder), and so does less data, monotonically.

Table 13: **Changing the training subjects moves the components about as much as reseeding, whereas changing the stories or reducing the data moves them more.** Mean Hungarian-matched cosine between models of the two conditions, and the seed-only reference within them (3 seeds per condition). A rotation null gives 0.04 on the semantic decoder. These models are scored at their best validation epoch, the only checkpoint saved for them.

	Sem. decoder		Cortical map	
Perturbation	Cosine	Reference	Cosine	Reference
Subjects, 3 vs 6	0.68	0.67	0.64	0.63
Subjects, 2 vs 6	0.66	0.68	0.63	0.66
Stories, disjoint and size-matched	0.51	0.67	0.53	0.66
Half of the data	0.61	0.66	0.60	0.66
A quarter of the data	0.55	0.67	0.55	0.65
An eighth of the data	0.45	0.66	0.47	0.64

Models with a semantic objective have more stable cortical maps. Table 14 compares the three models of Table 1 and the random-encoder control at the final epoch. Cross-seed stability of the cortical maps is about equal for the coupled SAE and Sem→Brain (0.65) and lower for Brain→Sem (0.54). Brain→Sem, the only model without a semantic objective in the encoder, also differs in recipe and stopped earlier, after 51–64 epochs. Brain→Sem components are still far from arbitrary (0.54 against 0.32 for the random encoder). Conversely, cross-subject consistency is higher when the encoder is trained on brain data (0.205 and 0.196) than when it is not (0.177). The coupled SAE is the only model that does well on both measures, although the margins are small (3 seeds).

Table 14: **Only the coupled SAE does well on both cortical-map stability and cross-subject consistency.** 3 seeds, final epoch; the margins are small. Mean Hungarian-matched cosine between seed pairs for three aspects, and cross-subject consistency of the cortical maps (Appendix F.3; mean $\pm$ std over seeds). For Brain→Sem and the random encoder the semantic decoder is fit afterwards on the frozen code.

Model	Sem. decoder	Cortical map	Encoder row	Cross-subject r
Coupled SAE	0.64	0.65	0.41	0.205 ± 0.008
Sem→Brain	0.75	0.65	0.31	0.177 ± 0.002
Brain→Sem	0.59	0.54	0.47	0.196 ± 0.006
Random encoder	0.61	0.32	0.04	—

F.7 Speech-timing components: the five most frequently active fire most strongly on pauses and rapid speech

Five components are active far more often than the others. In seed 0 at the final epoch, c30 is active on 53% of training time points, c21, c47 and c48 on 41% each and c49 on 34%, against 6% for the median component (column Active of Table 15). Components c21 and c48 fire most strongly on rapid speech, and c30, c47 and c49 on silent TRs. The semantic-decoder columns of c21 and c47 have cosine -0.99, and those of c48 and c49 -0.97 (encoder rows -0.90 and -0.68); because the code is non-negative, the two ends of one direction are represented by two components. The delay scales of c30 are negative at all four delays (-1.05, -1.03, -0.88 and -0.95; Appendix F.8). Among the other 45 components, 38 have cross-seed stability above 0.5 for both the semantic-decoder column and the cortical map, and 5 have coherence $z < 2$. Of these 45, 21 are DMN-affiliated, 18 are FPN-affiliated and 16 have a significant Neurosynth match.

F.8 Example components

Fig. 2 in the main text shows four example components (c8, c23, c2, c45). Here we show the other components named in the main text: c5 (§4.2), c20, c28 and c46 (§4.3), and one of the five speech-timing components (c30), together with c24. The other four speech-timing components (c21, c47, c48, c49) are not shown; Table 15 gives their values. Each figure has the layout of one row of Fig. 2 and prints the values of Table 15 (seed 0, final epoch).

Of the 50 learned components, 46 have all-positive delay scales, so their activation predicts an *increase* in brain response. Three components have mixed-sign delay scales with means near zero: c44 (quoted dialogue, mean $= -0.046$), c21 (speech-timing component, mean $= -0.021$) and c48 (speech-timing component, mean $= +0.042$). One component, the speech-timing component **c30** (mean $= -0.980$), has negative delay scales at all four delays (2–8 s), so its activation predicts a decrease in brain response.

