Brain-to-Text Decoding with Brain Atlases and Brain Foundation Models

Haruka Akama[1], Ryo Yoshida[1], Max Müller-Eberstein[1,2], Yohei Oseki[1]
[1]The University of Tokyo, Japan
[2]IT University of Copenhagen, Denmark
{haruka-akama, yoshiryo0617, iijimax, oseki}@g.ecc.u-tokyo.ac.jp

Abstract

Brain-to-text decoding using pre-trained Large Language Models (LLMs) as decoders has recently begun to enable open-ended, sentence-level generation directly from neural recordings. However, despite these decoder-side advances, a critical bottleneck still lies in the encoder which compresses extremely high-dimensional fMRI data into compact brain representations. This study investigates the contrast between three primary encoding strategies: an *expert-driven mapping*, **Brain Atlas**, of fMRI signals to 424 brain regions; *data-driven mappings*, namely **PCA**-reductions, as typically used in prior work; and a *hybrid* **Brain Foundation Model** (**BFM**), combining an expert-driven atlas encoding with large-scale data-driven pretraining. Experiments on the Narratives fMRI dataset demonstrate that including an expert-driven mapping significantly outperforms the purely data-driven PCA-configurations across all evaluation metrics. Additionally, we find that adding the computationally expensive BFM on top of the heuristic Brain Atlas encoding yields no statistically significant gains. Our ablation analyses reveal that this divide is driven by Brain Atlas features smoothing cross-subject and temporal variance, while retaining a sparse importance profile which prioritizes signals from brain subnetworks related to language processing. These findings highlight expert-driven, spatially-aware feature aggregation as a key direction for future decoding performance gains.

Keywords: brain-to-text decoding, fMRI, brain atlas, brain foundation model, large language model

1. Introduction

Brain-to-text decoding aims to reconstruct natural language directly from neural activity. As such, it holds significant promise for Brain–Machine Interface (BMI) applications, particularly for individuals with severe motor impairments such as amyotrophic lateral sclerosis (ALS) or locked-in syndrome, who are unable to express their intended language through speech or movement. While early work focused on word-level recognition (Pereira et al., 2018), the recent adoption of Large Language Models (LLMs) as decoders has begun to enable open-ended, sentence-level generation from fMRI recordings (Ye et al., 2025).

Despite this progress on the decoder side, a critical bottleneck on the encoder side has gone largely unexplored. Prior work has primarily relied on Principal Component Analysis (PCA), or custom down-projections, to compress the extremely high-dimensional brain signal (often hundreds of thousands of voxels) into a compact representation that the decoder can process. As a result, it remains an open question as to how *encoder design* affects decoder performance, as well as how the *utilization* of the raw brain signals differs across each encoding method.

To shed light on these questions, this study contrasts three general strategies for encoding raw brain data from functional Magnetic Resonance Imaging (fMRI; Figure 1): an *expert-driven mapping*, **Brain Atlas**, based on brain atlases that define functional boundaries in the brain based on neuroscience literature; *data-driven mappings*, namely **PCA**-reductions of the fMRI data following prior work (Ye et al., 2025); and a *hybrid mapping*, namely the **Brain Foundation Model** (**BFM**; Caro et al., 2024), which further encodes the expert-driven Brain Atlas representation through a highly

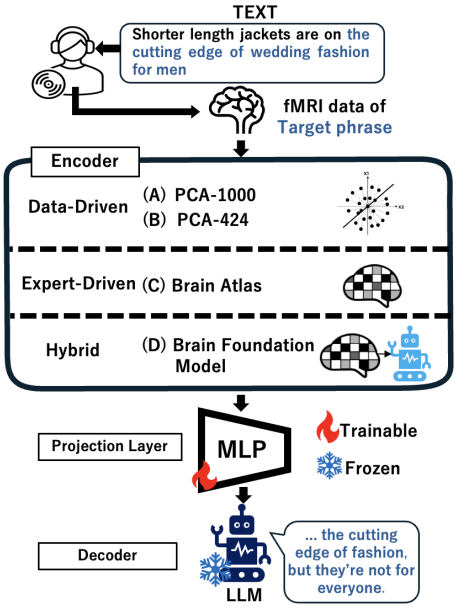


Figure 1: Overview of the brain-to-text decoding pipeline, within which we study four encoder designs: Purely data-driven (A) PCA-1000, (B) PCA-424; purely expert-driven (C) Brain Atlas; and a hybrid (D) Brain Foundation Model (Section 3).

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 124–144
16 May 2026.

data-driven contextualized embedding model, pretrained on 6,700 hours of fMRI recordings.

Towards a better understanding of how brain signal encoder design affects language decoder performance, this work contributes:

- Encoder design ablations covering expert-driven Brain Atlas mappings, data-driven PCA reductions, hybrid BFM embeddings, and two control settings (Section 3);
- Brain-to-text decoding experiments, which show that expert-driven Brain Atlas mappings outperform PCA and control baselines, as well as BFM on efficiency (Section 5);
- A regional ablation analysis, which shows that Brain Atlas dimensionality reduction exhibits a sparse contribution structure focused on language processing areas, whereas BFM and PCA features are broadly distributed and include negative contributions that impede performance (Section 6);

Our results establish expert-driven compression as a more effective design choice for brain-to-text decoding, and highlight the issues underlying current, purely data-driven encoder designs.

2. Related Work

Using non-invasive neural recordings, such as fMRI, to decode different types of information has been a formidable challenge across modalities, owing to the inherent trade-off between temporal and spatial resolution in neuroimaging and the significant modality gap between neural activity and externalized concepts. Due to the high information density of brain data, the direct reconstruction of concrete concepts has only recently been enabled with the advent of more advanced statistical machine learning methods. This includes the classification of mental states (Haynes and Rees, 2006), as well as early formulations of how to potentially decode image information from the visual cortex (Thirion et al., 2006).

For the reconstruction of linguistic information, initial studies similarly focused on the closed-set problem of word-level classification (Mitchell et al., 2008), with later studies introducing phoneme-level reconstruction, in order to enable the decoding of unseen words (Pei et al., 2011). For an even higher level of flexibility, Pereira et al. (2018) further introduced universal decoders generating semantic vectors via ridge regression. At the sentence level, Tang et al. (2023) added a selective decoding framework to a pre-trained language model to enable higher-quality decoding. In the end, all of these methods nonetheless rely on candidate sets to delimit the decoding space, and do not perform fully free-form text generation.

With the advent of more capable decoder models, the open-endedness, granularity, and quality of reconstructed outputs has increased dramatically. The introduction of diffusion models (Ho et al., 2020) in computer vision has lead to full brain-to-image decoding pipelines with high fidelity (Takagi and Nishimoto, 2023; Chen et al., 2023; Wang et al., 2024). Meanwhile, the introduction of LLMs has begun to enable open-ended brain-to-text decoding at the sentence level: Ye et al. (2025) proposed one of the first open-ended generation pipelines, which map PCA-reduced fMRI data to the embedding space of an LLM, in order to decode brain activity from reading and listening to stories back to full sentences. To verify that the model exploits brain-specific information, they additionally introduced a permuted control model with randomly mismatched signals. The unpermuted model significantly outperformed this control with a win rate of 66.5%, confirming that the decoder leverages genuine neural information. In our experiments, we adopt this pipeline as our direct baseline.

While advancements on the decoder side have driven continual improvements to generation quality, the encoder side has remained fairly constant. Due to the extremely high dimensionality of fMRI data (e.g., 250k voxels for a resolution of $3\times3\times4$ mm), typically, some form of dimensionality reduction (e.g., PCA) is applied, before the resulting compressed representation is passed through a learned projection (e.g., small MLP), which maps the brain activation vector to the decoder's embedding space. This configuration has largely remained unchanged across all aforementioned configurations, yet presents a critical and understudied bottleneck to the quality of the decoded outputs. As such, this work explicitly studies the effects of different encoder design choices on the task of open-ended, sentence-level brain-to-text decoding.

3. Methodology

We first outline the brain-to-text decoding pipeline of Ye et al. (2025), and define four encoder designs that are either data-driven, expert-driven, or both, to evaluate in the subsequent experiments.

Task Formulation. A raw fMRI scan is represented as a four-dimensional array $\mathcal{X} \in \mathbb{R}^{X\times Y\times Z\times T}$, where (X, Y, Z) denote the spatial dimensions and T the number of time steps. The spatial dimensions are first flattened to yield a matrix $M \in \mathbb{R}^{T\times V}$ ($V = X \cdot Y \cdot Z$); taking a single row gives $\mathbf{x} \in \mathbb{R}^V$, the fMRI signal at one time step. Given $\mathbf{x}$ and a text context $y_{<t} = (y_1, \ldots, y_{t-1})$, the goal of brain-to-text decoding is to generate the

subsequent tokens $y_t, y_{t+1}, \ldots$ that approximate the speech perceived by the subject.

As illustrated in Figure 1, the decoding pipeline consists of three stages:

1. **Encoder** $f_{\text{enc}} : \mathbb{R}^V \to \mathbb{R}^d$ reduces the high-dimensional fMRI signal to a compact d-dimensional representation $\mathbf{z} = f_{\text{enc}}(\mathbf{x})$. In this work we compare four encoder designs that differ in their dimensionality reduction strategy, i.e., whether they are purely data-driven, expert-driven, or hybrid.

