SCRIBE: SCALABLE CONTEXTUAL REPRESENTATIONS OF INTERPRETABLE BRAIN-GROUNDED SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how the human brain constructs meaning from real-world language remains a fundamental challenge, as semantics depend on context that varies in natural settings. To enable the study of real-world linguistic meaning in the brain, we introduce SCRIBE, an approach for modeling context-sensitive meaning. Grounded in a neuroscience-informed semantic space and developed using an ensemble of open-source large language models, SCRIBE dynamically annotates meaning in context to produce interpretable, word-level semantic representations for an arbitrary text. Experiments using human judgments, controlled synthetic contexts, and brain activity recorded while participants listen to naturalistic stories, show that SCRIBE captures systematic context-dependent semantic shifts and predicts neural responses in language-selective brain regions, significantly outperforming non-contextual semantic representations. Variance partitioning further shows that SCRIBE captures unique brain-predictive information beyond syntactic features, supporting its ability to encode contextual semantic information relevant to the brain. Together, these results establish SCRIBE as a scalable and interpretable foundation for studying natural language semantics in the brain.

1 INTRODUCTION

Understanding how the human brain constructs meaning from language is a central goal in cognitive neuroscience and artificial intelligence. Semantic representations, which capture the meanings of words and concepts, are inherently dynamic and context-dependent. For example, the word "knife" may evoke harm and negative affect in a sentence such as "he cut himself with a knife", but a more neutral, functional meaning in "he chopped vegetables with a knife". While this context awareness is a defining property of natural language, it also makes semantics difficult to study in the brain. The effectively unbounded range of contexts makes it challenging to design experiments or obtain human annotations of semantics in context that capture meaning as it occurs in real-world language.

Prior work has primarily proposed semantic representations along two distinct directions. One line of research studies isolated concepts using feature-based representations grounded in cognitive and neuroscience theories (Binder et al., 2016; McRae et al., 2005; Devereux et al., 2014; Lynott et al., 2020). These representations are interpretable and have been successfully used to model semantic structure and predict brain activity (Mitchell et al., 2008; Anderson et al., 2017). However, these approaches typically rely on static, out-of-context representations of word meaning. Even when extended to sentence-level modeling (Anderson et al., 2017), they assume that word meanings remain fixed across contexts and combine them using simplified compositional schemes. In addition, these representations often rely on manual feature annotation and fixed vocabularies, making them difficult to scale to arbitrary natural language inputs.

A second line of work aims to study semantics using distributed representations derived from large language models (LLMs) (Devlin et al., 2019; Touvron et al., 2023; Jiang et al., 2023; Grattafiori et al., 2024). These representations are contextual, scalable, and have been shown to predict neural responses during natural language comprehension (Toneva & Wehbe, 2019; Pereira et al., 2018). However, because internal LLM representations entangle semantic, syntactic, and other lexical prop-

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.

erties of language, their alignment with the brain is difficult to attribute to semantic processing. As a result, they provide limited insight into how semantic meanings are represented in the brain.

In this work, we bridge these two approaches by introducing Scalable Contextual Representations of Interpretable Brain-grounded sEmantics (SCRIBE) to model context-dependent semantic representations that are both scalable and interpretable. SCRIBE builds on a neuroscience-informed semantic space derived from prior work (Binder et al., 2016), and extends it to natural language by using an ensemble of open-source LLMs to generate context-aware semantic features. We validate SCRIBE along four complementary dimensions. First, we show that SCRIBE scores significantly align with human ratings for isolated out-of-context concepts, establishing their semantic validity. Second, in a large-scale behavioral study with 500 recruited participants, we show that human judgments along interpretable semantic feature dimensions systematically vary with linguistic context, and that SCRIBE captures these shifts while closely aligning with human ratings across contextual conditions. Third, we demonstrate that contextual SCRIBE representations significantly outperform their out-of-context counterparts at predicting neural responses during naturalistic story comprehension and capture unique brain-predictive information beyond the closest interpretable stimulus representation, QAemb (Singh et al., 2025). Finally, using variance partitioning with independently defined syntactic features, we show that SCRIBE retains more unique brain-predictive variance than QAEmb while exhibiting less overlap with syntactic features in language-related regions, supporting greater semantic specificity of its neural predictivity. Together, these results establish SCRIBE as a scalable, interpretable, context-aware foundation for studying how the brain represents meaning in natural language.