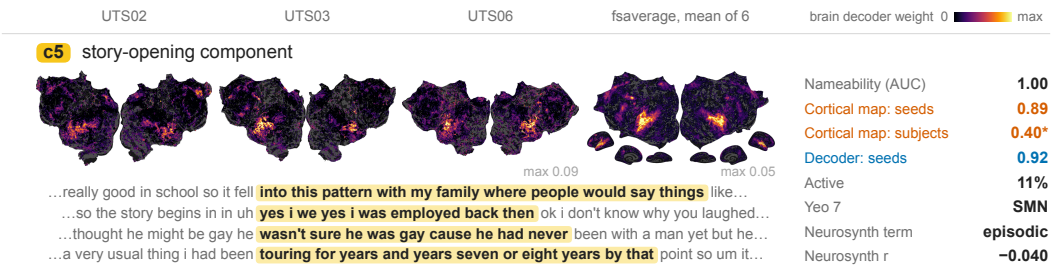


Figure 4: **Component c5 marks story openings, a non-semantic regularity; its somatomotor-affiliated map is stable across seeds (0.89) and has no significant Neurosynth match.** Seed-0 coupled SAE. Title: the judge's description (*story-opening component*). Maps: the component's cortical map (brain-decoder weight; Appendix F.3) in UTS02, UTS03 and UTS06, each on its own flattened cortex, and averaged over the six subjects on fsaverage (flatmap with lateral, medial and ventral views). The three subject maps share one colour scale and the fsaverage map has its own, because the average is weaker than any single subject; the maximum of each scale is printed below the maps. Grey: no value (unselected voxels; on fsaverage also the weakest 5%). Below: four top-activating excerpts, with the firing TR highlighted. Side column: nameability (AutoInterp AUC); cross-seed stability of the cortical map ("Cortical map: seeds"); cross-subject r of the cortical maps ("Cortical map: subjects"; *: $q < 0.05$); cross-seed stability of the semantic decoder ("Decoder: seeds"); active % of time points; significant Yeo 7 networks (SMN = somatomotor, DMN = default mode, FPN = frontoparietal, DAN = dorsal attention, VAN = ventral attention); best Neurosynth term and its map correlation r (*: $q < 0.05$).

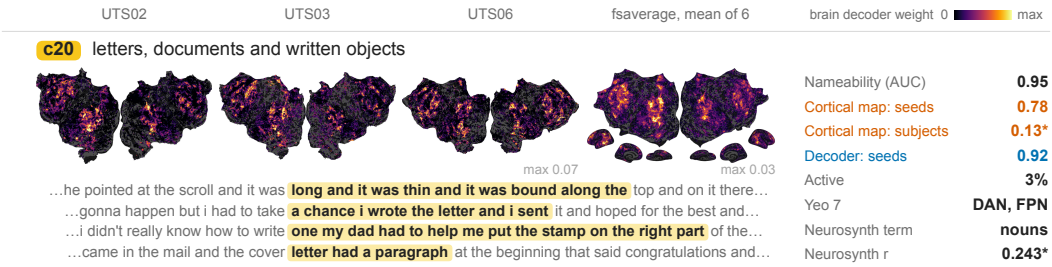


Figure 5: **Component c20 (*letters, documents and written objects*) is a lexical component with a dorsal-attention and frontoparietal map that matches the Neurosynth term "nouns"; its cross-subject consistency is low but significant (0.13).** Layout and side column as in Fig. 4.

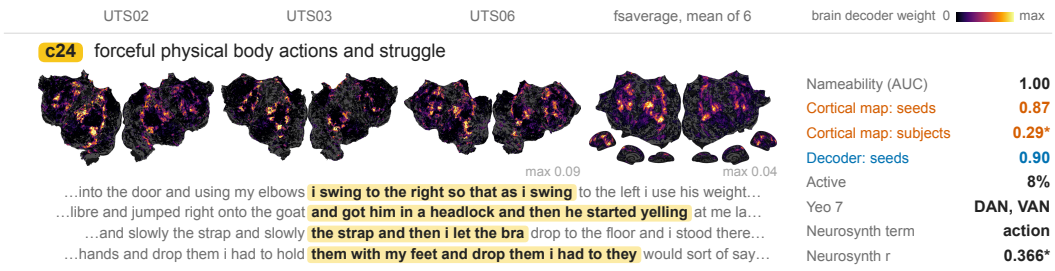


Figure 6: **Component c24 (*forceful physical body actions and struggle*) is a lexical component whose dorsal- and ventral-attention map matches the Neurosynth term "action" ($r = 0.366$).** Layout and side column as in Fig. 4.