2. **Projection** $g : \mathbb{R}^d \to \mathbb{R}^{d_{\text{LLM}}}$ maps the encoder output into the LLM's embedding space, yielding the brain embedding $\mathbf{z}_{\text{brain}} = g(\mathbf{z})$.

3. **LLM Decoder** generates text autoregressively, conditioned on $\mathbf{z}_{\text{brain}}$ and the text context. Added special tokens $\langle$`brain/`$\rangle$ and $\langle$`/brain`$\rangle$ mark the boundaries of the brain-derived embedding in the LLM input sequence.

Since neural activity during language comprehension is a continuous, distributed process across multiple cortical regions (Heald et al., 2023), as well as due to high amounts of noise during the recording process, information loss during the encoding phase can become a critical bottleneck for decoding performance. Therefore, the central design question of this work is the choice of encoder f_{enc}. Specifically, how to most effectively compress the high-dimensional fMRI signal while preserving the linguistically relevant neural information needed for language generation.

Encoder Designs. We organize our experimental configurations along two primary axes: **data-driven** dimensionality reduction, and **expert-driven** anatomical mapping. Within this space, we evaluate the following four encoder configurations:

- **(A) PCA-1000 (Data-driven):** fMRI signals are reduced to 1,000 dimensions using PCA, then mapped by the projection layer. This configuration is used in Ye et al. (2025) (Section 3.1).

- **(B) PCA-424 (Data-driven):** fMRI signals are reduced to 424 dimensions using PCA, then mapped by the projection layer. This configuration serves as a matched-dimension PCA baseline to (C) Brain Atlas (Section 3.1).

- **(C) Brain Atlas (Expert-driven):** fMRI signals are converted into 424 mean activations grouped by parcellating the brain into the corresponding regions of interest (ROIs) using the AAL-424 atlas (Nemati et al., 2020), then mapped by the projection layer (Section 3.2).

- **(D) Brain Foundation Model; BFM (Hybrid):** The same 424 ROI-wise signals as in (C) are fed to a contextual embedding model, pre-trained on fMRI data (Caro et al., 2024), followed by the projection layer (Section 3.3).

In terms of purely data-driven approaches, PCA-based dimensionality reduction is often used in prior work, including Ye et al. (2025). In contrast, brain atlases provide an expert-driven approach that parcellates the brain into anatomically informed ROIs, thereby preserving functional and anatomical organization. Additionally, the recent introduction of pre-trained brain foundation models such as Caro et al. (2024)'s BFM offers a hybrid encoding approach, combining an initial expert-driven atlas mapping with large-scale data-driven pretraining.[1] Despite these fundamental differences in design principles, the impact of the encoding strategy on decoding performance has not yet been investigated in prior work. We argue that one key to improving brain-to-text decoding lies in the design of this encoding stage.

Control Settings. To additionally evaluate the effectiveness of using any brain signal, we further employ two control designs based on Ye et al. (2025):

- **(E) LLM-only:** The LLM receives text embeddings only, with no fMRI input, i.e., special tokens with no intermediate brain embeddings. This condition isolates the contribution of brain signals by providing a text-only upper bound.

- **(F) Random:** fMRI signals are replaced with pre-generated random 424-dimensional vectors, which are passed through the projection layer. Thus, the model architecture is identical to (C), but the brain input contains only noise. This condition tests whether the model exploits any meaningful brain signal.

3.1. PCA Baselines

Principal Component Analysis (PCA) is a parameter-free dimensionality reduction method, that is driven purely by the variance characteristics of its input data. As it is used in Ye et al. (2025), it further serves as our direct baseline.

Before PCA, fMRI data are normalized: for each voxel its temporal mean is subtracted, and the result is divided by a per-voxel global standard deviation estimated by pooling all subjects' recording per dataset. Given the normalized time series $\tilde{X} \in \mathbb{R}^{T_{\text{all}} \times V}$ (all subjects' timesteps for one dataset

[1] While it would be of interest to investigate BFM without the initial brain atlas mapping, this is the only input format accepted by the pre-trained model.

concatenated), PCA finds $d{+}1$ orthogonal directions $W = [\mathbf{w}_1, \ldots, \mathbf{w}_{d+1}] \in \mathbb{R}^{V\times(d+1)}$ that maximize explained variance:

$$\mathbf{w}_k = \underset{\|\mathbf{w}\|=1,\ \mathbf{w}\perp\mathbf{w}_{1:k-1}}{\arg\max} \ \mathrm{Var}\left(\mathbf{w}^\top \tilde{X}^\top\right).$$

The first principal component (PC_1), which predominantly captures globally correlated physiological noise (e.g., heartbeat, respiration; see Carbonell et al., 2011; Birn, 2012; Murphy and Fox, 2017), is discarded. Each normalized volume is then projected onto a d-dimensional representation using PC_2 through PC_{d+1}:

$$\mathbf{z} = W_{2:d+1}^\top(\tilde{\mathbf{x}} - \boldsymbol{\mu}) \in \mathbb{R}^d,$$

where $\boldsymbol{\mu} \in \mathbb{R}^V$ is the column mean of $\tilde{X}$ and $W_{2:d+1}$ collects columns 2 through $d{+}1$ of W. We evaluate two output dimensionalities: (A) $d = 1{,}000$, replicating Ye et al. (2025); and (B) $d = 424$, matching the Brain Atlas feature dimensionality.

A defining property of PCA is that its components are determined purely by variance maximization across the full voxel space, without reference to any anatomical parcellation. This means that it can be applied in a purely data-driven manner, without any expert annotation. However, it may also capture high-variance components that reflect non-neural sources—such as head-motion artifacts and physiological noise—rather than functionally organized neural signals.

3.2. Brain Atlas

Generally, a brain atlas provides a static mapping of normalized voxel positions to anatomically informed brain regions. For our experiments, we adopt the Automated Anatomical Labeling 424 atlas (AAL-424; Nemati et al., 2020), which parcellates the brain into $K = 424$ non-overlapping ROIs within the normalized Montreal Neurological Institute (MNI) coordinate system (Figure 2). Since AAL-424 is also used in the initial encoding step of the BFM (Caro et al., 2024), this configuration allows us to compare the performance of an expert-driven mapping with and without an additional data-driven encoder model. Unlike PCA, whose partition of the voxel space is determined entirely by data-driven variance maximization, the atlas partition is fixed *a priori* based on anatomical boundaries (major sulci) grounded in neuroscience knowledge.

Formally, a brain atlas defines a partition $\mathcal{A} = \{R_1, R_2, \ldots, R_K\}$ of the voxel index set, where $R_k \subseteq \{1, \ldots, V\}$ is the set of indices assigned to the k-th ROI. The R_k are disjoint, and $\bigcup_k R_k \subseteq \{1, \ldots, V\}$. For each ROI, the encoder computes the mean activation over its constituent voxels:

$$z_k = \frac{1}{|R_k|} \sum_{v\in R_k} x_v, \quad k = 1, \ldots, K.$$

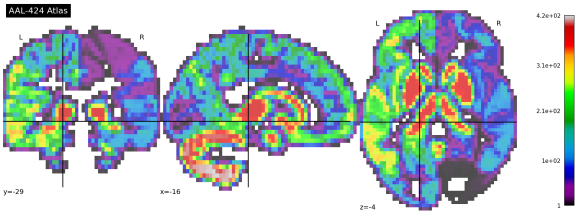


Figure 2: The AAL-424 brain atlas in MNI space (coronal, sagittal, and axial slices). Each color denotes a distinct ROI indexed from 1–424. The fine-grained parcellation covers both cortical and subcortical structures.

The full encoder output is the concatenation of all ROI signals:

$$\mathbf{z} = [z_1, z_2, \ldots, z_K]^\top \in \mathbb{R}^K, \quad K = 424.$$

As with PCA, the resulting ROI signals are normalized per dataset and per subject: for each ROI, its temporal mean (within the recording) is subtracted and the result is divided by a per-ROI global standard deviation estimated from that subject's recordings within the dataset. Crucially, each dimension z_k corresponds to a specific, anatomically labeled brain region, making the resulting representation directly interpretable. Additionally, the normalization to MNI space corrects for cross-subject morphological variability, ensuring that ROI R_k refers to the same anatomical structure across individuals and enabling meaningful cross-subject aggregation.

Beyond the ROI-based AAL-424 parcellation, the Akiki-Abdallah hierarchy groups the same 424 ROIs into 24 modules (AA-24; Akiki and Abdallah, 2019) nested within 7 canonical parent networks (Visual, Ventral Salience, Central Executive, Sensorimotor, Default Mode, Dorsal Salience, and Subcortical), which we use for our region-level ablation analysis in Section 6.

3.3. Brain Foundation Model

In order to encode brain recordings in a denser manner, recent work has proposed mirroring masked language models (Devlin et al., 2019) and vision transformers (Dosovitskiy et al., 2021) to train brain foundation models (BFMs). In this work, we leverage a 13M-parameter BFM by Caro et al. (2024), pre-trained on approximately 6,700 hours of non-linguistic fMRI data through self-supervised learning. This particular BFM offers a hybrid expert and data-driven configuration for our experiments, as it first maps the raw fMRI data to the AAL-424 atlas, before embedding it further. In the pretraining phase, activation patches from random brain regions are masked, and the model learns to reconstruct them from the remaining context, thereby capturing inter-regional functional relationships and

spatiotemporal co-activation patterns without labeled data.