2 Related work

Feature-based interpretable semantic representations. A central question in cognitive science is how to represent meaning, with a long line of work modeling concepts through explicit semantic features derived from human judgments and experiential attributes (McRae et al., 2005; Devereux et al., 2014; Lynott et al., 2020; Hoffman et al., 2013). Binder et al. (2016) consolidated this line of work into a comprehensive, neurobiologically grounded semantic space spanning sensory, motor, affective, social, and cognitive dimensions. Such interpretable representations have been used to model brain activity at both the word and sentence level (Mitchell et al., 2008; Fernandino et al., 2015; Anderson et al., 2017). However, these approaches rely on out-of-context representations and manually annotated datasets with limited vocabularies, limiting their scalability to settings where meaning varies across contexts. Our work leverages LLMs to generate context-aware semantic representations, enabling a scalable and semantically grounded feature space.

LLM-based contextual semantic representations. A second line of work models brain activity using contextualized LLM embeddings LLMs (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Goldstein et al., 2022), which capture a wide range of linguistic and semantic phenomena (Jawahar et al., 2019; Brown et al., 2020). Despite their impressive brain alignment, LLMs' distributed and high-dimensional representations are difficult to interpret and to relate explicitly to semantic processing in the brain. In fact, prior work has shown that the brain alignment of pretrained LLMs depends in part on syntax (Oota et al., 2023; Fresen et al., 2024). In contrast, we leverage LLMs as semantic annotators to map concepts to an interpretable and cognitively grounded feature space, which can then be used to predict brain activity.

Bridging interpretability and contextual semantic representations. Previous work has shown that semantic properties can be recovered from LLM representations using probes (Chersoni et al., 2021; Turton et al., 2021; 2020). However, these approaches do not assess whether modern LLMs can directly annotate such properties through through prompting, analagous to the human semantic annotation procedure used to construct semantic feature spaces. Moreover, although LLM representations are inherently context sensitive, these approaches average them across contexts rather than modeling how semantic properties vary with context (Turton et al., 2021). More recently, structured prompting has enabled LLMs to directly produce explicit, interpretable features, for example question-answering-based embedding methods such as QAEmb (Singh et al., 2025). QAEmb represents stimuli using binary features obtained from LLM responses to yes/no questions derived from plausible neurolinguistic theories. This makes QAEmb the closest methodological comparison to SCRIBE: both approaches use LLMs as annotators to transform natural-language stimuli into

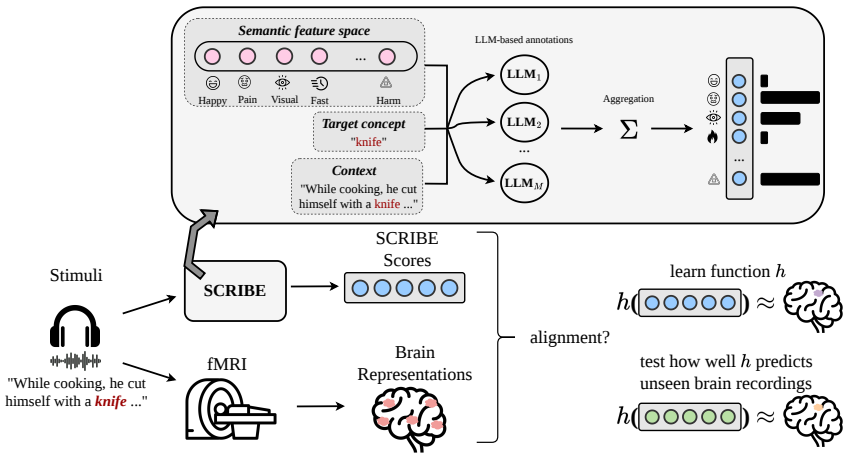


Figure 1: SCRIBE constructs interpretable, context-aware semantic features that are used to predict brain activity, enabling the modeling of the dynamic nature of meaning in the brain. It combines three information sources: a cognitively grounded semantic feature space (Binder et al., 2016), a target concept, and the linguistic context in which the concept appears. For each semantic property in the feature space, LLMs are prompted independently to assign a score reflecting how strongly the property applies to the concept in context. These property-wise scores are then aggregated across models to produce the final representations, which is used as input to a brain encoding model that predicts fMRI responses.

explicit, cognitively motivated, interpretable word-level features for brain encoding. They differ, however, in what these features are designed to represent. QAEmb provides a general framework spanning neuroscience-motivated linguistic properties, whereas SCRIBE is specifically designed to represent conceptual semantics within the independently defined neurocognitive semantic space of Binder et al. (2016), using LLMs to estimate how semantic properties vary with context.

