INTERPRETING BRAIN RESPONSES TO LANGUAGE WITH SPARSE FEATURES FROM LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A central goal of cognitive neuroscience is to characterize the features that are represented by human language cortex. Artificial language models (LMs) have emerged as a powerful tool to address this challenge, but studies relating biological and artificial representations are often criticized as relating one black box to another. The present work introduces **Augmented Sparse Encoding Models**, an encoding framework that replaces dense LM hidden states with hierarchically-organized sparse autoencoder (SAE) features, while explicitly including surprisal as a predictor. Using this approach, we (i) produce *interpretations* of neural responses and (ii) test whether model-brain alignment reflects primary or idiosyncratic variation in LM representations. Using a high-field 7T fMRI dataset of eight participants listening to 200 linguistically diverse sentences, we first validate our modeling framework by recovering previous interpretations of voxel populations tuned to processing difficulty and meaning abstractness. We then interpret a previously-uncharacterized (but reliable) voxel population and find that it is tuned to people-related content. Next, we show that the fronto-temporal human language network is predicted by a common set of features across its constituent regions, with variable sensitivities to these features. Finally, we show that brain responses are best explained by the features that tend to capture the most general information encoded in LM representations, suggesting a nontrivial correspondence between brain and LM language representation.

1 INTRODUCTION

Humans effortlessly map speech signals to complex meanings through language, but the representations that support this process in the brain remain poorly characterized. Language models (LMs) have become a central tool for probing the representations underlying language processing (Jain & Huth, 2018; Caucheteux & King, 2022). At the same time, this line of work has been criticized as relating one black box (an LM) to another (the brain) without yielding clear claims about what neural populations represent or compute. Recently, two developments make it increasingly feasible to gain scientific insight from LM encoding models. On the LM side, sparse autoencoders (SAEs) provide a tool for decomposing dense LM hidden states into latent features that are often more identifiable and easier to interpret than residual stream features (Bricken et al., 2023). On the neuroscience side, decades of neuroimaging work have delineated a fronto-temporal language network (Binder et al., 1997; Fedorenko et al., 2024), and additional work has made progress in characterizing the response properties of these regions. One line of work shows that responses in language regions track linguistic processing difficulty (as quantified by LM surprisal, for instance; Henderson et al., 2016; Shain et al., 2020; Wehbe et al., 2021; Heilbron et al., 2022; Tuckute et al., 2024), while another shows tracking of concrete vs. abstract meanings during sentence or narrative comprehension (Botch & Finn, 2024; Tuckute et al., 2025; with related work on single-words, e.g., Binder et al., 2005; West & Holcomb, 2000; Fernandino et al., 2015).

In the present study, we leverage these advancements to propose and validate a new class of encoding models: **Augmented Sparse Encoding Models**.[1] Compared to standard LM encoding models, our framework introduces two changes: we project dense residual-stream LM features into a sparse, hierarchically organized SAE basis, and we augment that basis with an explicit feature that captures

[1] Code available here.

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

processing difficulty (surprisal). **Our first goal is to produce interpretations of voxel response tuning.** Next, because the SAE basis is hierarchically organized from general, primary features to fine-grained, idiosyncratic features, it enables us to go beyond claims from prior LM encoding work: whereas they show that LM representations can predict neural responses, they do not characterize the properties of the features that drive alignment. **Our second goal, therefore, is to identify whether brain alignment relies on primary or idiosyncratic LM feature dimensions**.

Using Augmented Sparse Encoding Models, we make the following contributions:

1. SAE features can predict voxel responses to language as accurately as dense LM residual-stream features, while also providing interpretations that affirm and extend prior neuroscience findings on processing difficulty-tuned and content-tuned voxels (Section 4).
2. We discover and interpret previously uncharacterized voxel populations (Section 5).
3. We show how different regions of the human language network respond differentially to processing difficulty vs. content features across three independent datasets spanning sentence and story comprehension. We further characterize the features that are prevalent across brain regions and individuals (Section 6).
4. We show that brain responses to language are predicted by primary, general features of LM representations rather than idiosyncratic ones—revealing a deeper correspondence between artificial and biological language representations than implied by measures of encoding accuracy alone (Section 6).

2 Related Work

Organization of the Language Network A set of frontal and temporal brain areas in the left hemisphere—the "language network"—supports language understanding and production across input modalities (Binder et al., 1997; Deniz et al., 2019; Lipkin et al., 2022; Hu et al., 2023; Fedorenko et al., 2024). Within this system, prior work has emphasized processing difficulty and predictability as robust drivers of univariate response magnitude during comprehension (Shain et al., 2020; Wehbe et al., 2021; Tuckute et al., 2024). A separate line of work has studied abstract–concrete semantic organization in the brain, mostly in single-word experiments (Binder et al., 2005; West & Holcomb, 2000; Fernandino et al., 2015); one recent study links these lines of work by showing that sentence-evoked voxel responses are jointly organized along dimensions of processing difficulty and content (specifically meaning-abstractness) (Tuckute et al., 2025). So, although the location of the language network and its broad response properties are now well established, how linguistic information is represented at a finer grain across voxels—within and across regions, and across individuals—remains less well understood.

Mechanistic Interpretability Efforts to interpret the algorithms and representations learned by trained neural networks have coalesced into a set of techniques and paradigms that comprise the field of mechanistic interpretability (Elhage et al., 2021; Geiger et al., 2025). In particular, the rediscovery of sparse dictionary learning (Olshausen & Field, 1996) has driven progress on unsupervised methods for interpreting and controlling model behavior (Fel et al., 2023; Ameisen et al., 2025; Bricken et al., 2023). Sparse dictionary learning methods, such as SAEs, attempt to learn overcomplete dictionaries of features, such that dense representations can be approximated by a small number of monosemantic (and therefore interpretable) features.

LM Encoding Models Standard LM encoding studies have established that LM representations can predict brain responses to language at the granularity of fMRI voxels, M(EEG) sensors, and intracranial recordings (Jain & Huth, 2018; Caucheteux & King, 2022; Hosseini et al., 2024; AlKhamissi et al., 2025). A growing body of work has aimed to distill and interpret the properties of LM representations that enable LMs to capture neural responses to language. One direction restricts or perturbs the linguistic information (e.g., syntactic vs. semantic information) a model is trained on or can use during inference (Pasquiou et al., 2023; Merlin & Toneva, 2024; Kauf et al., 2024). A second direction isolates specific internal LM components, such as attention weights rather than standard residual-stream embeddings (Lamarre et al., 2022; Kumar et al., 2024). A third direction develops explicitly human-readable or sparse, identifiable feature spaces for voxel-wise encoding models (Benara et al., 2024; Singh et al., 2025; Zeng & Gallant, 2025). Our work is closest to the third direction. Most closely related, Zeng & Gallant (2025) apply sparse dictionary learning to

word-level embeddings in narratives. We differ by (i) using SAEs over contextual LM hidden states, and (ii) leveraging the hierarchical structure of one class of SAE architecture to ask *which levels of the LM feature hierarchy* best align with brain responses. Concurrent and complementary work employs SAEs to study semantic organization in the cortex (Guo et al., 2026) and to dissociate temporal windows and broad cortical regions using intracranial measurements (Kleinman & Goldstein, 2026).

3 METHODS

Brain Data We used brain responses from prior work (Tuckute et al., 2025), in which eight proficient English speakers underwent ultra-high-field 7 Tesla (7T) fMRI while listening to 200 linguistically diverse sentences, each repeated three times in pseudorandomized order. The fMRI blood-oxygenation-level-dependent (BOLD) response amplitude was modeled with a General Linear Model (Prince et al., 2022), yielding a single beta value per voxel per sentence trial (relative to a fixation baseline). We averaged over the three repetitions of a sentence. We restricted analyses to voxels with a noise ceiling signal-to-noise ratio > 0.4 (NCSNR; computed from stimulus repetitions as in Allen et al. 2022), ensuring that voxel selection and subsequent interpretation are based on voxels with reliable responses. Given the left-lateralization of language (Lipkin et al., 2022), all analyses are restricted to the left hemisphere. Analyses were conducted on the fsaverage surface, but for simplicity we refer to surface vertices as "voxels" throughout the paper.

We select voxels of interest in two main ways: one based on two principal components (PCs) of sentence-evoked responses identified by Tuckute et al. (2025), and the other based on whether a voxel is part of the fronto-temporal language network (Fedorenko et al., 2010). For the PC-based selection, we leverage a recent finding that two PCs capture most generalizable variance in sentence-evoked voxel responses (Tuckute et al., 2025; and related work Shain et al., 2020; Wehbe et al., 2021; Botch & Finn, 2024): PC1 reflects *processing difficulty* (characterized by frequency and surprisal) and PC2 reflects concrete–abstract *content* (the degree to which a sentence's meaning is grounded in perceptual experience). Based on each voxel's correlation with these two PCs, we define five populations: **Hard-to-Process** (high PC1), **Easy-to-Process** (low PC1), **Abstract** (high PC2), **Concrete** (low PC2), and **Ghost** (weak loading on both). For the language-network selection, we use five canonical left-hemisphere functional regions of interest (fROIs) defined via a *sentences* $>$ *non-words* contrast (Fedorenko et al., 2010), selecting the top 10% significant voxels within three frontal brain parcels—inferior frontal gyrus (IFG), its orbital portion (IFGorb), and middle frontal gyrus (MFG)—and two temporal parcels—anterior temporal (AntTemp) and posterior temporal (PostTemp) (see Appendix A for details). This dataset—combining 7T resolution with highly reliable sentence-level measurements—provides an ideal testbed for the voxel-level analyses we pursue below. We additionally use two independent 3T datasets for replication, see Appendices H and I.

Augmented Sparse Encoding Models Encoding models are defined by two components: (i) a feature extractor that transforms language stimuli into a feature vector, and (ii) a linear readout that maps that feature vector onto neural responses to the same linguistic stimuli. LM encoding models typically use pretrained LMs as feature extractors, aggregating the intermediate residual-stream representations of an LM to create one feature vector per sentence (See Fig. 1B, Bottom). Linear readouts are typically regularized linear regressions. See individual study sections for details on the linear readouts that we employ.

In this work, we introduce two changes to the typical framework to make the resulting LM encoding model more interpretable. First, we employ pretrained SAEs to transform dense, difficult-to-interpret residual-stream LM representations into a much larger, sparser, and (ideally) more interpretable feature basis. Specifically, we pass LM hidden states (i.e., residual stream representations) into the encoder of an SAE and compute the mean representation over tokens in a sentence in this sparse basis (Fig. 1B, Middle). Our second change is the inclusion of a dedicated feature that encodes the average surprisal (i.e., negative log probability) of a sentence, as computed by an LM (See Fig. 1B, Top). Surprisal is widely used as a measure of sentence processing difficulty, and has been shown to correlate with behavioral and neural signatures of processing difficulty in human participants (Wilcox et al., 2020; Michaelov et al., 2024). Together, these two changes enable us to (i) distinguish whether neural responses are best explained by processing difficulty vs. representa-

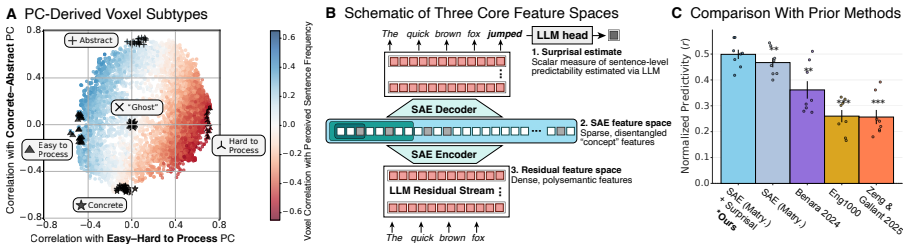

Figure 1: Overview of methods. **(A)** Each dot is a voxel from one sample participant, projected onto two PCs from prior work. Voxels selected for analysis in Study 1 and 2 are annotated. Color denotes each voxel's correlation with sentence-frequency annotations derived from humans (Tuckute et al., 2025). **(B)** Visualization of LM feature spaces. Prior LM encoding models rely on residual stream features, whereas Augmented Sparse Encoding Models use SAE feature spaces along with surprisal, which is measured using the mean of LM output probabilities over tokens in a sentence. **C** Comparison with prior interpretable encoding methods, evaluated on the larger set of language-fROI voxels analyzed in Section 6.

tional content obtained from the LM/SAE, and (ii) attempt to directly interpret the SAE features that explain neural responses.