Following the intended design of BFM, we discard the reconstruction head used in pretraining and use only the frozen Transformer backbone as a feature extractor over AAL-424 parcel-level fMRI signals. The Brain foundation model processes each recording by dividing it into non-overlapping chunks of 200 timepoints. Within each chunk, the time series is further segmented into non-overlapping temporal patches of 20 timepoints. Let $\mathbf{H}_{v,p} \in \mathbb{R}^{d_b}$ denote the final-layer hidden vector for voxel (parcel) index $v = 1, \ldots, V$ at patch index p, after removing the CLS token. We then apply mean pooling over the spatial (voxel/parcel) dimension to obtain a fixed-length patch representation $\mathbf{z}_p \in \mathbb{R}^{d_b}$:

$$\mathbf{z}_p = \frac{1}{V} \sum_{v=1}^{V} \mathbf{H}_{v,p}.$$

For each sample, the patch representation corresponding to that sample is used as a single brain token and finally passed to the projection layer. Through this configuration, we examine the effects of leveraging external data-driven knowledge for the informativeness of brain activation encoding.

4. Experimental configuration

Next, we describe the datasets, models, training procedure, and evaluation metrics used to compare the effect of the four encoder designs (A–D) defined in Section 3 on brain-to-text decoding performance.

Models. Following the sentence-level brain-to-text decoding pipeline of Ye et al. (2025), the projection consists of two hidden layers with ReLU activations. Similarly, the final LLM decoder uses `GPT-2 small` (Radford et al., 2019) (124M parameters) across all experimental settings. The brain foundation model (Caro et al., 2024) comprises approximately 13M frozen parameters and was used exclusively in configuration (D).

Training Data. For a direct comparison with Ye et al. (2025), we utilize data from 27 subjects in the *Narratives* dataset (Nastase et al., 2021), consisting of fMRI recordings of individuals listening to approximately 4.6 hours of auditory stories. Specifically, we used five stories: *pieman*, *lucy*, *notthefallintact*, *slumlordreach*, and *tunnel*. Recordings were acquired on a Siemens Skyra 3T scanner with a spatial resolution of 3×3×4 mm and a repetition time of 1.5 s. For each subject, data were split into training (60%), validation (20%), and test (20%) sets based on temporal order. The dataset underwent standard preprocessing via `fMRIPrep 20.0.5` (Esteban et al., 2019), including motion correction and normalization to `MNI152NLin2009cAsym` space.

Training Procedure. The parameters of the LLM decoder are kept frozen, with only the projection layer and the embeddings of the special tokens ⟨`brain/`⟩ and ⟨`/brain`⟩ around $\mathbf{z}_{\text{brain}}$ being subject to optimization. In configuration (D), the parameters of the brain foundation model are also kept frozen.

The input embedding sequence $\mathbf{X}$ to the LLM is defined as:

$$\mathbf{X} = \left[\langle \texttt{brain/} \rangle,\ \mathbf{z}_{\text{brain}},\ \langle \texttt{/brain} \rangle,\ \mathbf{e}_1, \ldots, \mathbf{e}_{t-1} \right]$$

where $\mathbf{z}_{\text{brain}}$ denotes the encoder output used as the brain embedding (i.e., PCA features in configurations (A) and (B), brain atlas-based features in (C), and BFM-derived features in configuration (D)). Here, $\mathbf{e}_1, \ldots, \mathbf{e}_{t-1}$ are the token embeddings of the preceding context provided to the decoder. We minimize the following cross-entropy loss to maximize the likelihood of the ground-truth sentence continuation:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M} \sum_{t=1}^{M} \log\, p(y_t \mid y_{<t}, \mathbf{X})$$

In addition to encoding linguistically relevant neural information, $\mathbf{z}_{\text{brain}}$ needs to conform to the decoder model's embedding space. While this alignment can be learned jointly with the main $\mathcal{L}_{\text{CE}}$ objective, Ye et al. (2025) find that some of their configurations benefit from an additional warm-up phase, in which the projection layer is first trained to reduce the distance to the target embedding space. As such we report additional ablations including this warm-up phase in Appendix A.2.

Hyperparameters. Once again following Ye et al. (2025), the mini-batch size was set to 1, and the maximum number of epochs was set to 100. We used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-3}. To prevent overfitting, the dropout rate was set to 0.5. Training was terminated if no improvement in validation performance was observed for 10 consecutive epochs.

Evaluation Metrics. We employ the following standard metrics for text generation: **BLEU-1** (Papineni et al., 2002) to evaluate n-gram overlap based on "Precision" with a brevity penalty; **ROUGE-1/L** (Lin, 2004) to evaluate "Recall" at the levels of unigrams and the longest common subsequence; and **WER** (Morris et al., 2004) to measure the word-level error rate based on substitutions, deletions, and insertions. For testing the statistical significance of metric differences, we employ the two-sided Wilcoxon signed-rank test with Pratt handling

of zeros (Pratt, 1959), continuity correction, and Benjamini-Hochberg FDR correction for multiple comparisons (Benjamini and Hochberg, 1995).

5. Results

Table 5 presents the results for our four encoder configurations. Across all metrics, the two encoder designs incorporating expert-driven brain atlas inputs (C and D) consistently outperform the purely data-driven PCA approaches (A and B). Surprisingly, the hybrid BFM approach yields no statistically significant gains over using only the brain atlas mapping. Both Brain Atlas and BFM further significantly outperform the control settings (E and F), implying that information from the brain signal is being used to improve decoding performance.

We note that the absolute metric levels (e.g., WER ≈ 0.94) are consistent with the current state of the art in non-invasive fMRI-to-text decoding (Tang et al., 2023; Ye et al., 2025), and reflect inherent limitations of the recording modality: fMRI integrates neural activity via an indirect hemodynamic proxy, which is limited to a time-resolution of 1.5s per volume. In stark contrast, upper bound performance on brain-to-text decoding can reach down to 3% WER (Makin et al., 2020), but only if invasive electrodes are directly implanted in the brain (iEEG with a time-resolution of 200ms), the task is speaking the text (and not listening), and the output vocabulary stems from a closed set.

Brain Atlas significantly outperforms both PCA baselines, and control settings. The Brain Atlas encoding approach achieves BLEU-1 = 0.1276, ROUGE-1 = 0.1186, ROUGE-L = 0.1115, and WER = 0.9439. Compared to PCA-1000, which achieved the highest performance in Ye et al. (2025), Brain Atlas sees improvements of $+16.9\%$ in BLEU-1, $+14.0\%$ in ROUGE-1, and $+13.5\%$ in ROUGE-L. Similarly, the relative improvement to PCA-424 is $+11.2\%$, $+8.1\%$, and $+7.6\%$ respectively. At the same time, Brain Atlas also outperforms both control settings, showing that the brain signal it encodes contributes to improvements in decoding performance. All improvements remained significant after FDR correction at $q < 0.05$ (see * in Table 5). To better understand why the Brain Atlas encoding yields these gains, we conduct multiple ablation studies in Section 6.1.

BFM shows no significant improvement over Brain Atlas. BFM, which utilizes an initial brain atlas mapping before further embedding the data using a pre-trained Transformer, achieved BLEU-1 = 0.1282, ROUGE-1 = 0.1195, ROUGE-L = 0.1121, and WER = 0.9435. Compared to Brain Atlas alone, this leads to marginal numerical increases of $+0.5\%$ (BLEU-1), $+0.8\%$ (ROUGE-1), and $+0.5\%$ (ROUGE-L). However, none reached significance after FDR correction. Validation loss was also comparable at 4.870 (C) vs. 4.884 (D). As such, despite being the most data-intensive method, appending a pre-trained BFM does not appear to improve the quality of neural representations for language tasks. We further analyze potential reasons behind this discrepancy in Section 6.2.

PCA-424 significantly outperforms PCA-1000. Counter-intuitively, reducing the number of PCA components from 1,000 to 424 also yielded significant improvements across all metrics after FDR correction at $q < 0.05$ (see † in Table 5). This result further helps disentangle the contributions of dimensionality and encoding strategy to the overall A $\rightarrow$ C improvement. Since both B and C use $d = 424$ dimensions, the B $\rightarrow$ C gap directly measures the benefit of expert-driven encoding. In practice, the effect of changing the encoding strategy ($\Delta\text{Loss}_{\text{B}\rightarrow\text{C}} = -0.623$) is roughly twice as large as the dimensionality-reduction effect alone ($\Delta\text{Loss}_{\text{A}\rightarrow\text{B}} = -0.319$), with both differences remaining independently significant after FDR correction. In Section 6.1, we further examine the mechanistic basis behind the difference in performance between PCA-424 and PCA-1000, as well as between PCA and Brain Atlas.

Qualitative Examples. Table 2 presents two representative decoded outputs from Brain Atlas. Additional qualitative examples are provided in Appendix Table 5. In Example 1, the model correctly recovers *heart* and *pounding*; in Example 2, it correctly recovers *know*. In both cases, generation then drifts into repetitive continuation, a known degeneration pattern of autoregressive decoding (Holtzman et al., 2020; Welleck et al., 2020), which here likely indicates limited additional neural evidence beyond the initially matched tokens. While the overall quality of the results reflects the general difficulty of open-ended sentence recovery from neural signals (Tang et al., 2023; Ye et al., 2025), these examples, together with the quantitative metric improvements, suggest that expert-driven encoding supports the recovery of key lexical items from non-invasive neural signals.