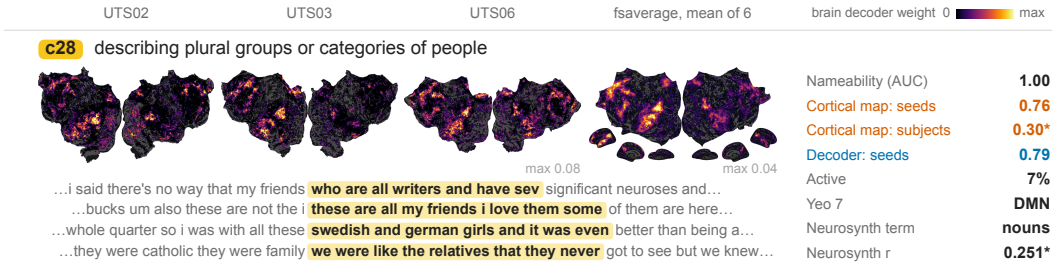


Figure 7: **Component c28 (*describing plural groups or categories of people*) has a default-mode map.** Layout and side column as in Fig. 4.

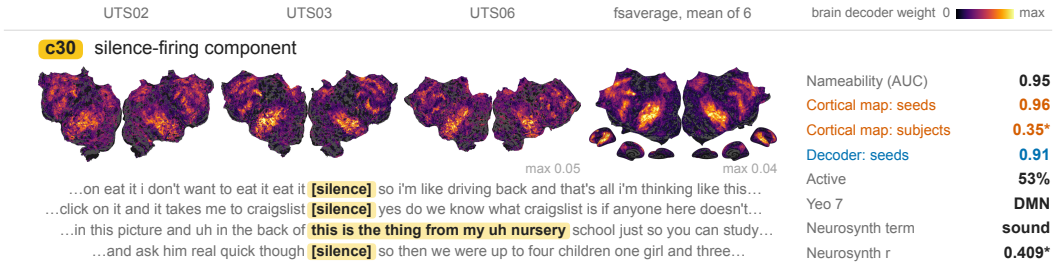


Figure 8: **Component c30 (*silence-firing component*) is a speech-timing component and the only component with all-negative delay scales.** Three of its four excerpts are silent TRs ("[silence]"). Because its delay scales are negative (mean $= -0.980$), its activation predicts a lower brain response; the maps show magnitudes. It is active on 53% of time points (Appendix F.7). Layout and side column as in Fig. 4.

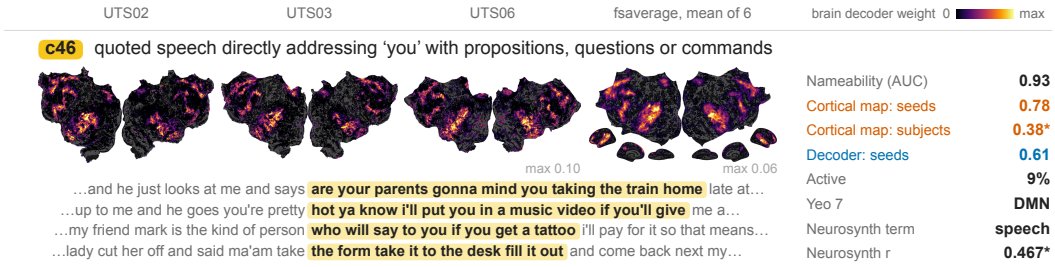


Figure 9: **Component c46 (*quoted speech directly addressing 'you' with propositions, questions or commands*) has a default-mode map that matches the Neurosynth term "speech" ($r = 0.467$).** Layout and side column as in Fig. 4.

F.9 Catalog of the 50 components

Table 15 lists all 50 components of the coupled SAE (seed 0). The interactive viewer at `https://alicialitrtwe.github.io/research/jointsae.html` shows each component in detail. Median AUC is 0.993 (Claude Fable 5.1 judge) and median coherence is $z = 4.4$. The judge labels 49 components monosemantic, 1 polysemantic and none dead.

Table 15: **All 50 components of the seed-0 coupled SAE with their per-component values.** Descriptions and AUC come from the Claude Fable 5.1 judge. We shortened the descriptions by dropping the judge's parenthetical examples; the released CSV has the full text. ‡ marks the 7 components whose best-match cosine, averaged over 19 reseeded models, is below 0.5 for the semantic-decoder column or the cortical map (Appendix F.6). AUC = AutoInterp AUC; z = coherence; NS = matched Neurosynth term, with its map correlation r and FDR-corrected q; Stab. = cross-seed stability of the cortical map; Cross-subj. r = cross-subject consistency of the cortical map (mean Pearson r over the 15 subject pairs; Appendix F.3), seed-0 value at the final epoch; * marks significance against the shuffled-component permutation null at BH-FDR $q < 0.05$ (42 of 50 components); Active = fraction of training time points at which the component is among the 5 active ones (seed 0, final epoch; median 6%); — = no significant network; DMN = default mode, FPN = frontoparietal, DAN = dorsal attention, VAN = ventral attention, SMN = somatomotor. Components c30 (53%), c21, c47, c48 (41% each) and c49 (34%) are the most frequently active. Their strongest moments, as speech-timing components, are pauses and rapid speech (Appendix F.7).