We instantiate these changes using the `gemma-2-2b` base model (Team et al., 2024), owing to its competitive language performance, relatively small size, and availability of pretrained SAEs. Following prior work on SAEs in gemma models, we primarily use representations from layer 12, though we replicate some of our findings with layer 14 (Appendix B). We analyze two different SAE architectures: JumpReLU and Matryoshka. The JumpReLU SAE is a common, high-performing variation of a standard SAE (Rajamanoharan et al., 2024; Lieberum et al., 2024). The Matryoshka SAE introduces additional structure into the SAE basis to learn more disentangled features at varying levels of granularity (Bussmann et al., 2025; Chanin et al., 2025). Specifically, it is trained in a hierarchical fashion that produces five nested feature sets: the first set can coarsely approximate LM residual states on its own, and each subsequent set is trained to reduce the residual reconstruction error given all earlier sets. As a result, early SAE features capture general, widely-applicable information, and later features capture progressively finer-grained features.

4 Study 1: Validating Augmented Sparse Encoding Models Using PC-Derived Voxel Subtypes

We first evaluate our Augmented Sparse Encoding Models on the four PC-derived voxel subtypes, defined using two brain-derived PCs (Tuckute et al., 2025): one corresponding to processing difficulty (easy-to-process to hard-to-process) and another corresponding to content (concrete to abstract sentence meanings) (Fig. 1A, Section 3). These PC interpretations were obtained via a small set of manually collected (human-derived) annotations. Thus, these voxel populations provide a useful validation setting: because their response properties have already been partially characterized, we can ask whether *interpretations* obtained from Augmented Sparse Encoding Models recover or contradict these prior accounts.

Analysis Methods We wish to identify a sparse set of features that best predict a voxel's response profile in order to gain insight into the underlying features driving that voxel's activity. To do so, we employ a two-stage analysis pipeline comprised of feature selection (using a LASSO regression) and then refitting (using a Ridge regression). We always include the surprisal feature in the Ridge regression, allowing for clean comparisons between full LM encoding model predictivity and a surprisal-only baseline. This pipeline is fully cross-validated (5-fold). The final predictivity score is defined as the mean Fisher-z-transformed correlation across folds, noise-ceiling normalized at the voxel level. See Appendix C for more details.

Finding 1: Sparse SAE features predict voxel responses as well as dense residual-stream features. From Fig. 2A, we see that SAE features tend to predict voxel responses to language stimuli nearly as well as dense LM residual features. In particular, features in the Matryoshka SAE basis

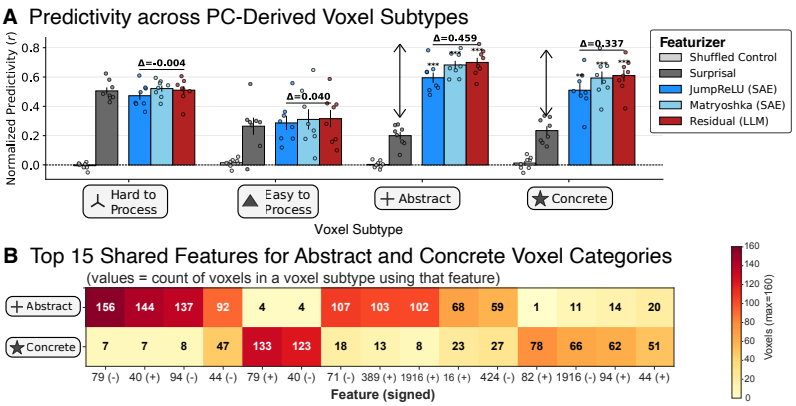

Figure 2: **(A)** Normalized encoding model predictivity of four voxel subtypes using different feature spaces. Processing-difficulty voxels show no predictivity benefit from LM representation features beyond surprisal, whereas abstract and concrete voxels benefit substantially. **(B)** Matryoshka SAE features selected by regressions on **Abstract** and **Concrete** voxels. Features form different subspaces for each subtype generalize across individual brains, and correspond to features that match—and extend—existing interpretations of these voxel populations.

achieve very similar performance as features in the dense residual LM basis, across all four voxel subtypes (on average, 18 Matryoshka SAE features were selected per voxel; Appendix D). The JumpReLU basis consistently yields feature sets that are somewhat worse for prediction. All feature spaces outperform a control that attempts to predict voxel responses from mismatched LM features (Hadidi et al., 2026). Compared to prior methods, our Augmented Sparse Encoding Models outperform several prior interpretable feature spaces (Benara et al., 2024; Huth et al., 2016; Zeng & Gallant, 2025) (Fig. 1C; See Appendix C for details).

Finding 2: Dissociation between processing difficulty-driven and content-driven voxel regimes. Fig. 2A shows a clear dissociation across the PC-derived subtypes: voxels in the **Hard-to-Process** subtype are captured by surprisal alone, with no significant improvement from SAE or residual-stream features. In contrast, voxels in the **Abstract** and **Concrete** subtypes benefit significantly from the inclusion of SAE or residual-stream features, above and beyond surprisal (Fig. 2A, $p < 0.001$). This finding recovers—at the single-voxel level—a dissociation between processing difficulty-driven and content-driven voxel populations during language comprehension. Consistent with prior work that summarized voxel responses via PCA or region-level averaging (Tuckute et al., 2025, also e.g., Wehbe et al., 2021), we show this here using Augmented Sparse Encoding Models.

Finding 3: Distinct features refine interpretations of content-driven voxels. We next ask which specific model features support predictivity of the content-driven voxel subtypes. SAE features can often (but not always) be interpreted by examining the natural-language contexts that maximally activate them, using resources such as Neuronpedia (Team, 2024). For this analysis, we focus on the Matryoshka SAE, as it (i) better matched the predictivity of the residual-stream feature space, (ii) is known to produce more interpretable features than JumpReLU SAEs (Bussmann et al., 2025), and (iii) results in more sparse feature bases than the JumpReLU SAE (Appendix D). For each voxel, we refit the encoding model on all sentences and examine the (signed) SAE features selected by the model. Because this feature basis is learned without supervision and contains tens of thousands of candidate features, agreement with prior interpretations would be nontrivial.

From Fig. 2B, we see that **Abstract** and **Concrete** voxels are consistently supported by different feature subspaces. Moreover, many of these features are selected *across* all eight participants (e.g., among the 160 voxels analyzed across eight participants—20 per subtype—feature 79 is selected in 156/160 abstract voxels; Fig. 2B). Thus, a small set of signed features is *consistently* selected across individual brains. See Appendix F for a robustness analysis. From Table 1 it is evident that the interpretations of these features broadly align with existing interpretations (Tuckute et al., 2025). For example, **Abstract** voxels are driven by features related to emotions (Feat. #40+; 144/160 voxels), and suppressed by features related to scenery (Feat. #79−; 156 voxels).

Feat. #	Feat. Label	Annotation	Max-Activating Examples
94	People	Descriptions of public figures involved in events	*- She worked under execute chef Walter Scheib III, who resigned...* *- Singer and former Haitian presidential candidate, picture here...*
79	Scenery	Descriptions of locations and scenes	*- lost control and crashed into a ditch* *- apartment complex's newly created dog park*
40	Emotions	Informal speech related to emotions and relationships	*- oh shit cause I still have mad feelings for this girl right?* *- Yeah I'm bummed. To the max.*
49	Pronouns	Possessive pronouns.	*his*, *their*, *our*, *my*
44, 71	Short Sentences	Greater activation on sentences with few tokens.	*This feature activates on the first token. We take the mean of SAE activations, so this tracks the number of tokens in a sentence.*

Table 1: Descriptions of Matryoshka SAE features that often predict voxel responses. Feature # corresponds to the dimension in the Matryoshka latent space. Feature Label and Annotation are manually annotated summaries of the semantics of the feature. Max-Activating examples demonstrate the natural-language contexts that most activate these features, from Neuronpedia (Team, 2024). See Appendix K for descriptions of more features.

Furthermore, this analysis also serves to nuance existing interpretations of these voxel subtypes. By characterizing voxels based on their projection along a PC, prior work was limited to looking for high-level, symmetric interpretations that cleanly contrast voxel populations at both ends of a PC. By attempting to create voxel-level interpretations, our encoding models can reveal subtle asymmetries that were previously obscured. For example, we find an asymmetry when investigating a people/event-related feature (Feat. #94), which suppresses **Abstract** voxels but is *not* very reliable for predicting **Concrete** voxels. Thus, what might appear to be a symmetric contrast when performing a PC-based analysis is, at the voxel level, asymmetric: some features drive one population without driving its counterpart to the same degree.

5 Study 2: Identifying and Interpreting Uncharacterized Voxel Populations

Having validated Augmented Sparse Encoding Models in Section 4, we now present a case study of deriving novel hypotheses about the tuning of previously uncharacterized voxel populations. We use the voxels present in the **Ghost** subtype, as described in Section 3. This population shows stable responses to language (i.e., high noise ceilings), but is not captured by the two main organizing PCs (Fig. 1A). In this section, we use Augmented Sparse Encoding Models to tackle both the problem of *identifying* coherent voxel subpopulations and then *interpreting* them.

Analysis Methods To identify subpopulations of voxels that are predicted by a coherent set of features within the **Ghost** subtype (G), we assess how well features that are found to predict one voxel (a *source* voxel, s) can generalize to predicting another voxel (a *target* voxel, t). For every $s \in G$, we identify a signed sparse feature basis that predicts s using all linguistic stimuli. Then, for every $t \in G$, we run a cross-validated and sign-constrained Ridge regression to predict t using exactly the set of (signed) features that are used to predict s, ensuring that features are being used in "the same way" for both the source and target voxels (i.e., directional feature agreement). See Appendix C for further analysis details.

Finding 1: A subset of the Ghost voxels form a coherent feature-defined subpopulation. Overall, many voxels in the **Ghost** subtype are not well predicted by surprisal or LM/SAE representations (Fig. 3A, and Appendix E). However, one voxel subpopulation stands out as an exception: using the feature generalization analysis described above, we identify a coherent subpopulation of 32 voxels (Fig. 3A, red square). Within this subpopulation, signed features that predict one voxel also generalize to other voxels, consistent with a shared underlying feature basis.