6. Analysis and Discussion

Our results indicate that expert-driven encoder designs significantly outperform purely data-driven approaches. Simultaneously, we find that adding the hybrid BFM on top of the Brain Atlas encodings does not significantly improve decoding performance, despite the added computational complexity. To better understand the reasons underlying

	Cond.	BLEU-1 ↑	ROUGE-1 ↑	ROUGE-L ↑	Loss ↓	WER ↓
Encoders	(A) PCA-1000	0.1092	0.1040	0.0982	5.812	0.9513
	(B) PCA-424	0.1148†	0.1097†	0.1036†	5.493†	0.9486†
	(C) Brain Atlas	0.1276*	0.1186*	0.1115*	4.870*	0.9439*
	(D) BFM	0.1282*	0.1195*	0.1121*	4.884*	0.9435*
Controls	(E) LLM-only	0.1154	0.1067	0.1045	5.079	0.9481
	(F) Random	0.1148	0.1094	0.1029	5.502	0.9507

Table 1: Brain-to-text decoding performance across encoder designs and control conditions. Conditions (E) and (F) serve as controls that establish, respectively, the language-model-only baseline and the effect of an uninformative brain embedding. The gap between (C) and (E) quantifies the information gain attributable to genuine neural signals (ΔLoss $= -0.209$ nats ≈ 0.30 bits/token). For statistical significance among encoder designs (A–D), * indicates significantly better than both (A),(B),(E),(F); † indicates significantly better than (A); all after FDR correction at $q < 0.05$.

Context	Prediction	Ground Truth
…earth in dead silence it is really quiet so i can really hear my	**heart pounding** and then i hear the sound of my **heart pounding** and then i hear the sound of my **heart pounding** and then i hear the sound of my **heart pounding**	**heart pounding**
…you the straps and you	**know** what i'm talking about you know what i'm talking about you know what i'm talking about you know what i'm talking about you know what i'm	**know** through the

Table 2: Qualitative output examples from the Brain Atlas encoder design. **Blue bold** indicates words matching the ground truth.

these findings, we compare the stability of each encoding strategy across subjects and time. Additionally, we analyze where in the brain the signal used by each encoder originates from.

6.1. How Does Expert Knowledge Improve Encoding Quality?

Inter-Subject Correlation (ISC) measures the Pearson correlation of brain encodings between pairs of subjects listening to the same auditory stimulus. High ISC indicates that the representation captures stimulus-driven, reproducible neural responses rather than subject-specific noise. For each story, we computed ISC across every subject pair, averaging across all dimensions with non-zero standard deviation. The mean ISC of Brain Atlas was 0.1090, approximately $10.8\times$ higher than PCA-424 (0.0101) and $22.7\times$ higher than PCA-1000 (0.0048). This large gap indicates that ROI-based averaging in Brain Atlas preserves neural responses that are reliably evoked across subjects by the same stimulus. Consequently, the projection layer receives a consistent, cross-subject signal that is easier to map to language representations.

Temporal Autocorrelation. Because fMRI measures neural activity indirectly via the BOLD (Blood-Oxygenation-Level-Dependent) signal, brain activations are intrinsically temporally smooth due to the hemodynamic response function (i.e., a gradual increase and decrease in oxygen levels). As such, a high lag-1 autocorrelation corr(x_t, x_{t+1}) for each active dimension indicates a more physiologically plausible signal. The mean autocorrelation of Brain Atlas was 0.5753, markedly higher than PCA-424 (0.4870) and PCA-1000 (0.3016). The decrease in correlation with increasing PCA dimensionality points to the progressive incorporation of high-frequency noise components into higher-order principal components—a factor which we analyze in greater detail next.

Additional Noise in PCA-1000 versus PCA-424. Despite $2.36\times$ as many dimensions, PCA-1000 gains only 7.3 percentage points of explained variance (the fraction of voxel variance captured by PCA), and even underperforms the random baseline. Per-voxel contribution maps in Appendix Figure 6, showing how much each voxel drives the components, further reveal that this marginal gain is concentrated at the extra-brain periphery. This indicates that head-surface artifacts and physiological noise rather than genuine neural signals are being captured by the additional feature dimensions. These noise-derived components elevate the overall contribution floor in PCA-1000, suppressing the contrast of truly informative regions (cerebellum, basal ganglia, occipital lobe) and make the decoder's learning problem harder (see Appendix A.3).

Expert versus Data-driven. Taken together, these two metrics demonstrate that dimensionality reduction preserving the anatomical spatial structure of the brain substantially contributes to brain-to-text decoding performance. PCA extracts information based on variance maximization, without explicitly accounting for anatomical topography or local spatial structure, allowing noise to dominate higher-order components. In contrast, Brain Atlas compresses signals to 424 dimensions via ROI-based averaging, preserving functional and anatomical organization. This delivers more consistent cross-subject neural responses and more physiologically plausible temporal dynamics than either PCA baseline. This anatomically coherent representation likely contributes to the stabilization of the projection into the LLM embedding space and enables higher-quality decoding.

6.2. Why Does BFM Not Improve over Brain Atlas?

Since Brain Atlas substantially outperforms both PCA baselines, a hybrid approach combining an expert-driven encoding with a data-driven brain foundation model pre-trained on large-scale neuroimaging data should intuitively yield further performance improvements. However, BFM provides no significant improvement over Brain Atlas.

Inter-Subject and Temporal Correlation. The mean ISC of BFM is 0.2718—approximately $2.5\times$ higher than Brain Atlas (0.1090)—yet this apparent advantage does not translate into better decoding performance. Similarly, BFM shows an extremely high lag-1 autocorrelation of 0.8987—far above the physiologically plausible BOLD range of Brain Atlas (0.5753) and the highest value among all configurations—with its temporal correlation uniquely narrow and concentrated near 1.0, with near-zero variance across dimensions (Appendix Figure 4). This indicates that BFM output features are uniformly over-smoothed and show essentially no dimension-wise variation in temporal regularity. Together with the ISC finding, this pattern suggests that BFM's representation collapses into a highly redundant signal. We hypothesize this behavior to be an artifact of BFM's attention design, which pools activations across both the temporal, and spatial dimensions (i.e., different brain regions). The high ISC and autocorrelation thus reflect over-smoothing rather than genuine neural consistency, directly limiting the projection layer's ability to map the representation to language space (see Appendix A).

Regional Contributions. To identify which brain regions drive decoding performance in Brain Atlas versus BFM representations, we conducted an ablation analysis, selectively masking activations from different brain regions. We group the original 424 ROIs into 24 functional subnetwork groups (e.g., Language Default Mode, Auditory Somatomotor, Subcortical Cerebellum) according to the AA-24 hierarchy (Akiki and Abdallah, 2019). Next, we mask each of these 24 groups and measured the resulting change in test loss ($\Delta\text{Loss}_g = \mathcal{L}_{\text{ablated}}(g) - \mathcal{L}_{\text{baseline}}$). Positive ΔLoss indicates a positive contribution to decoding, whereas negative ΔLoss indicates interference.

Brain Atlas exhibits a strikingly sparse importance profile, with only 5 of 24 groups showing non-zero ΔLoss (Figure 3). This means these top-5 groups contribute positively, while all other groups have no effect on decoding performance. Notably, the *Language Default Mode Network* (lang DM; superior temporal and inferior frontal gyri) ranks second at $\Delta\text{Loss} = 0.023 \pm 0.004$, confirming that Brain Atlas selectively captures signals from established neural correlates of language comprehension. This sparse, reliably positive profile simplifies the downstream projection task by providing a clean, discriminative input with no conflicting signals to suppress.

In contrast, BFM engages all 24 groups, with 8 showing *negative* ΔLoss, indicating that a substantial portion of its representation actively impedes language decoding (see Appendix A.4). Furthermore, while 3 out of 5 top contributing groups overlap across Brain Atlas and BFM, the latter's groups have a maximum contribution that is 46.5% smaller, while specifically not including the lang DM.

Future Directions for Hybrid Encoders. The success of Brain Atlas highlights expert-driven, anatomically grounded feature aggregation as a simple, performant, and robust encoder design. For hybrid encoders to match this, BFM representations would need equivalent spatial selectivity and linguistic focus. Pretraining on linguistic brain data, fine-tuning, and task-adapted brain atlases that more directly emphasize language-processing regions may all improve future hybrid encoder design, similarly to advances in the text-only modality (Devlin et al., 2019). Finally, a more granular encoding of the time-dimension—e.g., via a learned aggregation such as attention pooling—may also help retain important information for language decoding.