#	Description	AUC	z	Stab.	Cross-subj. r	Active	Yeo 7	NS term	r	q
0	people looking at each other, eye contact and facial expressions	1.00	5.0	0.73	0.25*	4%	DMN	arousal	−0.020	0.813
1‡	story-opening setup: a child's school and family expectations	0.94	8.8	0.47	0.08*	3%	DMN	autobiographical	+0.137	0.050*
2	clothing, outfits and physical appearance descriptions	1.00	9.1	0.87	0.28*	3%	DAN, FPN	color	+0.318	0.006*
3	physical places and venues where things are seen or encountered	0.87	3.3	0.97	0.06	11%	VAN	place	+0.021	0.454
4‡	dogs, puppies and baby animals	1.00	11.9	0.25	0.03	1%	FPN, DMN	categorization	+0.120	0.068
5	story-opening component	1.00	1.0	0.89	0.40*	11%	SMN	episodic	−0.040	0.918
6	place names and geographic locations people live in or move to	1.00	4.9	0.73	0.28*	6%	DMN	location	+0.005	0.682
7	crowds and groups acting collectively and noisily	0.99	6.5	0.58	0.17*	4%	DMN	music	+0.096	0.125
8	spoken numbers and quantities	1.00	1.8	0.91	0.20*	6%	FPN	language	+0.271	0.006*
9‡	film and camera/photography equipment	0.69	10.8	0.51	0.14*	4%	FPN	object	+0.121	0.282
10	religion, faith and cultural/group identity	0.81	3.9	0.58	0.24*	5%	DMN	social	+0.272	0.006*
11	sequences of physical actions by a third person, mostly a man	1.00	6.2	0.78	0.29*	3%	DMN	verbs	+0.139	0.012*
12	abstract reflective emotional language late in stories	0.98	4.5	0.68	0.15*	7%	FPN	emotional	+0.017	0.603
13‡	lists of required tasks and procedures	0.90	3.0	0.46	0.09*	5%	DMN	executive	+0.043	0.356
14	evaluative summary statements of an experience	1.00	2.9	0.58	0.14*	5%	—	autobiographical	+0.001	0.751
15	shared relationship and closeness using 'we'	0.98	9.3	0.63	0.15*	4%	DAN, FPN	familiarity	+0.067	0.116
16	first-person searching, sifting and looking around	1.00	4.5	0.60	0.22*	4%	FPN	multisensory	−0.025	0.841
17‡	therapy, counseling and self-help groups	1.00	10.6	0.41	0.09*	2%	—	stress	+0.025	0.642
18	food and drink items and meals	1.00	7.8	0.79	0.22*	4%	DAN, FPN	food	+0.094	0.158
19‡	Oregon Trail wagon-train scenario	0.95	6.7	0.30	0.06	1%	DMN	navigation	−0.120	1.000
20	letters, documents and written objects	0.95	7.0	0.78	0.13*	3%	DAN, FPN	nouns	+0.243	0.012*
21	rapid-speech component	0.94	5.5	0.98	0.03	41%	DMN	voice	+0.149	0.006*
22	first-person travel and locomotion from place to place	1.00	2.9	0.76	0.28*	7%	FPN, DMN	motion	−0.070	0.986
23	references to family members and personal relations	1.00	3.1	0.72	0.32*	6%	DMN	autobiographical	+0.433	0.006*
24	forceful physical body actions and struggle	1.00	7.2	0.87	0.29*	8%	DAN, VAN	action	+0.366	0.006*
25	introspective first-person feeling/thinking statements	0.98	4.1	0.88	0.22*	22%	DMN	autobiographical	+0.140	0.085
26	navy water-survival/parachute training procedures	1.00	24.3	0.51	0.11*	1%	DAN, FPN	planning	+0.092	0.187
27	spatial descriptions of rooms and built locations	1.00	3.6	0.77	0.35*	5%	DAN	spatial	+0.103	0.118
28	describing plural groups or categories of people	1.00	3.6	0.76	0.30*	7%	DMN	nouns	+0.251	0.047*