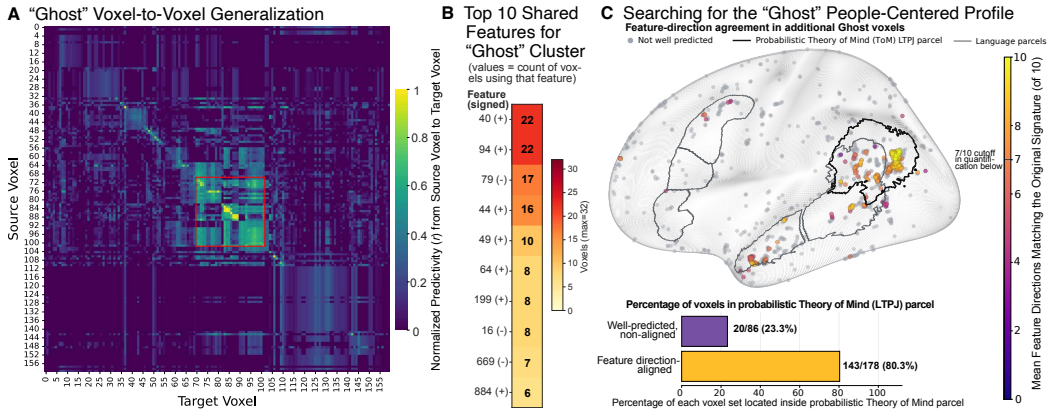

Figure 3: **(A)** Generalization heatmap for "Ghost" voxels (20 per participant, 160 in total), testing whether signed Matryoshka SAE features used to predict one voxel (source) generalize to predicting other voxels (target). Voxel order is determined by a hierarchical agglomerative clustering algorithm to cluster voxels with similar generalization profiles. Mutually well-predicted voxels are denoted by the red square. **(B)** Matryoshka SAE features selected by regressions on the red-square voxels in panel A. These features correspond to a "people-specific" voxel tuning. **C** We fixed the 10 signed features from panel A (i.e., the original Ghost cluster) and searched 1,490 additional Ghost voxels across participants. The surface map (group-average fsaverage7; see individual maps in Appendix G) shows well-predicted voxels colored by feature-direction with the 10-feature profile in panel B. The barplot quantifies the percentage of feature direction-aligned and well-predicted non-aligned voxels located within probabilistic theory of mind functional LTPJ (Dufour et al., 2013).

Finding 2: The coherent Ghost subpopulation is tuned to people-centered content. To interpret this voxel subpopulation, we fit Augmented Sparse Encoding Models to each voxel and examined the signed Matryoshka SAE features selected by the encoding models (i.e., the same analysis procedure as in Section 4). We find that the features used to predict this voxel subpopulation are typically driven by sentences that include "people-centered" features, such as discussions of relationships/emotions (Feat. #40+), descriptions of people taking concrete actions (Feat. #94+), or pronouns (Feat. #49+). Next, we asked whether this feature profile extends beyond the original cluster. We fixed this cluster's 10 signed features (Fig. 3B) and searched 1,490 additional voxels satisfying the **Ghost** criteria. Of these voxels, 275 were well-predicted by a ridge regression using the 10 features (NC-normalized cross-validated threshold of $r > 0.5$). We visualized these voxels on the brain surface, coloring each voxel by the number of feature directions (out of 10) that matched the original signature (Fig. 3C). Strikingly, highly aligned voxels were strongly concentrated in the left temporoparietal junction (LTPJ)—a region implicated in Theory of Mind reasoning Saxe & Kanwisher, 2003; Miao et al., 2026). For quantification, we used the probabilistic Theory of Mind LTPJ parcel from Dufour et al. 2013 and asked what proportion of feature-aligned voxels (defined as matching the original signature for at least 7 of the 10 feature directions, in five out of eight participants with more than three feature-aligned voxels) fell within this parcel: 80.3% fell within the LTPJ parcel, compared with 23.3% of well-predicted but non-aligned voxels. We emphasize that the signature generalizes beyond the original cluster to multiple participants despite no anatomical constraint on voxel selection, demonstrating how Augmented Sparse Encoding Models can recover coherent and interpretable representational structure shared across a subset of individuals.

6 Study 3: Characterizing the Feature Basis of the Fronto-Temporal Language Network

We next analyze the five canonical left-lateralized language regions (Fedorenko et al., 2024), using Augmented Sparse Encoding Models to ask whether the language network implements a largely shared representational code, or whether differences exist between regions and/or individuals.

Finding 1: Language regions show mixed tuning to processing difficulty and content, but not uniformly across regions. First, in line with Section 4, we find that Matryoshka SAEs still provide

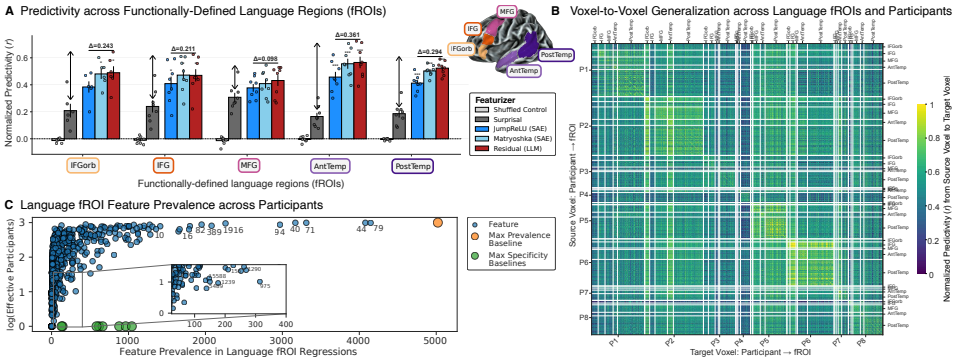

Figure 4: **(A)** Normalized encoding model predictivity of five language fROIs. **(B)** Generalization heatmap for language fROI voxels (2,296 voxels in total across eight participants), testing whether signed Matryoshka SAE features used to predict one voxel (source) generalize to predicting other voxels (target). Voxel order is sorted according to participant and fROI. **(C)** Quantification of the prevalence of features across participants. Each point is a feature (828 in total); higher values indicate broader sharing across individuals. Orange shows the theoretical maximum-prevalence baseline (a feature in every voxel regression), and green shows maximum-specificity baselines for features confined to a single participant (a feature in *all* regressions from exactly *one* participant).

a better feature basis than JumpReLU SAEs, and are on par with residual-stream features (Fig. 4A). Second, frontal fROIs are relatively well-captured by surprisal alone; in particular, the middle frontal gyrus (MFG) does not show a significant gain from SAE/residual-stream features beyond surprisal. In contrast, temporal fROIs show a significant boost from feature-based models above and beyond surprisal, indicating stronger sensitivity to content-linked structure during sentence comprehension (Fig. 4A, $p < 0.001$). We replicated these findings in two independent datasets with different participants, different MRI field strength, different stimuli, and experimental paradigm (Appendices H and I).

Finding 2: Language-network voxels draw on a common feature basis, with graded participant-specific variation. Next, we ask whether the human language network is organized around a shared feature basis or more sharply separated regional and participant-specific feature spaces. First, we find that the voxel-to-voxel generalization heatmap shows substantial transfer between feature bases across the language network (Fig. 4B), in stark contrast to the **Ghost** voxel-to-voxel heatmap (Fig. 3A). More specifically, we find that the heatmap does not reveal substantially higher transfer within-fROI than cross-fROI (within-fROI predictivity = 0.51 vs. cross-fROI predictivity = 0.50). This finding indicates substantial feature sharing across language regions, rather than region-specific feature bases. See Appendix J for further analysis details.

Second, we also find substantial transfer across *individuals*. Specifically, predictivity within-individuals is slightly higher than across individuals (within-participant predictivity = 0.61 vs. cross-participant predictivity = 0.49). To understand the feature sharing across participants further, we quantify how broadly individual SAE features are shared across individuals by plotting each feature's prevalence across language-voxel regressions against the entropy of its distribution across participants (Fig. 4C). This prevalence analysis revealed four regimes: a small number of features are extremely prevalent and broadly shared across all eight participants (upper-right portion of the plot; See Table 1 and Appendix L for interpretations). Many substantially less prevalent features are also reused across all participants despite appearing in far fewer voxel regressions. This suggests that the same type of heterogeneity *within* individual brains occurs *across* participants. A broad middle regime contains features with nontrivial prevalence reused across multiple participants, while a smaller low-entropy tail suggests participant-specific features (inset in Fig. 4C).

What are these shared features? The two most prevalent features are Feat. #79 and #44—Feat. #79 is content-linked, responding to locations and scenes, whereas Feat. #44 (and Feat. #71, another prevalent feature) appears to be driven by contexts with fewer tokens. Because each sentence contains the same number of words, this feature implicitly captures whether a sentence contains infrequent words that are divided into a larger number of tokens. These features may reflect additional form- or

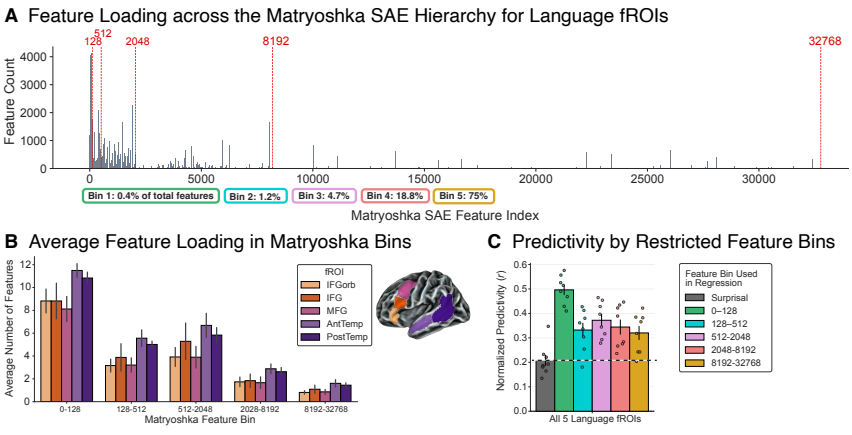

Figure 5: Augmented Sparse Encoding Models preferentially rely on general features to predict voxels across language fROIs, despite making up only 0.4% of all features. **(A)** Histogram of how often each Matryoshka features (ordered by feature index) are selected by the encoding models. **(B)** Average number of features per Matryoshka bin. **(C)** Performance of encoding models with feature sets restricted to individual Matryoshka feature bins.

difficulty-linked features beyond average sentence-level surprisal alone, as surprisal is explicitly included in all models. Other prevalent features captured people/emotions (Feat. #40), people/actions (Feat. #94), questions (Feat. #389), consequences/statements (Feat. #1916), and technical or specialized vocabulary (Feat. #82) (see Table 1). These widely-shared features represent only a portion of a much richer feature basis: in total, 828 unique (unsigned) features were selected at least once across language-network voxels. In summary, these results argue that the entire language network is organized around a shared, dominant feature set consisting of both content- and form-linked features. At a single-voxel level, this code consists of a small cross-participant core alongside a broader regime of features reused across only *subsets* of participants.

Finding 3: Human language network voxels are preferentially explained by general, widely-applicable LM features. Finally, we ask where brain-predictive features fall within the Matryoshka SAE hierarchy. As described in Section 3, the Matryoshka SAE imposes an ordered feature hierarchy: early latent features account for general, widely-applicable variance in LM representations (Bussmann et al., 2025). We find that language-network encoding models preferentially recruit the most general Matryoshka features (feature indices < 128; Fig. 5A,B,C, Appendix M; see also Appendix I). These findings indicate that brain-alignment with the human language network relies on the LM's general, widely-applicable features rather than with fine-grained, idiosyncratic features. These findings are replicated using the PC-derived voxel subtypes (Appendix M) and are not driven by a lack of active features in the finer-grained bins (Appendix N). These findings extend "standard" LM-encoding claims from predictivity to more specific correspondences between the properties of features that organize biological and artificial language systems.

7 Discussion

Augmented Sparse Encoding Models move beyond prior approaches that treat both models and brains as black boxes. Our empirical results show that small sets of features can be used to predict brain responses to language and separate processing difficulty-related signal from content. Within the language network, tuning is shared but not uniform: regions differ in their sensitivity to processing difficulty versus content, and content-linked features range from those broadly shared across individuals to individual-specific features. For extended discussion of language network organization, the "Ghost" population, and the relationship between LM feature granularity and alignment, see Appendix O. Our interpretations are constrained by the SAE feature basis, and some features remain opaque. However, other common critiques of SAEs (e.g., lack of causal claims; Wu et al., 2025) do not apply, as we merely require an unsupervised, interpretable feature basis. Other limitations stem from the brain data, as voxels encompass hundreds of thousands of neurons; it is possible that the more general Matryoshka SAE features may partly reflect this coarse spatial scale. However, *any* coarse feature is not inherently brain-predictive.