Our approach bridges these directions by using LLMs as context-aware semantic annotators within a cognitively grounded feature space, yielding representations that are simultaneously interpretable, context-aware, and explicitly semantic.

3 METHODS

We first describe the components of SCRIBE (Section 3.1), and then explain how these are extended to annotate natural stimuli and to predict brain activity (Section 3.2).

3.1 SCRIBE: INTERPRETABLE AND CONTEXT-AWARE SEMANTIC REPRESENTATIONS

SCRIBE maps a target concept to a set of interpretable semantic features (Figure 1). It relies on three information sources: (i) an interpretable semantic feature space (Binder et al., 2016) that specifies the set of properties to be estimated, (ii) the concept to be annotated, and (iii) the context that the annotation should be conditioned on. SCRIBE prompts an LLM to estimate the degree to which each property in the semantic space applies to the target concept, in a specific context. These property-wise scores are then aggregated across models to produce the final semantic representation.

3.1.1 COGNITIVELY-GROUNDED, INTERPRETABLE SEMANTIC FEATURE SPACE

Semantic feature space. Our approach builds on the semantic feature space introduced by Binder et al. (2016), which provides a structured and interpretable representation of conceptual meaning. This feature space defines a set of 65 semantic properties that describe concepts in terms of neurocognitively grounded attributes, spanning sensory (e.g., *color*, *brightness*), motor (e.g., *body parts*, *movement*), affective (e.g., *pleasantness*), and higher-level cognitive dimensions (see Appendix A.1). These dimensions are derived from a synthesis of findings in cognitive neuroscience

and psychology, providing a neurocognitively grounded basis for representing meaning. Each dimension corresponds to a well-defined and interpretable aspect of semantic content.

This feature space has two key properties that are central to our approach: it is inherently interpretable, and it is grounded in neurocognitive evidence, making it well-suited for studying semantic processing in the brain. In SCRIBE, we treat this space as a fixed semantic basis over which meaning can be expressed and systematically modulated by context.

Human semantic ratings for out-of-context concepts. In addition to defining the semantic feature space, Binder et al. (2016) introduce a dataset of human ratings for out-of-context concepts. In this dataset, participants rated how strongly each of the 65 properties applies to a given concept on a 0–6 scale, covering 535 English words (see Appendix A.1 for details). Aggregating these ratings yields continuous, multi-dimensional representations that serve as reference semantic profiles for concepts in the absence of context. In SCRIBE, we use these human ratings as a reference for out-of-context meaning, defined over this semantic feature space.

3.1.2 LLM-BASED SEMANTIC ANNOTATIONS

A central component of SCRIBE is the use of LLMs to estimate semantic feature scores, replacing manual annotation with a scalable and automated procedure. This design enables efficient and scalable estimation of semantic features across diverse concepts and contexts (Grattafiori et al., 2024; Jiang et al., 2023), while preserving the interpretability of the underlying semantic space. Concretely, given the interpretable semantic space and a target concept, SCRIBE prompts the model independently for each semantic property in the feature space, asking it to estimate how strongly the property applies, producing a semantic vector by combining these property-wise scores. The prompt mirrors the annotation protocol of Binder et al. (2016) (see Appendix A.2.1), with prompt ablations showing that performance is largely robust to variations (Appendix A.2.2). To generate the semantic scores, we use recent open-source instruct-based LLMs: `llama3.1-8b`, `mistral-7b`, `qwen2.5-7b` and `llama3.1-70b` (Grattafiori et al., 2024; Jiang et al., 2023; Hui et al., 2024). SCRIBE then aggregates annotations across models by averaging their semantic representations to improve robustness, mitigating model-specific biases and producing more stable representations.

3.1.3 CONTEXT-AWARE SEMANTIC REPRESENTATIONS

A key contribution of SCRIBE is that it explicitly models context-dependent meaning. Because semantic meaning is inherently dynamic and shaped by linguistic context, assigning semantic property scores to concepts in isolation is insufficient. SCRIBE addresses this by conditioning semantic feature estimation on surrounding text, allowing the same concept to be assigned different semantic profiles depending on its usage. Crucially, by leveraging LLMs for this context-aware annotation, SCRIBE enables scalable estimation of semantic features across naturalistic stimuli, making it possible to model meaning as it occurs in real-world language. Concretely, given the interpretable semantic space, a target concept and the sentence in which it appears, the LLM is prompted for each semantic property to estimate how strongly it applies to the concept *in that context*, using the shared semantic feature space (see Appendix A.2.3 for prompt details). This yields a context-dependent semantic vector, while preserving the interpretability of the underlying feature space. By leveraging LLMs for this context-aware annotation, SCRIBE enables scalable estimation of semantic features over naturalistic stimuli, allowing meaning to be modeled as it occurs in real-world language.