Continued on next page

#	Description	AUC	z	Stab.	Cross-subj. r	Active	Yeo 7	NS term	r	q
29	mostly the repeated shouted command 'pick up the gun'; no unified concept	0.84	0.3	0.89	0.09*	8%	SMN	execution	+0.033	0.521
30	silence-firing component	0.95	4.0	0.96	0.35*	53%	DMN	sound	+0.409	0.006*
31	quotative 'I was like / he was like' reported dialogue and reactions	0.87	4.2	0.86	0.37*	12%	DMN	speech	+0.323	0.006*
32	medical procedures and hospital treatment	1.00	7.5	0.50	0.06	3%	FPN, DMN	execution	−0.061	0.987
33	foreign-language phrases, unusual words and proper names	0.85	1.0	0.66	0.31*	17%	DMN	lexical	+0.400	0.006*
34	biographical career/education milestones	1.00	9.0	0.51	0.16*	1%	—	autobiographical	−0.028	0.954
35‡	religious and funeral ritual actions	1.00	12.7	0.40	0.03	2%	FPN, DMN	verbs	+0.165	0.012*
36	narrator framing and transition phrases	0.93	1.0	0.68	0.30*	8%	FPN	syntactic	−0.025	0.784
37	urgent fast-paced first-person movement sequences	0.91	7.1	0.91	0.11*	8%	VAN, FPN	movement	−0.068	1.000
38	unusual tokens	0.99	3.4	0.90	0.24*	17%	—	nouns	+0.378	0.006*
39	describing a person's personality and character traits	1.00	3.8	0.65	0.05	3%	—	affective	+0.070	0.225
40	states of physical or emotional suffering	0.97	4.3	0.64	0.15*	4%	FPN	anxiety	+0.031	0.542
41	first-person self-reflective statements about own tendencies and feelings	0.98	4.1	0.68	0.29*	6%	DAN, FPN	autobiographical	−0.130	1.000
42	habitual or hypothetical 'you would / we would' procedural narration	0.99	4.3	0.54	0.21*	4%	DMN	episodic	+0.124	0.164
43	casual vernacular quoted dialogue with 'he's like / I'm like'	0.98	10.5	0.64	0.21*	7%	DMN	imitation	+0.034	0.567
44	quoted questions, requests and instructions in dialogue	0.92	4.8	0.98	0.01	25%	DMN	speech	+0.305	0.006*
45	natural landscape and scenery descriptions	1.00	8.0	0.78	0.24*	4%	DAN, FPN	imagery	+0.149	0.006*
46	quoted speech directly addressing 'you' with propositions, questions or commands	0.93	3.0	0.78	0.38*	9%	DMN	speech	+0.467	0.006*
47	silence-firing component	1.00	2.6	0.95	0.11*	41%	VAN	sound	+0.191	0.025*
48	rapid-speech component	1.00	1.6	0.98	0.17*	41%	SMN	voice	+0.506	0.006*
49	silence-firing component	1.00	2.5	0.70	0.20*	34%	VAN	sound	+0.256	0.016*

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 new 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 (§5) states the scope limits (one dataset, one language model and layer, one test story), that the components are descriptive and only partly stable across seeds rather than a mechanistic decomposition, and that the margins between the three models rest on 3 seeds. The remaining limits are stated where they arise: coupling costs some brain prediction relative to brain-only training (§4.1); across 20 seeds the coupled SAE retains 82% of ridge, with the three reported seeds among the higher ones (§4.1, Appendix A); AutoInterp saturates for two of the three models (Table 1, §4.5); the low-data margin for new subjects rests mostly on one of the two new subjects (§4.6); and the zero-padding of three training subjects and its effect are reported in Appendix A.4. Other naturalistic datasets remain to be tested.

   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 the appendices, 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 release code, configuration files, and scripts with the paper, together with instructions for data access, preprocessing, training, evaluation, and figure generation; see Appendix A (*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 the mean ± population standard deviation over 3 seeds (20 seeds for the coupled SAE's headline brain score and the stability analyses), the transfer curves report the standard error over distinct training-story draws, and the cortical analyses report permutation-based significance with multiple-comparison correction where appropriate; see §4 and Appendices F.1, F.2, F.5, E, F.3, F.4 and F.6.

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 A (*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: §4.6 notes a practical benefit in translational settings where long scanning sessions are impractical, and the Discussion (§5) notes a potential negative impact in the form of increased privacy risk from inferring sensitive information from neural data.

    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 A (*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 (free license key), `himalaya` (BSD-3-Clause), the Neurosynth database (ODbL 1.0), NiMARE (MIT), and `all-mpnet-base-v2` (Apache-2.0).

    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 code, configuration files, and scripts together with documentation and reproduction instructions; see Appendix A (*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 components admit predictive natural-language descriptions on held-out contexts; see §3, §4, and Appendix F.1.

    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.