AI USE STATEMENT

In this work, we used generative AI tools for designing and providing feedback on research methodology or experiments, implementing methods, and reformatting datasets. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, supporting qualitative or thematic analysis, or interpreting results. Assisting in the writing of proofs and assisting with translation are not applicable to this work.

Additionally, we used generative AI tools for creating and modifying scientific figures, suggesting experimental hyperparameters, editing code, and brainstorming. We have reviewed all AI-assisted work. Specifically, we manually reviewed and tested all code for correctness and cross-checked the results in each figure against the corresponding raw values. We did not use AI tools to draft the original manuscript draft. We only used AI tools at a review-stage for writing-related edits and suggestions. We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

ETHICS STATEMENT

All human neuroimaging data used in this work were de-identified. No raw participant data or personally identifiable information are shared as part of this study. In addition to the primary 7T dataset, we analyzed two previously published, publicly available datasets that were distributed in de-identified form. Behavioral ratings were obtained from previously published work and were analyzed only in aggregate form, without information linking responses to individual participants. All human-participant data used in this work were collected under protocols approved by the relevant institutional review boards or ethics committees, with informed consent obtained from participants.

REPRODUCIBILITY STATEMENT

In support of reproducibility, we provide all code and anonymized data for reproducing our findings, as well as extended details on voxel selection (Appendix A) and all methods (Appendix C).

REFERENCES

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. From language to cognition: How llms outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24332–24350, 2025.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in neural information processing systems*, 34:8332–8344, 2021.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fMRI. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4533e4a352440a32558c1c227602c323-Paper-Conference.pdf.

Richard Antonello, Chandan Singh, Shailee Jain, Aliyah Hsu, Sihang Guo, Jianfeng Gao, Bin Yu, and Alexander Huth. Generative causal testing to bridge data-driven models and scientific theories in language neuroscience. *arXiv preprint arXiv:2410.00812*, 2024.

Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019. URL https://www.science.org/doi/10.1126/science.aav9436.

Vinamra Benara, Chandan Singh, John X Morris, Richard J Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. Crafting interpretable embeddings for language neuroscience by asking llms questions. *Advances in neural information processing systems*, 37:124137, 2024.

Jeffrey R Binder, Julie A Frost, Thomas A Hammeke, Robert W Cox, Stephen M Rao, and Thomas Prieto. Human brain language areas identified by functional magnetic resonance imaging. *Journal of neuroscience*, 17(1):353–362, 1997.

Jeffrey R Binder, Chris F Westbury, Kristen A McKiernan, Edward T Possing, and David A Medler. Distinct brain systems for processing concrete and abstract concepts. *Journal of cognitive neuroscience*, 17(6):905–917, 2005.

Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. A functional dissociation between language and multiple-demand systems revealed in patterns of bold signal fluctuations. *Journal of neurophysiology*, 112(5):1105–1118, 2014.

Thomas L Botch and Emily S Finn. Neural representations of concreteness and concrete concepts are specific to the individual. *Journal of Neuroscience*, 44(45), 2024.

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

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. URL https://www.nature.com/articles/s42003-022-03036-1.

David Chanin, Tomáš Dulkáš, Adria Garriga-Alonso, and Johnny Lin. A bunch of matryoshka saes, 2025.

Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, and Evelina Fedorenko. Multilingual computational models reveal shared brain responses to 21 languages. *bioRxiv*, pp. 2025–02, 2025.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

Nicholas Dufour, Elizabeth Redcay, Liane Young, Penelope L Mavros, Joseph M Moran, Christina Triantafyllou, John DE Gabrieli, and Rebecca Saxe. Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PloS one*, 8(9):e75468, 2013.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194, 2010. URL https://journals.physiology.org/doi/prev/20100421-aop/pdf/10.1152/jn.00032.2010.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203: 104348, 2020.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312, 2024. URL https://www.nature.com/articles/s41583-024-00802-4.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2711–2721, 2023.

Leonardo Fernandino, Colin J Humphries, Mark S Seidenberg, William L Gross, Lisa L Conant, and Jeffrey R Binder. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia*, 76:17–26, 2015.

Bruce Fischl, Martin I Sereno, Roger BH Tootell, and Anders M Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4):272–284, 1999.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. The natural stories corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. Sparse autoencoders map brain-llm alignment onto cortical semantic topography. *arXiv preprint arXiv:2605.23035*, 2026.

Nima Hadidi, Ebrahim Feghhi, Bryan H Song, Idan A Blank, and Jonathan C Kao. Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. *Nature Communications*, 2026.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.

John M Henderson, Wonil Choi, Matthew W Lowder, and Fernanda Ferreira. Language structure in the brain: A fixation-related fmri study of syntactic surprisal in reading. *NeuroImage*, 132: 293–300, 2016.

Eghbal Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024. URL https://www.biorxiv.org/content/10.1101/2024.12.26.629294v1.abstract.

Jennifer Hu, Hannah Small, Hope Kean, Atsushi Takahashi, Leo Zekelman, Daniel Kleinman, Elizabeth Ryan, Alfonso Nieto-Castañón, Victor Ferreira, and Evelina Fedorenko. Precision fmri reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, 33(8):4384–4404, 2023.

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

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical-semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *Neurobiology of Language*, 5(1):7–42, 2024.

Timna Wharton Kleinman and Ariel Goldstein. Back to the feature: Toward a feature-centric account of brain–lm alignment. In *ICLR 2026 Workshop on Representational Alignment (ReAlign)*, 2026.

Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15(1):5523, 2024. URL https://www.nature.com/articles/s41467-024-49173-5.

Mathis Lamarre, Catherine Chen, and Fatma Deniz. Attention weights accurately predict language representations in the brain. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4513–4529, 2022.

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*, pp. 278–300, 2024.

Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, et al. Probabilistic atlas for the language network based on precision fmri data from> 800 individuals. *Scientific data*, 9(1):529, 2022.

André Longon, David Klindt, and Meenakshi Khosla. Superposition disentanglement of neural representations reveals hidden alignment. *arXiv preprint arXiv:2510.03186*, 2025.

Yifan Luo, Yang Zhan, Jiedong Jiang, Tianyang Liu, Mingrui Wu, Zhennan Zhou, and Bin Dong. From atoms to trees: Building a structured feature forest with hierarchical sparse autoencoders. *arXiv preprint arXiv:2602.11881*, 2026.

Ziming Mao, Jia Xu, Zeqi Zheng, Haofang Zheng, Dabing Sheng, Yaochu Jin, and Guoyuan Yang. Sparse autoencoders bridge the deep learning model and the brain. *arXiv preprint arXiv:2506.11123*, 2025.

Monika S Mellem, Kyle M Jasmin, Cynthia Peng, and Alex Martin. Sentence processing in anterior superior temporal cortex shows a social-emotional bias. *Neuropsychologia*, 89:217–224, 2016.

Gabriele Merlin and Mariya Toneva. Language models and brains align due to more than next-word prediction and word-level information. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18431–18454. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.1024/.

Zizhuang Miao, Heejung Jung, Philip A Kragel, Ke Bo, Patrick Sadil, Martin A Lindquist, and Tor D Wager. Common and distinct neural correlates of social interaction processing and theory of mind in narratives. *Nature Communications*, 2026.

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of language*, 5(1):107–135, 2024.

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

Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, and Christophe Pallier. Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax, and context. *Neurobiology of Language*, 4(4):611–636, 2023.

Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11: e77599, 2022. URL https://elifesciences.org/articles/77599.

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.

Jennifer M Rodd, Olivia A Longe, Billi Randall, and Lorraine K Tyler. The functional organisation of the fronto-temporal language system: evidence from syntactic and semantic ambiguity. *Neuropsychologia*, 48(5):1324–1335, 2010.

Andrea D Rowe, Peter R Bullock, Charles E Polkey, and Robin G Morris. Theory of mind'impairments and their relationship to executive functioning following frontal lobe excisions. *Brain*, 124(3):600–616, 2001.

Rebecca Saxe and Nancy Kanwisher. People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4):1835–1842, 2003. doi: 10.1016/S1053-8119(03)00230-1.

Rebecca Saxe and Anna Wexler. Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, 43(10):1391–1399, 2005.

Rebecca Saxe, Matthew Brett, and Nancy Kanwisher. Divide and conquer: a defense of functional localizers. *Neuroimage*, 30(4):1088–1096, 2006.

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307, 2020.

Cory Shain, Alexander Paunov, Xuanyi Chen, Benjamin Lipkin, and Evelina Fedorenko. No evidence of theory of mind reasoning in the human language network. *Cerebral Cortex*, 33(10): 6299–6319, 2023.

Chandan Singh, Richard J Antonello, Sihang Guo, Gavin Mischler, Jianfeng Gao, Nima Mesgarani, and Alexander G Huth. Evaluating scientific theories as predictive models in language neuroscience. *bioRxiv*, 2025.

Donald T Stuss, Gordon G Gallup Jr, and Michael P Alexander. The frontal lobes are necessary for theory of mind. *Brain*, 124(2):279–286, 2001.

Jerry Tang and Alexander G Huth. Semantic language decoding across participants and stimulus modalities. *Current Biology*, 35(5):1023–1032, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Neuronpedia Team. Neuronpedia: An open platform for mechanistic interpretability. https://www.neuronpedia.org, 2024.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024. URL https://www.nature.com/articles/s41562-023-01783-7.

Greta Tuckute, Elizabeth J Lee, Yongtian Ou, Evelina Fedorenko, and Kendrick Kay. A two-dimensional space of linguistic representations shared across individuals. *bioRxiv*, 2025. doi: https://doi.org/10.1101/2025.05.21.655330.

Leila Wehbe, Idan Asher Blank, Cory Shain, Richard Futrell, Roger Levy, Titus von der Malsburg, Nathaniel Smith, Edward Gibson, and Evelina Fedorenko. Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, 31(9):4006–4023, 2021.

W Caroline West and Phillip J Holcomb. Imaginal, semantic, and surface-level processing of concrete and abstract words: an electrophysiological investigation. *Journal of Cognitive Neuroscience*, 12(6):1024–1037, 2000.

Ethan G Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P Levy. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, 2020.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Alicia Zeng and Jack Gallant. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. *bioRxiv*, pp. 2025–11, 2025.

A Detailed Voxel Selection Procedure

We selected voxels of interest in two main ways: one based on the two principal components (PCs) of sentence-evoked responses identified by Tuckute et al. (2025), and the other one based on whether a voxel is part of the fronto-temporal language network (Fedorenko et al., 2010).

PC-derived subtypes (Hard-to-Process, Easy-to-Process, Abstract, Concrete). We define these four subtypes via a three-step procedure which was applied separately for each of the eight participants. We selected $n = 20$ voxels per subtype.

Step 1: Identify voxels significantly predicted by either of the two PCs. We restrict to left-hemisphere voxels that are significantly predicted by a two-PC linear encoding model (PC 1 and PC 2 as regressors; cross-validated $R^2 > 1.06\%$; see Tuckute et al. 2025 for details). This step ensures that the voxels entering subtype assignment are meaningfully captured by at least one of the two PCs.