7. Conclusion

This study investigated the effect of different encoder design strategies on brain-to-text decoding, contrasting data-driven PCA baselines against an expert-driven, anatomically informed mapping via the AAL-424 brain atlas. We additionally evaluated a hybrid encoder that supplements the same atlas-

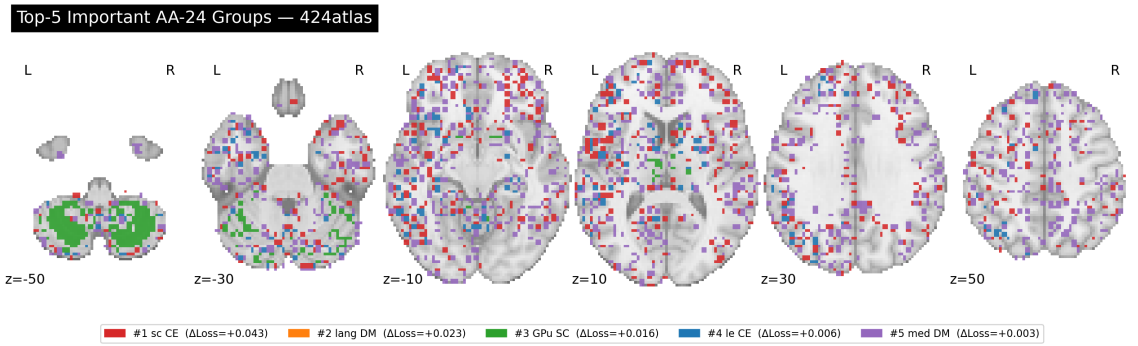


Figure 3: Top-5 contributing AA-24 groups for Brain Atlas (axial slices, $z = -50$ to $+50$ mm): (1) Subcortical Cerebellum (sc CE, $\Delta\text{Loss} = +0.043$, red); (2) Language Default Mode Network (lang DM, $\Delta\text{Loss} = +0.023$, orange); (3) Globus Pallidus/Putamen (GPu SC, $\Delta\text{Loss} = +0.016$, green); (4) Left Cerebellum (le CE, $\Delta\text{Loss} = +0.006$, blue); (5) Medial Default Mode Network (med DM, $\Delta\text{Loss} = +0.003$, purple). The spatial distribution confirms anatomically coherent localization of the most predictive regions.

based representations with a pre-trained brain foundation model, combining expert-driven and large-scale data-driven encoding.

The Brain Atlas approach significantly outperformed both control settings, as well as both PCA baselines on every evaluation metric, demonstrating that its expert-driven, anatomically grounded dimensionality reduction actively encodes brain signal information which substantially benefits decoding performance. Inter-subject correlation and temporal autocorrelation analyses further confirm that ROI-based averaging preserves more stimulus-driven, cross-subject reproducible neural responses and more BOLD-like temporal dynamics than either PCA baseline.

Introducing a BFM on top of the Brain Atlas encoding, however, yielded no statistically significant improvements. Regional ablation analysis reveals that BFM features exhibit a dense, mixed-sign contribution profile that impedes the projection layer, in contrast to the sparse, reliably positive profile of Brain Atlas. This highlights developing language-specialized BFMs as an important future challenge.

Taken together, our findings suggest that encoder-side representation quality—particularly expert-driven, anatomically grounded dimensionality reduction—plays a more decisive role in brain-to-text decoding than model scale alone. More broadly, the results suggest that the choice of dimensionality-reduction strategy is a key factor to improving brain-to-text decoding performance.

8. Limitations

First, due to implementation constraints, our experiments were conducted on five stories sourced from the Narratives dataset (Nastase et al., 2021), using a single decoder (`GPT-2 small`). While we were able to closely reproduce, and thus compare our methods with, the results of (Ye et al., 2025), generalizability to other fMRI datasets, stories with a different narrative structure, as well as additional languages, or decoder architectures remains to be established. Similarly, we employ only one type of brain atlas (AAL-424). This allows us to compare the effect of adding a BFM, which uses the same initial anatomical mapping (Caro et al., 2024). Given AAL-424's effectiveness, future work would likely benefit from investigating other expert-driven mappings of functionally grouped brain regions.

Second, the task of brain-to-text decoding currently lacks standard training procedures. As such, we make certain design decisions based on prior work, but do not exhaust all hyperparameter possibilities. Specifically, for extracting a brain signal embedding from BFM, the mean pooling strategy used to aggregate its spatial token representations is a simple but potentially suboptimal design choice. As proposed in Section 6.2, alternative aggregation strategies such as attention pooling thus warrant investigation in future work. Additionally, we use BFM in a frozen state. Depending on the downstream task, this is the intended use of the model (Caro et al., 2024), however, due to its non-linguistic pretraining, full fine-tuning, may still yield different results. Similarly, the projection layer has been found to benefit from an additional warm-up phase to explicitly align it with the embedding space of the decoder (Ye et al., 2025). In Appendix Table 3, we find that adding this representational alignment pretraining stage retains the performance ordering from our main results in Section 5 sans statistical significance. This suggests that the MSE alignment objective can compensate for encoder-side representation quality to some degree. Nonetheless, the fact that Brain Atlas outperforms the other encoder designs, even without additional pretraining, highlights the cost-effectiveness of an informative encoding strategy. More broadly, an important open

question is whether a fully learnable encoder could further improve performance beyond the frozen configurations studied here. Our frozen-encoder design isolates the effect of the encoding *strategy* itself from that of task-specific optimization, but end-to-end fine-tuning on language tasks may unlock complementary gains, particularly as larger linguistic fMRI datasets become available.

Third, all primary metrics are reported as averages across 27 subjects, and subject-level variance is not explicitly characterized. Although the Wilcoxon signed-rank test operates on per-subject scores and is therefore sensitive to the distribution of individual differences, it is possible that the observed group-level advantages are driven by a subset of subjects with particularly strong neural signal quality. Future work should examine per-subject performance distributions to assess the consistency of the reported effects.

Fourth, the BFM encoder processes brain recordings in non-overlapping chunks of 200 timepoints (300 s at repetition time (TR) = 1.5 s) via global self-attention. Consequently, the BFM representation associated with a given word w_t can incorporate neural signals from later timepoints within the same chunk, potentially spanning several minutes of subsequent story content. This introduces substantially greater temporal leakage than in the Brain Atlas and PCA baselines, which are computed independently at each TR. Strictly speaking, this means that BFM and Brain Atlas are not compared under equally causal configurations. However, our goal is not real-time decoding, but to test whether large-scale data-driven pretraining improves the quality of brain-to-text mapping. From that perspective, the null result is even more notable: despite enjoying a temporal information advantage, BFM still does not outperform Brain Atlas.

9. Ethics Statement

All fMRI data used in this study are drawn from the publicly available Narratives dataset (Nastase et al., 2021), which was collected under institutional ethical approval with informed consent from participants. No new human subjects experiments were conducted in this work. The dataset does not contain personally identifiable information.

10. Acknowledgements

We thank the anonymous reviewers for their helpful feedback. This work was supported by JSPS KAKENHI Grant Number JP24H00087, JST PRESTO Grant Number JPMJPR21C2, JST CREST Grant Number JPMJCR2565, JST BOOST Grant Number JPMJBY24B2, and Carlsberg Foundation Grant Number CF-25-0624.

11. Bibliographical References

Teddy J Akiki and Chadi G Abdallah. 2019. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific reports*, 9(1):19290.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Rasmus M Birn. 2012. The role of physiological noise in resting-state functional connectivity. *NeuroImage*, 62(2):864–870.

Felix Carbonell, Pierre Bellec, and Amir Shmuel. 2011. Global and system-specific resting-state fMRI fluctuations are uncorrelated: principal component analysis reveals anti-correlated networks. *Brain Connectivity*, 1(6):496–510.

Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed A Rizvi, Matteo Rosati, Christopher Averill, James L Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, and David van Dijk. 2024. BrainLM: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 2023. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22710–22720.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

John-Dylan Haynes and Geraint Rees. 2006. Decoding mental states from brain activity in humans. *Nature reviews neuroscience*, 7(7):523–534.

James B Heald, Daniel M Wolpert, and Máté Lengyel. 2023. The computational and neural bases of context-dependent learning. *Annual Review of Neuroscience*, 46(1):233–258.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Joseph G Makin, David A Moses, and Edward F Chang. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582.

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

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings of Interspeech 2004*, pages 2765–2768. ISCA.

Kevin Murphy and Michael D Fox. 2017. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *NeuroImage*, 154:169–173.

Samaneh Nemati, Teddy J Akiki, Jeremy Roscoe, Yumeng Ju, Christopher L Averill, Samar Fouda, Arpan Dutta, Shane McKie, John H Krystal, JF William Deakin, et al. 2020. A unique brain connectome fingerprint predates and predicts response to antidepressants. *IScience*, 23(1).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Xiaomei Pei, Dennis L Barbour, Eric C Leuthardt, and Gerwin Schalk. 2011. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8(4):046028.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

John W Pratt. 1959. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):655–667.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language-models.pdf. OpenAI technical report.

Yu Takagi and Shinji Nishimoto. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14453–14463.

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

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. 2024. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020.

Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. 2025. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1):346.

12. Language Resource References

Esteban, Oscar and Markiewicz, Christopher J and Blair, Ross W and Moodie, Craig A and Isik, A Ilkay and Erramuzpe, Aliaga and Kent, James D and Goncalves, Mathias and DuPre, Elizabeth and Snyder, Madeleine and others. 2019. *fMRIPrep: a robust preprocessing pipeline for functional MRI*. Nature Methods. PID https://doi.org/10.1038/s41592-018-0235-4. Software resource.