3.2 MODELING SEMANTIC REPRESENTATIONS IN THE BRAIN

We next use SCRIBE representations as inputs to brain encoding models predicting fMRI activity during natural language comprehension, allowing us to assess their neural predictivity.

Brain encoding. We evaluate brain alignment using a subset of the *Moth Radio Hour* dataset (LeBel et al., 2023), which consists of fMRI recordings from 8 participants listening to autobiographical stories ($10 - 15$ minutes each), with images acquired every 2s (TR $= 2.0$s). This dataset provides naturalistic, context-rich narratives and is widely used in prior work (Moussa et al., 2024; Antonello et al., 2023; Oota et al., 2024), enabling direct comparison with existing models, including interpretable approaches such as QAEmb (Singh et al., 2025). We use 9 stories for training

and 1 story for evaluation (see Appendix A.4.1 for details). We adopt a standard voxel-wise encoding framework following prior work (Oota et al., 2024; Antonello et al., 2023). First, contextual word-level features are obtained from SCRIBE (Sections 3.1 and 3.2.1). These features are temporally aligned to the fMRI signal using Lanczos interpolation and augmented with time-lagged copies spanning the preceding 4s to account for the hemodynamic response. The resulting feature vectors are used to train linear encoding models, regularized with the ridge penalty, to predict voxel-wise neural responses. The ridge regularization parameter is chosen via cross-validation.

Noise ceiling. To account for variability in fMRI data, we use voxel-wise noise ceilings estimates following (LeBel et al., 2023; Moussa et al., 2024). We restrict all brain-encoding analyses to voxels with a noise ceiling above 0.4. Prediction accuracy is measured on held-out data using the Pearson correlation between predicted and observed responses, normalized by the voxel-wise noise ceiling. Statistical significance of differences is assessed using the Wilcoxon signed-rank test. Finally, to focus on semantically relevant brain regions, we report results both across all retained voxels and within language-selective regions defined by a standard cortical multi-modal parcellation (Glasser atlas) (Glasser et al., 2016). See Appendix A.4 for more details.

Baseline. We compare SCRIBE with QAEmb (Singh et al., 2025), the closest existing interpretable stimulus representation to our approach (see Section 2 for a discussion). QAEmb constructs interpretable word-level representations using LLM responses to neuroscience-motivated yes/no questions. QAEmb representations consists of 35 binary features automatically selected from an initial pool of 606 questions, with each dimension corresponding to a specific question. We use the publicly released QAEmb annotations.

3.2.1 Annotation design: word types and context size

Applying SCRIBE to annotate *semantics* in natural language requires specifying both *which words* are annotated and *how much contextual information* is incorporated.

Word types. In standard brain encoding pipelines, representations are computed for all tokens in a sequence, including both content and function words (Toneva & Wehbe, 2019). However, not all tokens contribute equally to semantic information. Function words (e.g., articles, prepositions, auxiliaries) carry limited lexical meaning and may introduce noise. We therefore restrict semantic annotation to content-bearing words, assigning zero-valued feature scores to all others. To determine the optimal set of annotated tokens, we perform an ablation over part-of-speech categories, considering nouns, proper nouns, verbs, adjectives, adverbs, numerals, and pronouns, as well as more restrictive subsets (e.g., nouns only, or nouns, verbs, and adjectives as in Binder et al. (2016)). Results (Appendix B.3.1) show that, for both individual models and the ensemble, using the broader set yields the best performance, while more restrictive configurations lead to reduced performance, likely due to increased sparsity in the resulting representations.

Context size. For each target word, the amount of surrounding context used for semantic estimation can vary. In this work, we explore a range of context sizes, from small windows of k surrounding tokens that capture local context to larger windows that include all preceding context up to the occurrence of the target word. Across both individual models and the ensemble, we find that a context window of $k = 10$ tokens yields the best performance (see Appendix B.3.1). This pattern is consistent with prior work (Jain & Huth, 2018; Toneva & Wehbe, 2019), which shows that increasing context improves encoding performance up to a point, after which the gains saturate.