Step 2: Assign voxels to subtypes based on PC correlations. For each remaining voxel, we compute its Pearson correlation with the sentence-level PC1 and PC2 scores (vectors of length 200, i.e., the number of stimuli in the experiment). The threshold $|r| = 0.14$ marks the correlation values expected by chance for a two-tailed $p < 0.05$ threshold based on 200 samples. We assign voxels to subtypes as follows:

Subtype	PC 1 criterion	PC 2 criterion	Ranking
Hard-to-Process	$r_{\mathrm{PC1}} > 0$	$\lvert r_{\mathrm{PC2}}\rvert < 0.14$	top n by r_{PC1} (descending)
Easy-to-Process	$r_{\mathrm{PC1}} < 0$	$\lvert r_{\mathrm{PC2}}\rvert < 0.14$	top n by r_{PC1} (ascending)
Abstract	$\lvert r_{\mathrm{PC1}}\rvert < 0.14$	$r_{\mathrm{PC2}} > 0$	top n by r_{PC2} (descending)
Concrete	$\lvert r_{\mathrm{PC1}}\rvert < 0.14$	$r_{\mathrm{PC2}} < 0$	top n by r_{PC2} (ascending)

Table 2: Criteria used to define the four voxel subtypes.

Step 3: Ensure voxel reliability. We require an NCSNR of at least 0.4. If fewer than $n = 20$ voxels meet this threshold for a given subtype, we incrementally relax it in steps of 0.02 down to a minimum of 0.20 (thresholds: 0.40, 0.38, 0.36, . . . , 0.20). This relaxation is occasionally necessary given that some participants have more reliable voxels than others (see the NCSNR values for each voxel subtype in Table 3). No anatomical constraint is imposed beyond restricting to the left hemisphere. Across all subtypes, the $n \times 4 = 80$ selected voxels are mutually exclusive (i.e., no voxel appears in more than one subtype).

Ghost voxels. Ghost voxels are selected separately and do *not* require significant prediction by the two-PC encoding model, since by definition they are not well captured by either PC. Starting from all left-hemisphere voxels with NCSNR > 0.4, we identify voxels whose correlations with both PC 1 and PC 2 fall within $(-0.14,\ 0.14)$ (i.e., chance-level). Among these voxels, we rank by the sum of absolute correlations ($|r_{\mathrm{PC1}}| + |r_{\mathrm{PC2}}|$) and select the $n = 20$ voxels with the lowest combined absolute correlation, i.e., those closest to the origin in PC-correlation space (see Figure 1A).

Language-network fROIs. Using a standard approach in cognitive neuroscience, we use an independent "functional localizer" experiment to select voxels that are responsive to language (Saxe et al., 2006; Fedorenko et al., 2010). For each participant, we identify voxels falling within five predefined left-hemisphere parcels—IFG, IFGorb, MFG, AntTemp, and PostTemp—and select the top 10% of voxels by t-statistic from a standard *sentences* $>$ *nonwords* localizer contrast (Fedorenko et al., 2010). As in all other analyses, we require NCSNR > 0.4 for the language fROI voxels.

B Replication Using `gemma-2-2b` Layer 14

In this section, we replicate our primary quantitative regression results in Section 4 using LM representations from layer 14 (rather than layer 12) of `gemma-2-2b`. From Fig. 6, one can see that our results are qualitatively similar to those in Fig. 2A.

Table 3: Summary statistics for the five voxel subtypes across eight participants.

Subtype	NCSNR	Lang t-stat	N voxels/participant
Hard to process	0.558 ± 0.012	1.874 ± 0.660	20.0 ± 0.0
Easy to process	0.366 ± 0.026	-0.467 ± 0.177	20.0 ± 0.0
Abstract	0.515 ± 0.016	5.292 ± 0.697	20.0 ± 0.0
Concrete	0.490 ± 0.036	2.522 ± 0.343	20.0 ± 0.0
Ghost	0.428 ± 0.005	1.045 ± 0.312	20.0 ± 0.0

Table 4: Summary statistics for language fROI voxels (NCSNR > 0.4) across eight participants.

fROI	NCSNR	Lang t-stat	N voxels/participant
IFGorb	0.474 ± 0.008	6.402 ± 0.823	74.7 ± 13.3
IFG	0.472 ± 0.009	5.935 ± 0.931	62.6 ± 17.7
MFG	0.463 ± 0.015	6.844 ± 0.740	90.8 ± 26.2
AntTemp	0.478 ± 0.009	7.518 ± 0.743	74.5 ± 18.1
PostTemp	0.486 ± 0.011	9.000 ± 0.738	343.8 ± 66.1

C Additional Methodological Details for All Analyses

Computational Costs Every analysis presented in this work was performed using CPUs. A single Nvidia TitanRTX GPU was used to featurize stimuli.

Cross-Validated Predictivity Analysis For each voxel, we use 5-fold cross validation over sentence stimuli to estimate the performance of our encoding models in Sections 4 and 6. For each train/test split, we employ a two stage processing pipeline. The first stage consists of feature selection. For Matryoshka and JumpReLU feature sets, we first filter to the 8000 most promising features using F-tests from univariate linear regression on each feature. This removes features that do not correlate with voxel responses. Next, we fit an L1-penalized (LASSO) regression. We use a nested 5-fold CV to identify the LASSO α hyperparameter, searching over a log-spaced range of 10 values from .01 to 1. After identifying the α hyperparameter, we fit the LASSO regression to the entire train split to identify a sparse feature basis.

The second stage consists of fitting a generalizable regression model to the sparse feature basis identified in the first stage. We employ a Ridge regression for this purpose. Regardless of the results of the LASSO feature selection, we always include the surprisal feature in the Ridge regression. This procedure ensures that the feature set is never empty, and allows for clean comparisons between full LM encoding model predictivity and a surprisal-only baseline. We use a nested leave-one-out cross validation to identify the optimal α hyperparameter, searching over a range from .01 to 100000. We refit the Ridge regression on the entire train split using the optimal alpha in order to predict test split responses.

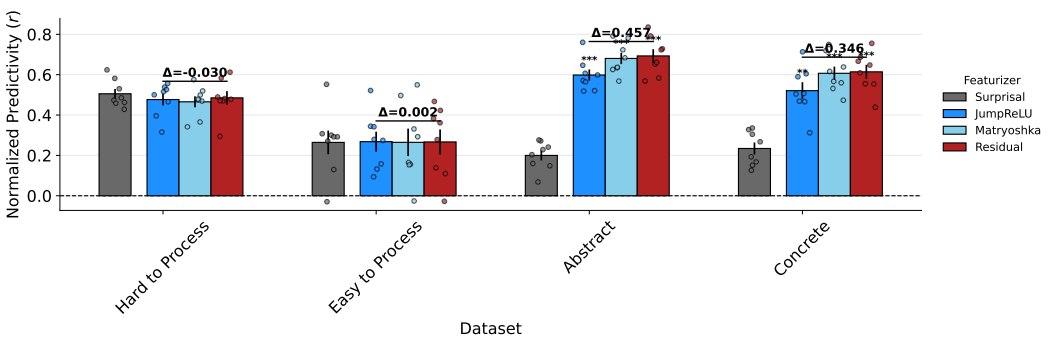

Figure 6: Replication of Fig. 2A using representations from layer 14 of `Gemma-2-2b`

Within each fold, brain responses were z-scored using training-set statistics only, and surprisal was z-scored using the training set only, preventing information leakage across folds. The final predictivity score was defined as the mean Fisher-z-transformed correlation across folds, noise-ceiling normalized at the voxel level.

For testing statistical significance among different featurizers (Figs. 2, 4, 6, 9), we use a paired t-test on the predictivity values obtained across all eight participants. For each figure, * = $p < .05$, ** = $p < .01$, *** = $p < .001$

Generalization Analysis Generalization analyses described in Sections 5 and 6 proceeded as follows. At a high level, the goal is to understand which voxels in a population are predicted by the same sets of features. To do this, we first identify a sparse set of features that predict one voxel, a *source voxel*, and then assess how well the identified (signed) feature set predicts each other voxel in the population.

In service of this goal, we identify features that predict a source voxel by first filtering features using an F-Test, then fitting a LASSO regression for feature selection, and finally fitting a Ridge regression to arrive at a final set of coefficients and features. Regardless of the results of the LASSO feature selection, we always include the surprisal feature in the Ridge regression. We run this pipeline using all sentence stimuli, resulting in a sparse, signed set of features that predict the source voxel.

Next, we wish to assess the degree to which this sparse, signed feature set can predict each of the other voxels in the population. We call each of these voxels *target* voxels. We use 5-fold cross validation over sentence stimuli to assess how well the (signed) features predict a target voxel. Within each of these folds, we use an 80-20 train/test split (rather than a computationally expensive nested 5-fold CV) to identify the optimal α value for a sign-constrained Ridge regression, searching over a range from .01 to 100000. After identifying the optimal α value for a fold, we refit a sign-constrained Ridge regression and predict test split voxel responses.

Feature Interpretation Analysis The feature interpretation analyses employed by Sections 4, 5, and 6 proceed by first filtering features using an F-Test, then fitting a LASSO regression for feature selection, and finally fitting a Ridge regression to arrive at a final set of coefficients and features. To arrive at a feature set that predicts voxel responses to all stimuli, we run this pipeline over all sentence stimuli (rather than performing a cross-validated analysis). After fitting the Ridge regression, we note the sign of each selected feature.

Other Methods Here we provide technical details on our comparison to existing methods in Fig. 1C. For Zeng & Gallant (2025), we featurized stimuli using their pretrained sparse concept featurizer dictionary, which contains 1000 sparse features. For Benara et al. (2024), we used the 29 questions for their final reported fMRI model (from their larger 674-question candidate pool). Following their ensemble procedure, each (sentence, question) pair was evaluated using `Mistral-7B-Instruct-v0.2` zero-shot and `Llama-3-8B-Instruct` with both zero-shot and five-shot prompts. Each configuration produced a binary yes/no response, and the three responses were averaged to produce one 29-dimensional QA feature vector per sentence. Finally for Eng1000, we created one average feature vector per sentence using the 985 features from Huth et al. (2016).

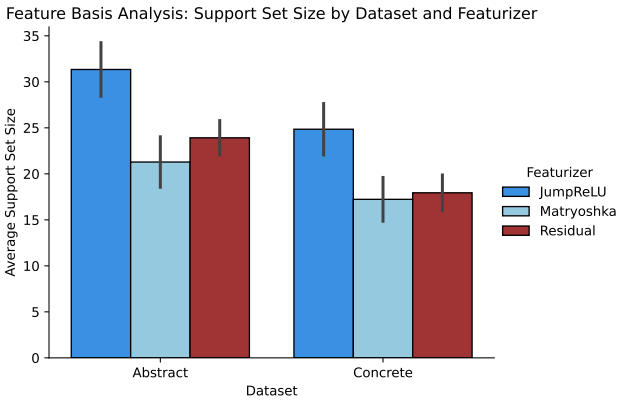

Figure 7: Average support set sizes (number of features selected by LASSO regression) for **Abstract** and **Concrete** voxel subtype regressions for all three feature sets used in Section 4.

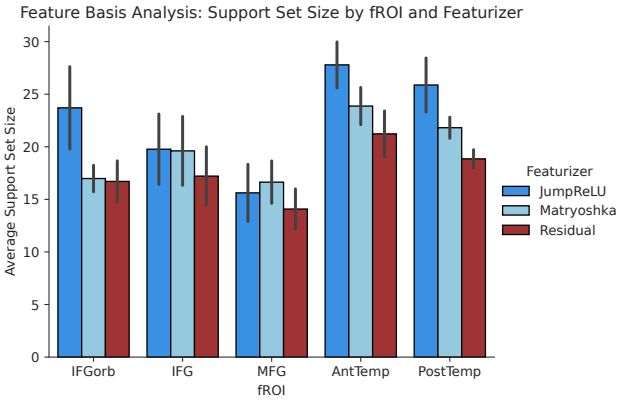

Figure 8: Average support set sizes (number of features selected by LASSO regression) for fROI regressions for all three feature sets used in Section 6.