Nastase, Samuel A and Liu, Yun-Fei and Hillman, Hanna and Zadbood, Asieh and Hasenfratz, Liat and Keshavarzian, Neggin and Chen, Janice and Honey, Christopher J and Yeshurun, Yaara and Regev, Mor and others. 2021. *The Narratives fMRI dataset for evaluating models of naturalistic language comprehension*. Scientific Data. PID https://doi.org/10.1038/s41597-021-01033-3. Dataset resource.

Appendix

A. Additional Ablation Analyses of Encoder Signal Quality

A.1. Details on Inter-Subject and Temporal Correlation

The performance gap between Brain Atlas and the PCA baselines in Table 5 motivates a closer look at the statistical quality of each encoder's input representation. We measure two complementary properties—inter-subject correlation (ISC) and lag-1 temporal autocorrelation—that respectively reflect stimulus-driven reliability across subjects and physiological plausibility of the BOLD signal profile (Figure 4). Brain Atlas achieves the highest ISC (0.1090) and autocorrelation (0.5753), confirming that expert-driven parcellation yields cleaner, more reproducible signals than either PCA baseline. Notably, BFM's extreme autocorrelation (0.8987) and elevated ISC (0.2718) indicate over-smoothing rather than genuine neural fidelity, consistent with its failure to improve over Brain Atlas in decoding performance (Section 6.2).

A.2. Projection pretraining

We additionally test whether warm-up pretraining of the projection layer—an MSE-based phase that first aligns the projection layer's outputs to the LLM embedding space (Ye et al., 2025)—can compensate for encoder-side signal quality differences (Table 3). With warm-up, the performance rank of all four encoders remains the same as without, however, they lose statistical significance. This suggests that representational alignment can partially substitute for encoder quality. The fact that the Brain Atlas can achieve these performance gains without added training complexity nonetheless highlights the importance of including expert knowledge in the encoder's design.

A.3. PCA-424 versus PCA-1000

Signal-to-Noise Ratio. The story-averaged cumulative explained variance ratio (EVR) of PCA-424 was 63.9%, compared to 71.2% for PCA-1000 (Figure 5). Despite a 2.36-fold increase in the number of components, the gain in EVR is only 7.3 percentage points, i.e., 576 additional components account for merely 7.3% of total variance. To determine what this additional variance represents, we analyzed per-voxel contributions to PCA components. Mean per-voxel contributions were comparable between PCA-424 and PCA-1000 for both brain voxels (5.6 vs. 5.4×10^{-4}) and non-brain voxels (3.6 vs. 3.7×10^{-4}). However, the contribution difference map (PCA-1000 $-$ PCA-424; Figure 6, bottom; red = PCA-1000 dominant, blue = PCA-424 dominant) reveals that red predominates at the extra-brain periphery. This indicates that the additional components in PCA-1000 are largely derived from non-brain noise: by the variance-maximization principle, PCA assigns components to extra-brain voxels whose large variance arises from head-surface artifacts and physiological noise. The minimal information gain (7.3% EVR) is therefore further devalued by being concentrated in non-brain noise rather than genuine neural signals.

Contrast within the Brain. Voxel contribution maps (Figure 7) show that both PCA-424 and PCA-1000 concentrate high-contribution voxels in the same regions—the cerebellum, basal ganglia, and occipital lobe. The key difference lies not in which regions are targeted, but in the *contrast* between high and low-contribution regions. PCA-424 exhibits sharper contrast, with high-contribution voxels clearly prominent against surrounding areas. PCA-1000 shows a more homogeneous distribution in which the salience of high-contribution voxels

is relatively suppressed. This is consistent with the noise allocation finding: noise-derived components in PCA-1000 elevate the overall contribution floor within the brain, obscuring truly informative voxels. The decoder trained on PCA-424 inputs therefore receives clearer signals about which voxels to prioritize, contributing to more stable learning and improved decoding performance.

A.4. Details on Regional Ablations

To identify which brain regions contribute to decoding performance and to explain why BFM does not improve over Brain Atlas, we conducted a region-level ablation analysis. Activations in each of the 24 functional subnetwork groups (AA-24 hierarchy; Akiki and Abdallah, 2019) were selectively masked, and the resulting change in validation loss ($\Delta\text{Loss} = \mathcal{L}_{\text{ablated}} - \mathcal{L}_{\text{baseline}}$) was measured (Table 4, Figure 9). Positive ΔLoss indicates that the masked group is relied upon for decoding; negative values indicate that it actively impedes decoding. Brain Atlas exhibits a sparse, positive-only profile in which only 5 of 24 groups carry non-zero importance. Its second highest contributing group is the Language Default Mode Network (superior temporal and inferior frontal gyri), directly reflecting established neural correlates of language comprehension. In contrast, BFM distributes importance across all 24 groups with 8 showing negative ΔLoss, suggesting that its dense, entangled representation introduces irrelevant or conflicting signals that hinder the projection into language space. The brain maps below confirm the anatomically coherent localization of the most predictive regions for each encoder.

B. Qualitative Decoding Examples

To complement Table 5's quantitative evaluation, Tables 5–8 present randomly selected decoding outputs from all four encoder configurations alongside the corresponding ground-truth transcripts. Table 5 shows 20 examples from (C) Brain Atlas, while Tables 6, 7, and 8 each show 10 examples from (A) PCA-1000, (B) PCA-424, and (D) BFM, respectively. Examples are drawn from the held-out test portions of all five stories to provide a representative sample of decoder behavior across diverse narrative contexts. While BLEU-1 and ROUGE scores are modest in absolute terms (but equivalent to prior work, such as Ye et al., 2025), the qualitative outputs show partial thematic alignment with the reference texts, particularly in content words and broad topic area. A notable qualitative difference is that Brain Atlas outputs tend to exhibit less degenerate repetition than the other encoders: PCA-424, PCA-1000, and BFM predictions frequently collapse into repetitive token or phrase loops (e.g., wait wait wait …", it's not your fault it's not your fault …"), whereas Brain Atlas outputs more often form grammatically coherent sentences with varied vocabulary, suggesting greater output diversity and naturalness.

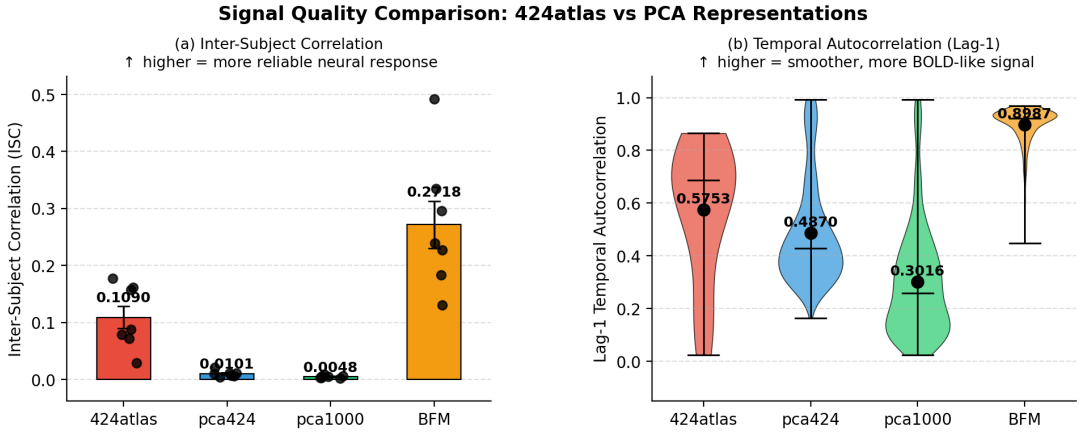


Figure 4: Signal quality comparison. **(a)** *Inter-Subject Correlation (ISC)*: ISC means are C = 0.1090, B = 0.0101, A = 0.0048, and D = 0.2718. Each dot represents the per-story ISC; error bars indicate standard error margin across stories. Higher ISC indicates more stimulus-driven, cross-subject reproducible neural responses. **(b)** *Lag-1 Temporal Autocorrelation*: means are C = 0.5753, B = 0.4870, A = 0.3016, and D = 0.8987. Violin plots reflect per-dimension distributions, with dots denoting means. Section 6 discusses how the mid-level correlations of Brain Atlas translate into better performance compared to PCA and BFM.

configuration	BLEU-1	ROUGE-1	ROUGE-L	Val. Loss	WER (↓)
(A) PCA-1000	0.1243	0.1170	0.1101	5.088	0.9444
(B) PCA-424	0.1252	0.1177	0.1106	5.019	0.9441
(C) Brain Atlas	0.1265	0.1180	0.1109	4.844	0.9441
(D) BFM	0.1255	0.1165	0.1093	4.868	0.9449

Table 3: Evaluation results with warm-up training (MSE alignment + cross-entropy) according to Ye et al. (2025). While the performance ordering of Table 5 is roughly maintained, no statistically significant pairwise differences are observed. This pattern suggests that warm-up absorbs signal-quality differences introduced by dimensionality reduction strategies.