4 Results

We validate SCRIBE along four directions. (1) *semantic validity*: we assess whether LLM-generated semantic scores align with human semantic ratings for isolated concepts, without contextual information. (2) *context awareness*: we test whether human semantic judgments vary with context and whether SCRIBE captures these shifts. (3) *neural predictivity*: we analyze how well these representations model brain activity during natural language processing, especially in semantic-relevant regions. (4) *semantic specificity*: using variance partitioning, we assess whether the brain-predictive information captured by SCRIBE is less shared with syntactic features than alternative representations, while retaining unique variance in language-related brain regions. Together, these analyses

Table 1: Alignment between SCRIBE scores and human ratings for isolated concepts (i.e., without additional context) (Binder et al., 2016). We report Pearson correlation (r) and mean squared error (MSE), with corresponding permutation-based p-values. Overall, SCRIBE aligns significantly with human semantic ratings, with the ensemble achieving the best performance.

Model	**Pearson** $\uparrow$	p_r	**MSE** $\downarrow$	p_{MSE}
Mistral 7B Instruct	0.62	< 0.001	2.16	< 0.001
Qwen 2.5 7B Instruct	0.69	< 0.001	1.32	< 0.001
LLaMA 3.1 8B Instruct	0.67	< 0.001	1.70	< 0.001
LLaMA 3.1 70B Instruct	0.74	< 0.001	1.53	< 0.001
Ensemble	**0.77**	< 0.001	**1.07**	< 0.001

provide a comprehensive evaluation of SCRIBE as a scalable, interpretable and context-aware approach to semantic representation.

Implementation details. All experiments were performed on internal compute clusters, using two NVIDIA H100 PCIe GPUs with $\sim$ 80GB GPU memory per device. The computation of the experiments took a total of $\sim$ 10 days.

4.1 SCRIBE GENERATES SEMANTIC FEATURES THAT ALIGN WITH HUMAN JUDGMENTS

We first evaluate whether SCRIBE can produce semantically meaningful representations in the absence of context, treating SCRIBE as a scalable proxy for human semantic ratings over isolated concepts. To this end, we compare SCRIBE scores for isolated concepts (i.e., without context) to human semantic feature ratings (Binder et al., 2016), using the original dataset of 535 words introduced in that work and described in Section 3.1.1, as a reference benchmark. For each concept, we quantify the similarity between the SCRIBE scores and human ratings using Pearson correlation and mean squared error (MSE). We assess statistical significance using permutation tests, obtained by randomly shuffling concept–property assignments 1000 times.

Table 1 summarizes the results. SCRIBE aligns significantly with human ratings across all evaluated models, with correlations above 0.6 and MSEs significantly below permutation baselines (see Appendix B.1.1 for further results). The ensemble shows the strongest alignment, reaching $r = 0.77$ and an MSE of 1.07, indicating that aggregation improves correspondence with human semantic judgments. Together, these results support SCRIBE as a scalable approximation of human semantic ratings for isolated concepts. We therefore use the ensemble throughout the remainder of the paper, with individual-model results reported in Appendix B.

4.2 SCRIBE CAPTURES CONTEXTUAL SHIFTS IN HUMAN SEMANTIC JUDGMENTS

We next investigate how human semantic judgments vary with linguistic context and whether SCRIBE captures these changes. Existing semantic property norms (Binder et al., 2016; Mitchell et al., 2008) characterize concepts in isolation, leaving unexplored how the same properties vary when concepts occur in different contexts. To study this systematically, we construct a controlled dataset in which context is manipulated to increase or decrease specific concept–property associations. For each concept–property pair, we generate two contextualized sentences using an instruction-tuned LLM (`llama3.1-8b`): a low-target context (L_{T}) designed to decrease the association and a high-target context (H_{T}) designed to increase it (see Appendix A.3 for details).

We first examine how these controlled contextual manipulations affect human semantic judgments. To this end, we conducted a large-scale human behavioral study, recruiting 500 participants to rate a balanced subset of concept–property pairs under the original out-of-context condition (O) and the corresponding L_{T} and H_{T} contexts (see Appendix B.2.1). To our knowledge, this provides the first human-annotated resource extending interpretable semantic feature norms from isolated concepts to context-dependent judgments. As shown in Table 2(a), human judgments change systematically with context, following the expected ordering $H_{\text{T}} > O > L_{\text{T}}$, with all directional comparisons significant ($p < .001$). These results provide direct evidence that semantic properties traditionally characterized through fixed, out-of-context concept norms are dynamically modulated by linguistic context. We will publicly release these contextual human ratings.