D Feature Set Support Size

In Fig. 7, we present the average support size (i.e., number of features) resulting from LASSO-based feature selection using residual stream, JumpReLU, and Matryoshka feature sets for **Abstract** and **Concrete** voxel subtypes. When including processing difficulty voxel subtypes, the average Matryoshka support size is 18 features. In Fig. 8, we do the same, except for each language fROI. We find that Matryoshka feature sets result in comparable support set sizes to residual stream regressions, and typically smaller support sets than JumpReLU regressions.

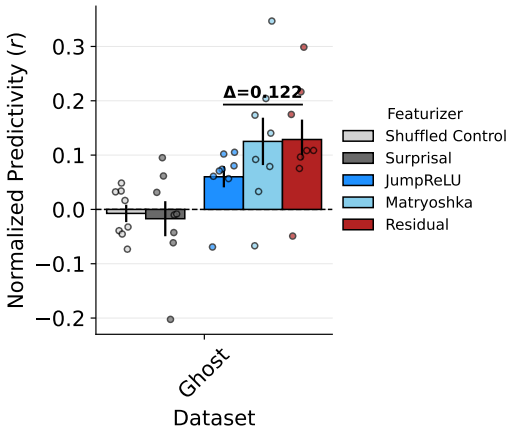

Figure 9: Regression predictivity for Ghost voxels using all feature sets.

E Ghost Voxel Regressions

In this section, we demonstrate that, on average, none of our feature sets can predict voxels in the **Ghost** subtype well. From Fig. 9, one can see that even the best performing feature sets (Residual and Matryoshka) only achieve Normalized Predictivity of approximately .1.

F Split-Half Qualitative Analysis

In this section, we perform a split-half analysis to assess the stability of the qualitative findings presented in Figure 2. Specifically, we split our stimuli into two disjoint sets, and reran the qualitative analysis presented in Figure 2 on each set independently. We identify the top 15 features in each split (identically to Figure 2), and find that their IoU is .58, with 19 features in the union. For each dataset in abstract, concrete, we then considered these 19 features and counted, separately within each split, the number of voxels for which each feature was selected. These feature selection frequencies were highly correlated across splits (r=.87 for abstract, r=.85 for concrete). Furthermore, the mean per-voxel IoU between splits is .24 for Abstract and .19 for Concrete. Despite there being more than 32K distinct features available to the model, both the aggregated feature-selection patterns and the individual voxel feature-selection patterns were reproducible across stimulus set splits.

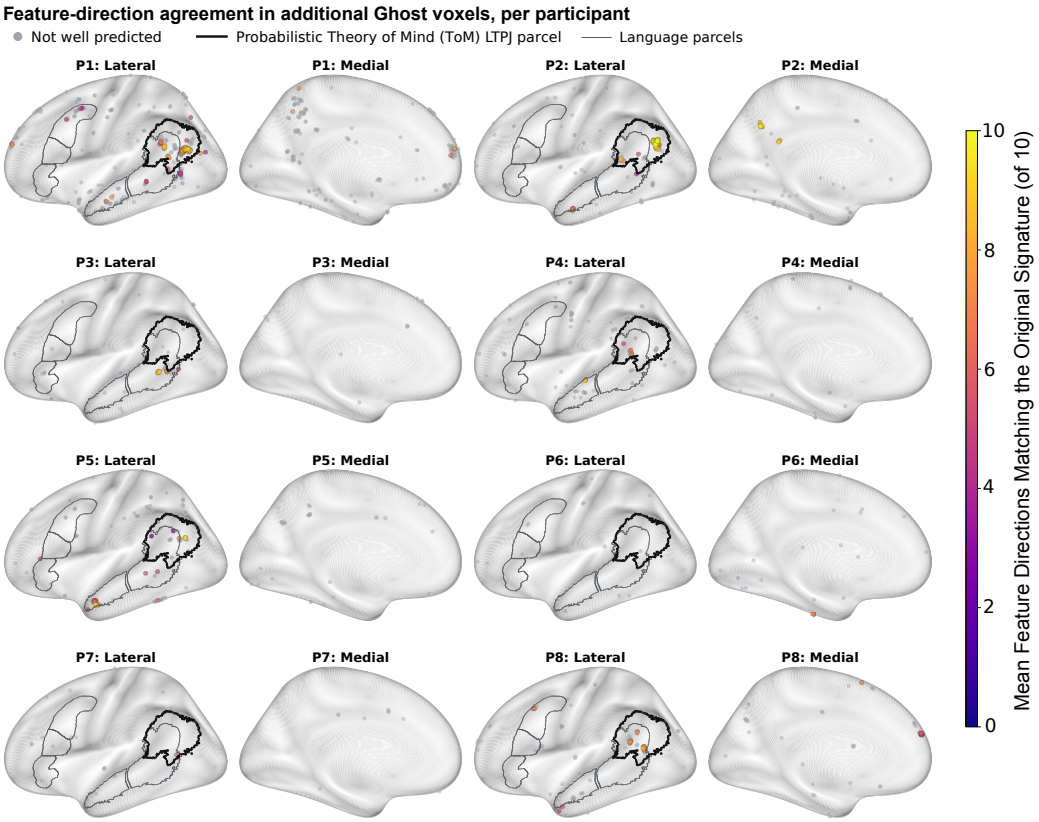

Figure 10: Participant-level maps from the additional **Ghost** voxel search (complementing the group map in **Fig. 3C**). Well-predicted voxels are colored by the number of feature directions (out of 10) matching the original people-centered signature; gray voxels were not well predicted. Black outlines indicate the probabilistic Theory-of-Mind LTPJ parcel (Dufour et al., 2013) and gray outlines indicate the language parcels.

G Participant-level Surface Maps for Additional Ghost Voxel Analyses

H Replication Using 3T Sentence-Level Brain Dataset

In this section, we replicate our main quantitative results from Section 6 using an independent sentence fMRI dataset (Tuckute et al., 2024). Five participants read 1,000 linguistically diverse sentences during 3T fMRI. Language fROIs were defined using the same localizer procedure as in the main analyses, and we refer the reader to Tuckute et al. (2024) for additional methodological details on this dataset. This dataset differs from the main 7T dataset in several respects: different participants, lower field strength (3T vs. 7T), different stimulus modality (reading vs. listening), different MRI preprocessing, and a different experimental design.

The experiment was designed for fROI-level rather than voxel-level analyses: it was acquired at 3T (vs. 7T) and contains no within-participant stimulus repetitions, so voxel-wise reliability cannot be estimated from repeated stimulus presentation (Allen et al., 2022).

Despite these differences, Fig. 11 shows the same qualitative pattern as Fig. 4: MFG exhibits the smallest difference between surprisal vs. other feature sets, and AntTemp exhibits the largest difference.

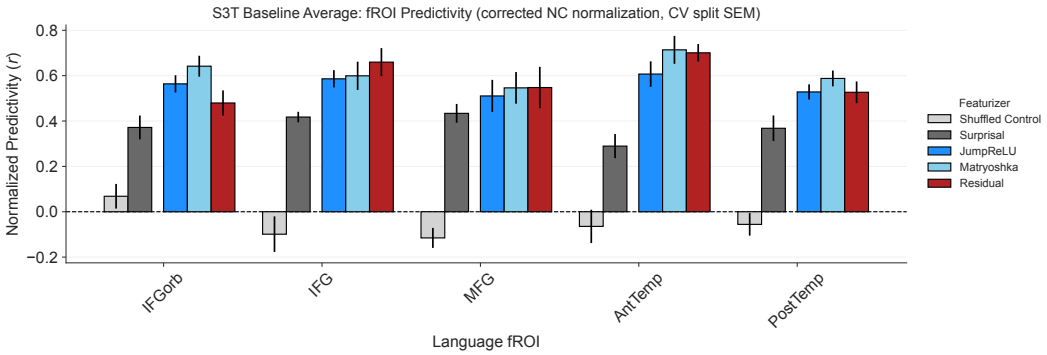

Figure 11: Replication of Fig. 4 on an independent dataset. Following the procedure in Tuckute et al. (2024), the brain responses in each language fROI were averaged across the five participants. Predictivity is reported as the mean Fisher-z-transformed correlation across 5 CV folds, normalized by the across-participant noise ceiling; see Tuckute et al. (2024) for details. Error bars show SEM across 5 cross-validation splits.

I Replication Using 3T Naturalistic Stories Brain Dataset

In this section, we replicate our main quantitative results from Section 6 using the Natural Stories fMRI dataset used by de Varda et al. (2025) (expanded from Blank et al. (2014)). The dataset comprises nine approximately five-minute spoken English stories (Futrell et al., 2018). We used the publicly released group-averaged response time series for five functionally defined left-hemisphere language fROIs: IFGorb, IFG, MFG, anterior temporal, and posterior temporal regions. Predictors were aligned to the released 2-s response bins, and we used the released response series with a three-TR hemodynamic shift for modeling (see details in de Varda et al. (2025)).

For feature extraction, we extracted `gemma-2-2b` layer-12 residual-stream representations and the corresponding Matryoshka and JumpReLU SAE features. Features were computed using a trailing 100-word causal context; subtoken representations were averaged within words and word representations within each 2-s bin. As in the main text analyses, we compared surprisal alone with models that augmented surprisal with selected residual-stream or SAE features.

Model performance was evaluated with nine-fold leave-one-story-out cross-validation. Within every outer fold, feature screening, feature selection, and hyperparameter selection used only the eight training stories. Candidate model features were first reduced to the 8,000 features with the strongest univariate associations and then selected using L1 regression with leave-one-training-story-out cross-validation. In line with our framework, surprisal was retained in each of the final models. The selected predictors were refit using ridge regression, with the ridge parameter chosen through an inner leave-one-story-out loop over the training stories. To minimize story-onset effects, the first 50 response bins (100 s) of each story were excluded during feature selection and inner and outer scoring to avoid "ramping-bias" (see Antonello et al. (2023)); the final ridge model itself was fit using all available bins from the training stories. Predictivity was quantified as the Pearson correlation between predicted and observed responses in the held-out story.

Despite major differences in experimental stimuli and methodological details, Fig. 12 replicates the key quantitative patterns from our main analyses: sparse SAE features substantially improve prediction over surprisal alone and the temporal fROIs show a larger boost from "content-based" feature predictors compared to frontal ones (e.g., MFG) (**Fig. 12A**). The predictivity of the SAEs (especially Matryoshka) approach the predictivity level of residual stream features (**Fig. 12A**). Moreover, coarse, earlier Matryoshka feature bins are preferentially selected in the regressions and more predictive (**Fig. 12B–C**). Finally, a small set of features is consistently selected across language fROIs (**Fig. 12D**), and the top-loading features are shared with those in the main 7T dataset (see **Fig. 13**) and **Fig. 4**).