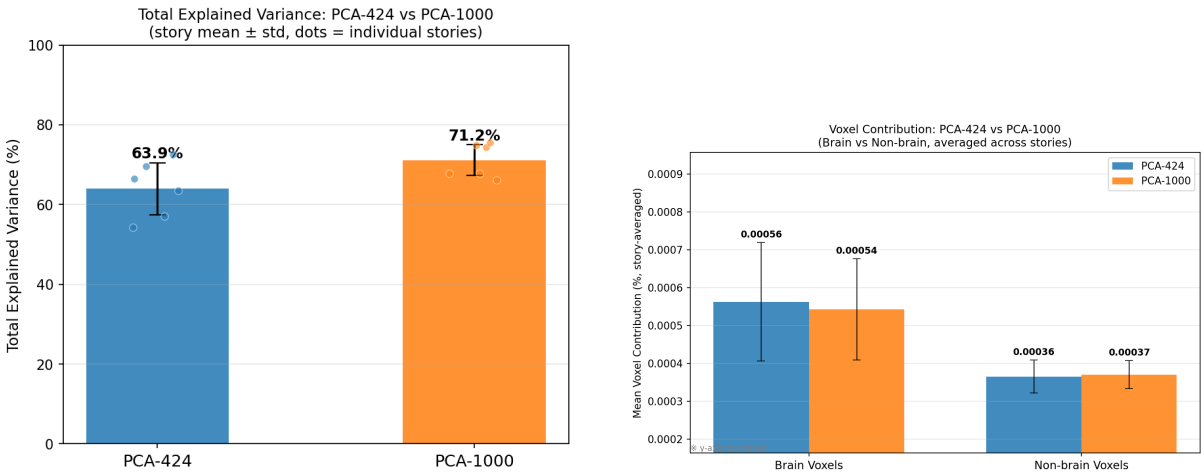


Figure 5: Left: Story-averaged total explained variance ratio (EVR) for PCA-424 (63.9%) and PCA-1000 (71.2%). Despite a 2.36-fold increase in dimensionality, the EVR gain is only 7.3 percentage points. Right: Mean per-voxel contribution for brain vs. non-brain voxels (PCA-424 and PCA-1000).

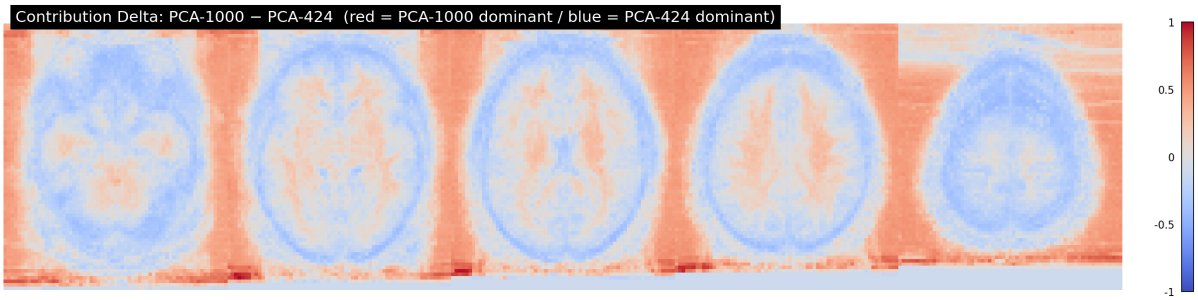
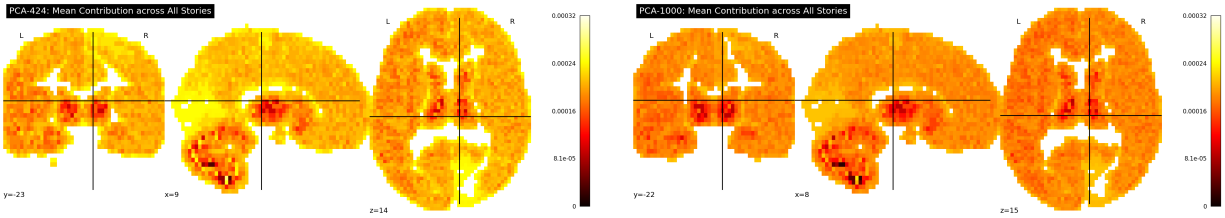


Figure 6: Contribution difference map (PCA-1000 − PCA-424). Red indicates PCA-1000-dominant regions; blue indicates PCA-424-dominant regions. Red concentration at the brain periphery shows that the additional PCA-1000 components largely reflect extra-brain noise.



(a) PCA-424 (b) PCA-1000

Figure 7: Mean voxel contribution maps across all stories for (a) PCA-424 and (b) PCA-1000. High-contribution voxels are concentrated in similar regions (cerebellum, basal ganglia, occipital lobe) in both configurations. PCA-424 shows sharper contrast between high- and low-contribution regions; PCA-1000 shows a more uniform, noise-elevated distribution.

Table 4: Top-5 AA-24 network groups ranked by ablation importance ($\Delta\text{Loss} = \mathcal{L}_{\text{ablated}} - \mathcal{L}_{\text{baseline}}$; mean $\pm$ standard error margin across 27 subjects). † indicates groups with positive ΔLoss in *both* models.

424atlas					**BFM**				
Rank	Group	ΔLoss			Rank	Group	ΔLoss		
1	sc CE†	$+0.043$	$\pm$	0.009	1	med DM†	$+0.023$	$\pm$	0.012
2	lang DM†	$+0.023$	$\pm$	0.004	2	sc CE†	$+0.016$	$\pm$	0.009
3	GPu SC†	$+0.016$	$\pm$	0.002	3	cent SM	$+0.011$	$\pm$	0.006
4	le CE†	$+0.006$	$\pm$	0.001	4	GPu SC†	$+0.010$	$\pm$	0.007
5	med DM†	$+0.003$	$\pm$	0.001	5	ri CE	$+0.006$	$\pm$	0.006

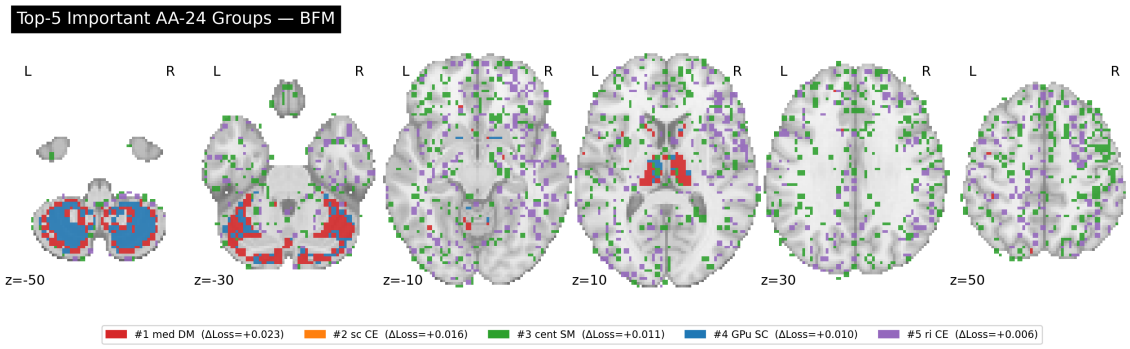


Figure 8: Top-5 contributing AA-24 groups for BFM projected onto MNI space (axial slices, $z = -50$ to $+50$ mm). Color indicates group identity (rank order): red = #1 med DM (Medial Default Mode Network, $\Delta\text{Loss} = +0.023$), orange = #2 sc CE (Subcortical Cerebellum, $+0.016$), green = #3 cent SM (Central Sensorimotor, $+0.011$), blue = #4 GPu SC (Globus Pallidus/Putamen, $+0.010$), purple = #5 ri CE (Right Cerebellum, $+0.006$). The spatial distribution confirms anatomically coherent localization of the most predictive regions.

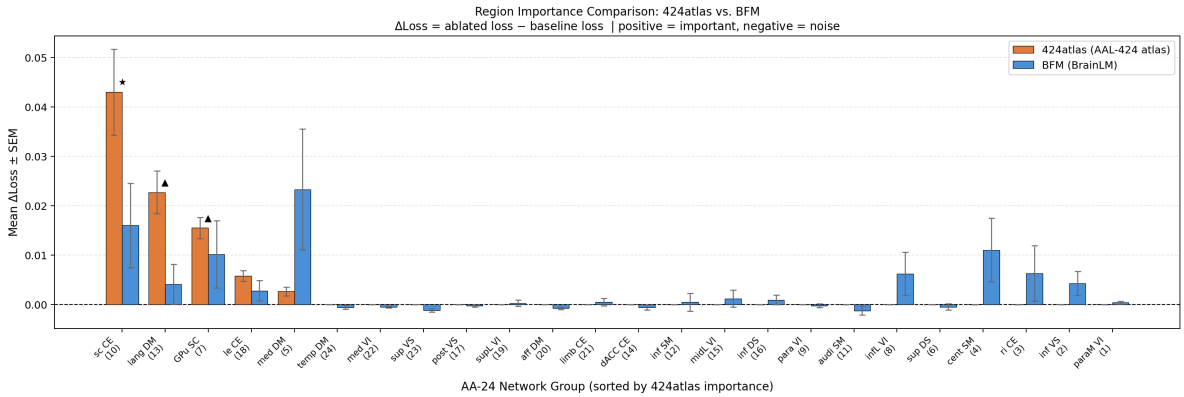


Figure 9: Ablation importance of AA-24 groups for Brain Atlas (orange, left bars in each pair) and BFM (blue, right bars in each pair). ΔLoss = ablated loss $-$ baseline loss; mean $\pm$ SEM across 27 subjects. Groups are sorted by Brain Atlas importance (left = most important). Positive values indicate reliance; negative values indicate noise contribution. Brain Atlas exhibits a sparse profile (5 of 24 groups non-zero), whereas BFM distributes importance broadly with 8 groups showing negative contributions. † Groups that are positive in both models.