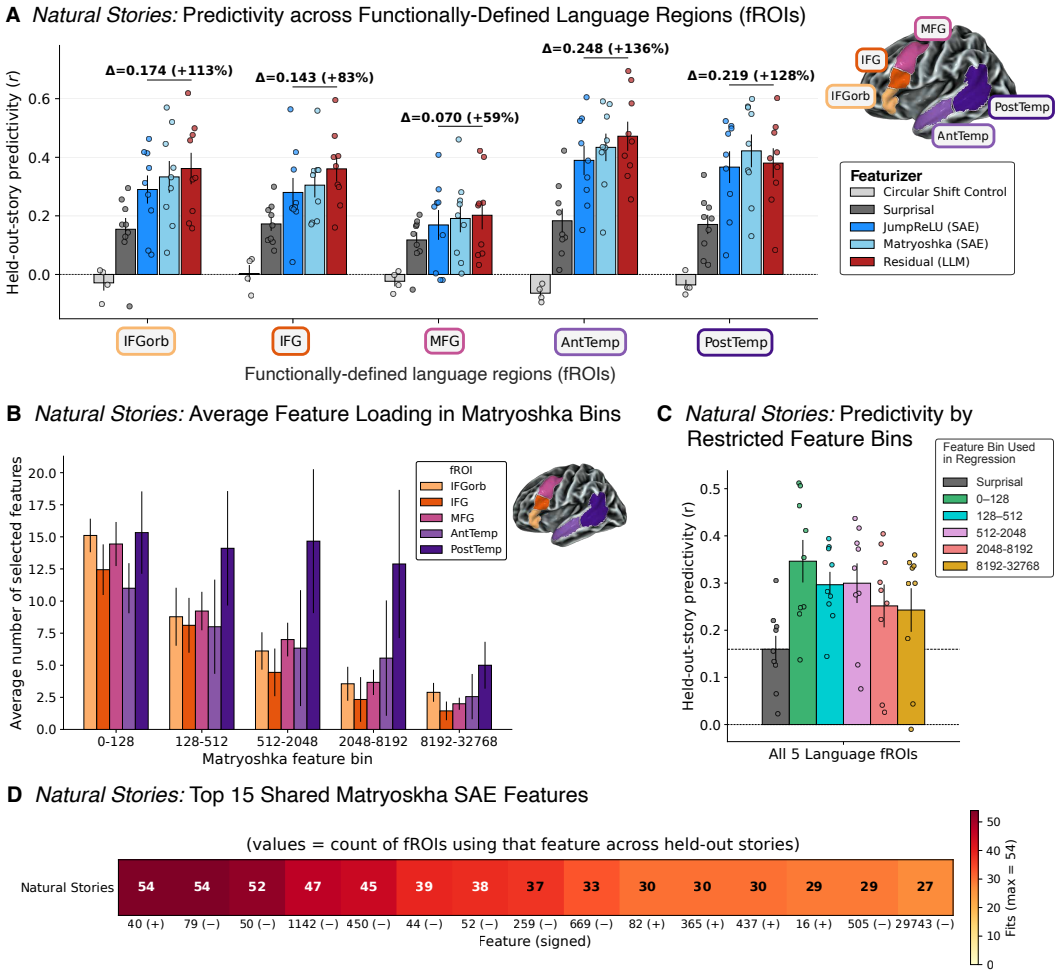

Figure 12: Replication of the language-network findings in the Natural Stories dataset, using naturalistic language comprehension. **(A)** Predictivity across language fROIs, replicating Fig. 4A. As a control, we used a circular shift baseline (de Varda et al., 2025). Bars show the mean held-out-story predictivity across the nine cross-validation folds, and error bars show the standard error of the mean across held-out stories. The Δ annotations report the raw and percentage increase from surprisal to the average predictivity of the three model-feature representations. **(B–C)** Matryoshka feature loading and restricted-bin predictivity, replicating Fig. 5B–C. Individual dots in Panel C denote individual held-out-story folds, bars denote their mean, and error bars show the standard error of the mean. **(D)** Most frequently shared Matryoshka features across language fROIs, complementing the feature-prevalence analysis in Fig. 4C. Cell color indicates the number of fROI-by-story fits in which each signed feature was selected. Six fROIs were used: IFGorb, IFG, MFG, anterior temporal, posterior temporal regions, and the aggregate of these five regions (i.e., the left-hemisphere language network).

J Detailed Information about Language fROI Voxel Selection for Generalization Analysis

Due to the computational cost of running the voxel-to-voxel generalization analyses for the language fROI voxels (which grows quadratically with the number of voxels; Section 6), we sub-sampled fROI voxels (see Table 4). The total number of language fROI voxels across participants was 5,021 voxels. The sub-sampling selection proceed as follows: First, we "deduplicated" voxels with identical brain response profiles within each participant (such duplicates arise from k-nearest-neighbor interpolation between native participant surfaces and the fsaverage template space (Fischl et al.,

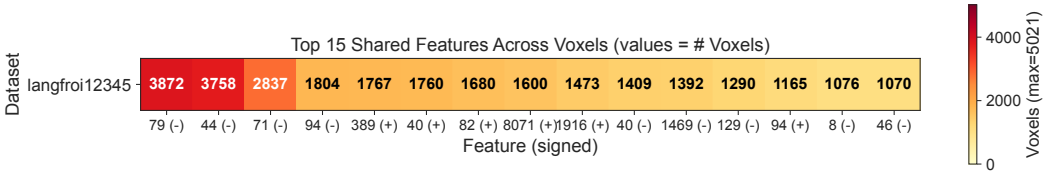

Figure 13: Signed Matryoshka feature prevalence across all participants and all language fROIs. We find many features from the most general Matryoshka bin, and many that arose when analyzing the voxel subtypes.

1999), where distinct fsaverage vertices can inherit the response of the same native-surface vertex), dropping 646 voxels (13%; 5,021 $\rightarrow$ 4,375). Second, we then kept the top 50% of voxels per (participant, fROI) combination, ranked by cross-validated NC-normalized predictivity (r) using the Matryoshka SAE features. Two participants had relatively few high-NCSNR voxels (P4: 117, P7: 113) and hence we did not further subsample these two participants' voxels. The final 2,296 voxels are distributed across participants and fROIs as reported in Table 5.

Table 5: Voxel counts for the fROI generalization analysis. Starting from the language fROI voxels (NCSNR $>$ 0.4), the top 50% by NC-normalized predictivity (r) were selected per participant $\times$ fROI, except for participants with fewer than 200 voxels (P4, P7), who were kept in full.

	Participant								
fROI	P1	P2	P3	P4	P5	P6	P7	P8	Total
IFGorb	52	39	39	0	16	24	0	21	191
IFG	54	42	53	9	1	48	16	14	237
MFG	64	98	25	7	59	11	16	59	339
AntTemp	34	47	11	21	38	79	28	28	286
PostTemp	215	237	127	80	172	210	53	149	1243
Total	419	463	255	117	286	372	113	271	2296

K Extended Feature Descriptions Table

This section contains an extended version of Table 1, which contains descriptions of Matryoshka features that often predict voxels.

L Language fROI Qualitative Feature Analysis

We visualize the prevalence of signed Matryoshka features across all participants and fROIs in Fig. 13. We first note that many of the most prevalent features are from the most general Matryoshka bin. Next, we note that many of these features appear in voxel subtype analysis in Section 4. These features broadly reflect the interpretations of the PCs that have been identified in Tuckute et al. (2025). Features -79, -94, +389, and +40 are all concordant with **Abstract** voxels, which are predominantly located in the language network. Features 44 and 71 are both driven more by sentences comprised of fewer tokens, and are thus suppressed by more complicated sentences. This reflects the prevalence of **Hard-to-Process** voxels in the language network.

Feat. #	Feat. Label	Annotation	Max-Activating Examples
94	People	Descriptions of public figures involved in events	- *She worked under execute chef Walter Scheib III, who resigned...* - *Singer and former Haitian presidential candidate, picture here...*
79	Scenery	Descriptions of locations and scenes	- *lost control and crashed into a ditch* - *apartment complex's newly created dog park*
40	Emotions	Informal speech related to emotions and relationships	- *oh shit cause I still have mad feelings for this girl right?* - *Yeah I'm bummed. To the max.*
49	Pronouns	Possessive pronouns.	*his*, *their*, *our*, *my*
44, 71	Short Sentences	Greater activation on sentences with few tokens.	*This feature activates on the first token. We take the mean of SAE activations, so this tracks the number of tokens in a sentence.*
389	Questions	Questions about ideas, processes, or existence.	- *Does anyone have a more precise idea?* - *Does one exist?*
1916	Consequences	Statements about consequences or results of processes.	- *from the initiation to the completetion of...* - *Consequently, it does not appear that the decline...*
16	Failure	Statements about destruction, failure, or error.	- *a slow-moving train-wreck that would do economic damage...* - *it will always crash halfway through...*

Table 6: Descriptions of more Matryoshka features that often predict voxel responses. Feature # corresponds to the dimension in the Matryoshka SAE latent space. Feature Label and Annotation are manually annotated summaries of the semantics of the feature. Max-Activating examples demonstrate the natural-language contexts that most activate these features, from Neuronpedia (Team, 2024).

M Additional Hierarchical Representational Alignment Analyses

In this section, we present additional analyses on the utility of different Matryoshka features for predicting voxel responses.

M.1 Extended fROI Analyses

We present an extreme case of the analysis presented in Fig. 5 — we compare the predictivity of a regression restricted to features from the most general bin of Matryoshka features (128 features) to the predictivity of a regression restricted to the union of all *other* feature bins ($>$ 30K features). In Fig. 14 find that the general feature bin enables better predictivity than the union of all other feature bins. This provides more support for the role of general features in predicting voxel tuning.

M.2 Replication using Subtypes Data

In this section, we replicate our feature granularity findings using the voxel subtypes data that were analyzed in Section 4. Figs. 15, 16, and 17 replicate the results presented in Fig. 5, and Fig. 18 replicates the results in Fig. 14.

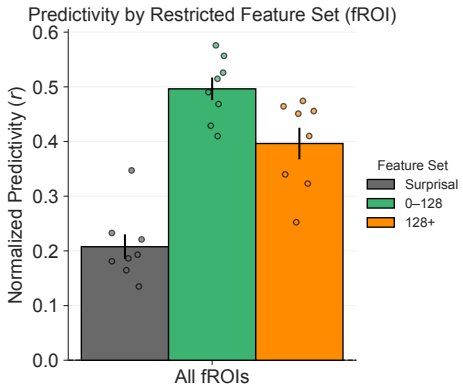

Figure 14: Predictivity of regressions restricted to using features from the most general Matryoshka bin vs. the union of all other features. We find that general features enable better predictivity.

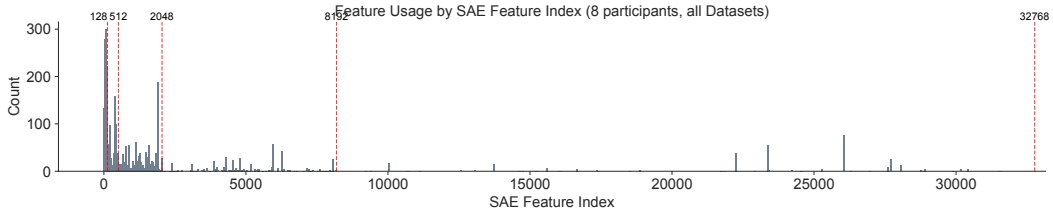

Figure 15: Histogram of how often each of the Matryoshka features is selected by encoding models in the voxel subtypes data.

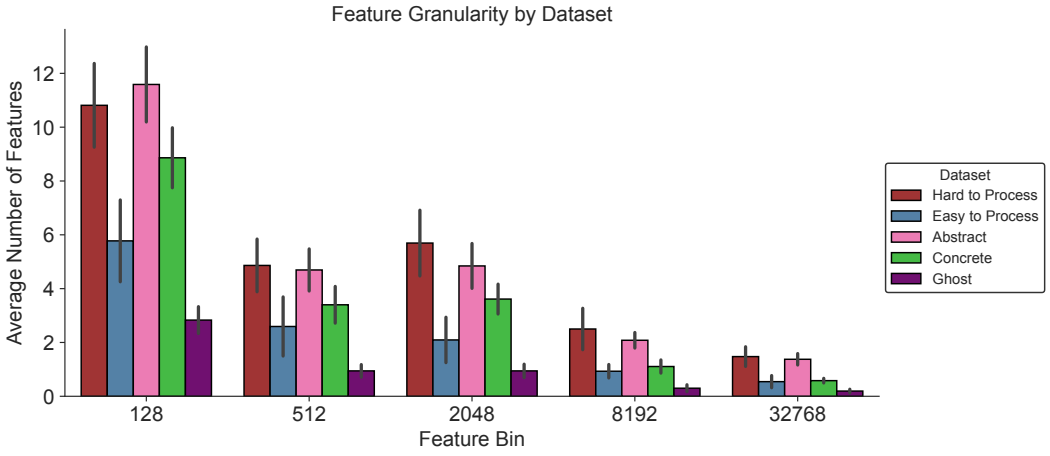

Figure 16: Average support set sizes (number of features selected by LASSO regression) per Matryoshka bin when predicting voxels in each subtype.

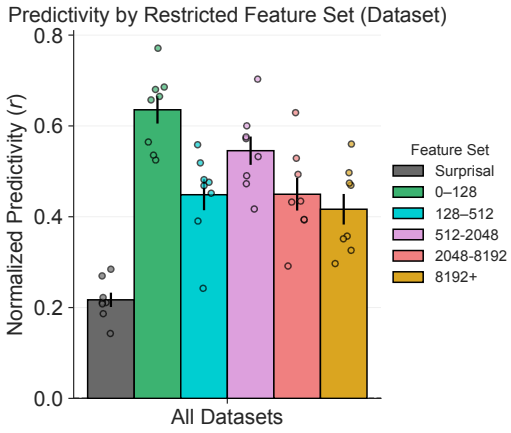

Figure 17: Performance of encoding models with feature sets restricted to individual Matryoshka feature bins.

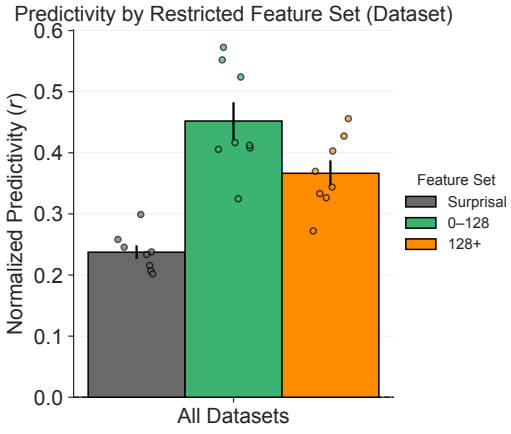

Figure 18: Predictivity of regressions restricted to using features from the the most general Matryoshka bin vs. the union of all other features. We find that general features enable better predictivity for voxels in the PC-derived subtypes.

Table 7: Per-bin statistics for the Matryoshka SAE on the 200 sentences used in the main analyses encoding models. "N non-zero" is the count of the number of features that activate on at least one sentence; the parenthetical is the percentage of the total number of features in the bin. "Min / Max fires" is the smallest / largest number of sentences (out of 200) any single non-zero feature in the bin fires on.

Bin	Range	N feats	N non-zero	Min fires	Max fires
1	0–128	128	122 (95.3%)	1	200
2	128–512	384	342 (89.1%)	1	200
3	512–2048	1,536	1,186 (77.2%)	1	199
4	2048–8192	6,144	3,397 (55.3%)	1	197
5	8192+	24,576	5,921 (24.1%)	1	200

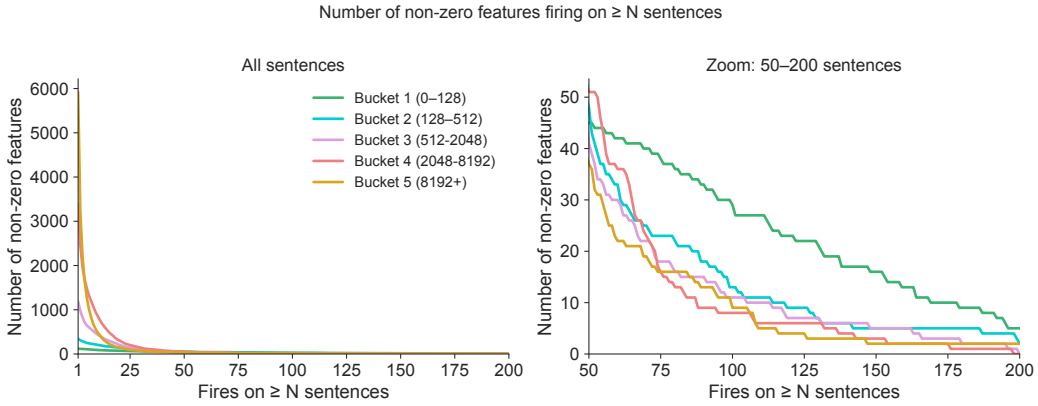

Figure 19: Number of non-zero Matryoshka features firing on at least N sentences, per bin. Left: full range, $N \in [1, 200]$. Right: zoom on $N \in [50, 200]$.

N Matryoshka SAE Feature Summary Statistics

In this section, we provide summary statistics for the Matryoshka SAE features (Gemma-2-2B, layer 12). Table 7 shows the count of non-zero features per Matryoshka bin on the 200-sentence stimulus set used in the main analyses, and Fig. 19 shows the number of non-zero features firing on at least N sentences as a function of N.

The table (7) shows that even in the fine-grained (late) bins there are still thousands of non-zero features (e.g., 3,397 in the 2048–8192 bin; 5,921 in 8192+), and at least one feature in every bin fires on nearly all 200 sentences ("Max fires" of 197–200). The right panel of Fig. 19 shows that around $N \approx 50$, all five bins converge to comparable non-zero feature counts (45, 48, 41, 51, 37 at $N{=}50$, respectively), even though the bins differ in raw size by orders of magnitude.

Bin-level predictivity differences are therefore not a trivial consequence of e.g., all-zero features in the late bins.

O Extended Discussion

O.1 The nature of linguistic representations: a shared feature basis

Prior work has shown that language-responsive regions show similar response profiles profiles during controlled and naturalistic language paradigms (Rodd et al., 2010; Fedorenko et al., 2020; Shain et al., 2020). Here, we extend this view by demonstrating that these profiles rely on a common, interpretable feature basis. Small sets of *signed* SAE features selected to predict one language-network voxel generalize strongly to voxels in other fROIs (Fig. 4B), unlike the **Ghost** population (Fig. 3A). This finding indicates that features can be reused across language fROIs. This full feature basis is

large—predicting voxels across all five language fROIs draw on many features (828 in total)—but far from random: a handful of SAE features dominate, including "processing difficulty-linked" features that track token count (and, implicitly, lexical frequency) and persist even with sentence-level surprisal as an explicit predictor (Appendix L). Though a shared basis set of features can predict voxels across fROIs, different fROIs are *not* uniformly tuned to these features. We find that processing difficulty features are useful for predicting voxels in all fROIs to different degrees (Fig. 4A, surprisal bar), and that there exists a basis of content-linked features that are useful for predicting voxels in most, but not all, fROIs.

Moreover, we find that a small set of core features also appear across *individuals* (Fig. 4C; in line with Tuckute et al., 2025). This core set of features is supplemented by many less prevalent features that are shared across most participants, as well as a broad set of features that are only shared within one or few individuals. While previous studies have shown that it is possible to generalize from one brain to another via LM residual-stream spaces (e.g., de Varda et al., 2025; Tuckute et al., 2024), these studies have been based on average fROI responses across the entire language network. The current analyses extend this prior work by generalizing at a single-voxel level instead of a network average (also see e.g., Tang & Huth, 2025), and further interpreting *which* features are shared. These results bring us closer to understanding how linguistic meaning is represented in individual brains. Future work with larger sets of stimuli can interpret these subset- and individual-specific features; one such direction is neural control (Bashivan et al., 2019; Tuckute et al., 2024; Antonello et al., 2024) in interpretable voxel subspaces, i.e., designing stimuli to drive targeted voxel populations.

O.2 People-centered content outside the dominant language dimensions

The Ghost analysis (Section 5) shows that Augmented Sparse Encoding Models can be used to identify and interpret voxel subpopulations that exhibit coherent feature tuning. We first identified a coherent cluster of **Ghost** voxels predicted by a people-centered feature profile, including features related to relationships, pronouns, and actions involving people.

We next asked whether the same people-centered feature profile extended beyond this initially identified cluster. In an anatomically unconstrained search across 1,490 additional Ghost voxels, we recovered the profile in five of eight participants (see Appendix G) and found that highly feature-aligned voxels were strongly concentrated in left temporoparietal junction (LTPJ). Within the five participants with more than three feature-aligned voxels, 80.3% of feature-aligned voxels fell within a probabilistic functional LTPJ Theory-of-Mind parcel (Dufour et al., 2013) (**Fig. 3C**), compared with 23.3% of well-predicted but non-aligned voxels. Importantly, no anatomical information was used to select these voxels.

These results provide a suggestive link to areas typically implicated in social cognition and Theory-of-Mind reasoning (Rowe et al., 2001; Stuss et al., 2001; Saxe & Kanwisher, 2003; Dufour et al., 2013; Miao et al., 2026). However, we emphasize that we did not include an independent Theory-of-Mind localizer in these participants, and therefore do not claim that these voxels are themselves functionally defined Theory-of-Mind regions. Similarly, failure to recover the profile clearly in the three out of eight participants could reflect either genuine inter-individual variability or limited sensitivity of the present dataset. We note that our analyses are restricted to the left hemisphere, where TPJ organization has been reported to be more variable across individuals than in the right hemisphere (Saxe & Wexler, 2005; Dufour et al., 2013).

This people-centered tuning profile also appears largely outside the core fronto-temporal language network. The strongest concentration of feature-aligned voxels lies adjacent to the probabilistic posterior temporal language parcel, rather than within it. This finding is consistent with prior work showing that the core language network is not tuned to social content or engaged by (non-verbal) theory-of-mind reasoning (Shain et al., 2023; Tuckute et al., 2024, cf. Mellem et al., 2016). At the same time, a smaller proportion of voxels with the people-centric tuning profile fell within the two temporal language parcels (Fig. 3C), suggesting that social tuning during comprehension of short sentences appears partly adjacent to the core language network and partly interspersed with temporal language areas. Follow-up work using independent language and Theory-of-Mind localizers will be important for more clearly distinguishing these functional populations.

Finally, because this tuning is spatially distributed and not aligned with the dominant PCs, it would be tricky to detect from anatomy or low-dimensional response structure alone. Augmented Sparse

Encoding Models isolate this people-centered voxel subpopulation through a small, signed feature basis, separating it from voxel populations tuned to processing difficulty or concrete–abstract content.

O.3 General LM features predict brain responses

The Matryoshka hierarchy analysis (Section 6) gives a more specific interpretation of why LM representations align with brain responses. The strongest alignment is *not* carried by the multiplicity of fine-grained features that contribute to an LM's high-dimensional feature space. Instead, brain-predictive features are concentrated among the set of features that most generically characterize LM representations. This finding suggests partial overlap in the features that organize internal language representations in the human language network and LMs, even though these systems presumably having very different means of arriving at these representations.

A priori, it was plausible that fine-grained features could have contributed substantially to predicting voxel responses. Various metrics that capture representation reconstruction steadily improve by incorporating such fine-grained features (Bussmann et al., 2025), and incorporating disentangled hierarchical features have resulted in improvements in sparse probing (Luo et al., 2026), and removing spurious correlations (Bussmann et al., 2025). However, our work suggests that most single voxels do not tend to represent fine-grained features, and are instead tuned to broader linguistic features. We do note that some fine-grained features predict responses in individual participants or small subsets of participants (Fig. 4C, inset), likely reflecting individual-specific representations of linguistic meaning.

Our claim about general SAE dimensions aligns with Antonello et al., 2021, who showed that low-dimensional structure in LM representations is reflected in brain responses. We extend this work by identifying the relevant dimensions in an interpretable sparse feature basis. We further leverage these features to identify voxel subpopulations and characterize how these feature bases are shared across regions and individuals.

Finally, our SAE feature granularity analyses connect to work in the visual domain. For instance, Longon et al. (2025) show that SAE-based disentanglement can increase model–brain alignment when superposition obscures shared latent structure (also see Mao et al. (2025)). Here, we use a hierarchical SAE basis to ask *which* levels of LM feature structure mediate brain alignment during language processing.