Table 5: Random qualitative examples (1–20) from (C) Anatomical Atlas (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…date today is june	i'm going to look at the wall and see if there's a hole in the wall we're going to find out what it's all about. oh yeah	fifteenth yet i could swear
Narratives_060	…to be having a lot of nightmares	when i woke up in the morning and i woke up in the morning and i woke up in the morning and i woke up in the morning and i woke up	lately really
Narratives_065	…get in bed and relax tell the truth is it going to do you ...	if you don't tell the truth **and** i'm going **to** get out of here so i'll go back **to** my room **and** wait for you **to** come in **and**	**to** sit around **and** stew
Narratives_016	…alan did not want to	be in the room so he said you know what i'm going to do and i'm going to go out there and i'm going to tell you this story	part with one of his
Narratives_144	…what he did in	the war **he** killed people **and** you know i was living in the city of alan you know i lived in the city of alan you know i went to	brazil **and he** said
Narratives_041	…no i don't believe it	's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible	swanson
Narratives_052	…with a clever imitation guy mary i'm going look around a little more well ...	**right** i'm going to look around a little more well your dinner will be ready all **right** i'm going to look around a little more well your dinner will be	**right** save it me there are
Narratives_066	…she learned some two hundred fifty signs	that she had learned from her father **and** he was **the** one who taught her how to teach her how to do it. i'm not sure if you've ever	**and the** big question is
Narratives_134	…well uh you know but	**uh** you know um i'm not going to be able to go out and **uh** um um um um um um um um um um um um um um um um um um	**uh** the stress of suddenly i was
Narratives_058	…morning guy woke up	in the middle of the night and he was lying on the floor with his hands behind his back and he had no idea what was going on so he decided to	screaming
Narratives_026	…i must warn you that it's useless to buck us	**out** of this place we've got to get away from here you're going to have to be careful what you're doing there's no way we're going to	come **out** peacefully
Narratives_141	…new york rules he would get g corp that actually has a seven twenty ...	**that** he would have to **go** to and i was the first one to come out of the door and it was a really cool thing to be able to get in	which were worth even though you can **go** do **that** great
Narratives_137	…thing you know a southern guy with an	old lady and i was like oh yeah you know what i'm going to tell you this is gonna be the story of my life and i'm going to tell	eight year perfect attendance pin from
Narratives_111	…bob mean so all your friends are a why member of and are you ...	'm going to the hospital and i'm going to a hospital and i'm going to an emergency surgery and i'm going to be in the hospital and i'm	then probably would've never said you know when i grow
Narratives_052	…these new products the dream guy	i've been dreaming about for a long time and i've been dreaming about all the new things i've been dreaming about since i've been asleep i've been	went home that night
Narratives_144	…of this building and i went normal life and my	dad **and i** went **to** school **and** my dad was a teacher **and i** was in college **and i** was the first guy **to** go **to** university **and i** was going **to**	**to** bob **and i** said bob you know those funny little sheltered ...
Narratives_140	…of an apartment	building and **i** was living **in** the same apartment building as bob and **i** had been living there for a long time and **i** knew that bob was going to be living	**in** new york **i** guess you weren't

Subject	Context	Prediction	Ground Truth
Narratives_041	…i think it's some sort of mass hypnosis or something	like that but it's not something that we've been able to find out what **it** means they're trying to get rid of **us** so i think it's a	well why doesn't **it** work for **us**
Narratives_084	…you know just talking made now to go to space i said have you ...	-saving **surgery** and you've been in space and you've been **to** space and you've been **to** space and you've been **to**	financial crime no no i used **to** threatening **surgery** so what do ...
Narratives_061	…recommend against crossing that line but i had just	found out that i was a student at **the** university so i decided to give it a shot and i went to **the** dean's office where i learned that i had	seen **the** line

Table 6: Random qualitative examples (1–10) from (A) PCA-424 (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…excuse me sir it's quite alright	oh wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait	do remember the phone call
Narratives_060	…mr in eh no he had an appointment with mr	mr he had an appointment with m mr he had an appointment with mr he had an appointment with mr he had an appointment with mr	the human research institute
Narratives_065	…no i didn't arthur he said his eyes on the	wall and he said yeah i don't know where **the** arthur dead man is they haven<brain/> oh yeah i've been told by **the** police that there's a	far dim end of **the** room
Narratives_016	…alan did not want to	get hit by the landlord so he died and that's the story **of** the tragedy **of** the tragedy **of** the tragedy **of** the tragedy **of** the tragedy **of** the tragedy **of**	part with one **of** his
Narratives_144	…connecticut and i work for the new york times	you know **i** work for the new ynew york times you know **i** work for the new york times you know **i** work for the new york	and **i** was working a story one
Narratives_041	…the power and wash out the memory of the day you	have to go back to the beginning of the day you have to go back to the beginning of the day you have to go back to the beginning of the day	your friend swanson
Narratives_052	…concrete more	concrete you see concrete more concrete underneath but underneath you see concrete more concrete underneath you see concrete more concrete underneath but underneath you ...	metal and here on the
Narratives_066	…year old she was eating at the	**table** and she was sitting down **with** her hands behind her head and she was eating a cup of coffee and she had just gone to sit down **with** her hands behind	**table with** us forks
Narratives_134	…trouble and um and bob realized he says you can't do ...	you can't do that and bob realized he realized he realized he realized he realized he realized he realized he realized he realized he realized he realized ...	be the first suspect
Narratives_058	…mary	look what is it you want to know what is it here i'm looking out the door and there is a huge pile of strange materials that has been put together	mary

Table 7: Random qualitative examples (1–10) from (B) PCA-1000 (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…excuse me sir it's quite alright	what are you going to **do** it's okay I'm going to get out of here hurry up and get out of here let's get out of here hurry up	**do** remember the phone call
Narratives_060	…mr in eh no he had an appointment with mr	mr mr mr mr mr mr mr mr mr mr mr mr mr mr mr	the human research institute
Narratives_065	…no i didn't arthur he said his eyes on the	wall and he was leaning against **the** wall looking up at **the** ceiling where **the** wall had fallen onto **the** wall there were no arthur's eyes on **the** wall but	far dim end of **the** room
Narratives_016	…alan did not want to	be the **one** that was on the roof **of** the building at the time, so he said he had to go up there and they were all over the place and	part with **one of** his
Narratives_144	…connecticut and i work for the new york times	**and** now **i** am **working** for the new y you know **i** work for the new y times **and** now **i** am **working** for the new york times **and**	**and i** was **working** a story one
Narratives_041	…the power and wash out the memory of the day you	have to be very careful you don day you have to be very careful you don day you have to be very careful you don day you have	your friend swanson
Narratives_052	…concrete more	than concrete more than concrete more<brain/> cement more than concrete more concrete more than concrete more **metal** more than concrete more concrete more than concrete ...	**metal** and here on the
Narratives_066	…year old she was eating at the	**table** she was eating at the **table** she was eating at the **table** she was eating at<brain/> **table** she was eating at the **table** she was eating at the **table**	**table** with us forks
Narratives_134	…trouble and um and bob realized he says you can't do ...	you can't do that and bob realized he realized he realized he realized he realized he realized he realized he realized he realized he realized he realized ...	be the first suspect
Narratives_058	…mary	what is it look what is it look what is it look what is it look what is it look what is it look what is it look what is it look	mary

Table 8: Random qualitative examples (1–10) from (D) Brain Foundation Model (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…excuse me sir it's quite alright	you've got to be careful what you're going to **do** it's not your fault it's not your fault it's not your fault it's not your	**do** remember the phone call
Narratives_060	…mr in eh no he had an appointment with mr	he had an appointment with mr he had an appointment with m guess what he was going to do it's been a long time since you've been here	the human research institute
Narratives_065	…no i didn't arthur he said his eyes on the	girl and he was like oh yeah i don't know what you're going to do but i'm going to go to **the** hospital and i'm going to get	far dim end of **the** room
Narratives_016	…alan did not want to	go to the hospital but he was going to the hospital and he said i'm going to the hospital and i'm going to the hospital and i'm going to	part with one of his
Narratives_144	…connecticut and i work for the new york times	**and** i've been **working** for the new y you know it's **a** lot of money to be able to work in the city **and** uh yeah so I'm	**and** i was **working a** story one
Narratives_041	…the power and wash out the memory of the day you	've been waiting for the day you've been waiting for the day you've been waiting for the day you've been waiting for the day you've been waiting for	your friend swanson
Narratives_052	…concrete more	than anything else you've ever seen in **the** history of **the** world it's been a long time since we've had a chance to have an event like this one	metal and here on **the**
Narratives_066	…year old she was eating at the	kitchen **table** she was sitting on the couch and she was asleep she was asleep she was asleep she was asleep she was asleep she was asleep she was	**table** with us forks
Narratives_134	…trouble and um and bob realized he says you can't do ...	you can't do that and he says yeah i'm going to **be the first** one to go and um and uh and um and um and um and um and	**be the first** suspect
Narratives_058	…mary	**mary mary mary mary mary mary mary mary mary** m oh my god oh my god oh my god oh my	**mary**