Table 2: Human contextual judgments and their alignment with the SCRIBE ensemble. **(a)** Mean semantic feature scores for human judgments and SCRIBE across the original (O), low-target (L_{T}), and high-target (H_{T}) conditions, with one-sided Wilcoxon signed-rank tests for the predicted directional contrasts. **(b)** Alignment between SCRIBE and human ratings in each condition, measured by Pearson correlation (r) and mean squared error (MSE), with one-sided permutation tests.

(a) Contextual modulation

Annotator	$\mathbf{L_T}$	**O**	$\mathbf{H_T}$	$\mathbf{L_T < O}$	$\mathbf{H_T > O}$	$\mathbf{H_T > L_T}$
Human	2.746	3.308	4.286	$< .001$	$< .001$	$< .001$
Ensemble	1.574	2.842	3.534	$< .001$	$< .001$	$< .001$

(b) Human–SCRIBE alignment

Condition	Pearson ↑	p_r	MSE ↓	p_{MSE}
O	0.743	.001	1.673	.001
$\mathbf{L_T}$	0.686	.001	2.878	.001
$\mathbf{H_T}$	0.625	.001	2.472	.001

We then ask whether SCRIBE captures the same context-dependent changes observed in human judgments. We observe in Table 2(a) that SCRIBE reproduces the human ordering $H_{\mathrm{T}} > O > L_{\mathrm{T}}$, with all directional comparisons significant ($p < .001$). Importantly, SCRIBE not only preserves the human ordering across contexts but also aligns with human ratings at the level of individual concept–property pairs. As shown in Table 2(b), human and SCRIBE scores are significantly correlated in all three conditions and have significantly lower MSE than chance.

Together, these results establish two complementary findings: semantic property judgments are systematically modulated by linguistic context in humans, and SCRIBE captures these context-dependent changes while maintaining substantial agreement with human judgments at the level of individual concept–property pairs.

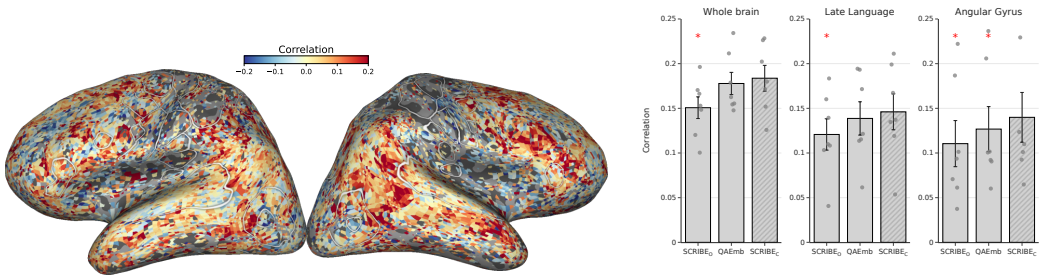
4.3 Context-aware SCRIBE scores are good predictors of brain activity

We next assess whether the contextual semantic scores produced by SCRIBE capture aspects of meaning reflected in brain activity.

Contextualization improves brain alignment of SCRIBE scores. We isolate the contribution of context to the neural predictivity of SCRIBE by comparing $\mathrm{SCRIBEO}$ (out-of-context), which estimates semantic scores from the target concept alone, with $\mathrm{SCRIBEC}$ (contextual), which additionally conditions scores on the surrounding text. All results use semantically meaningful word types and a context window of $k = 10$ tokens (see Section 3.2.1).

Figure 2a shows the spatial distribution of voxel-wise differences in brain alignment for a representative subject, with higher alignment for $\mathrm{SCRIBE_C}$ concentrated in language-related regions and more pronounced in the right hemisphere, consistent with the patterns observed in other subjects (Appendix B.3.2). We further quantify these observations in Figure 2b. Across participants, $\mathrm{SCRIBE_C}$ significantly outperforms $\mathrm{SCRIBE_O}$ both across the whole brain and within the examined language-related ROIs, demonstrating that incorporating linguistic context improves the neural predictivity of SCRIBE. These results show that incorporating context in SCRIBE leads to a better account of neural activity than static, out-of-context features.

SCRIBE predicts unique variance in semantic brain regions. We further compare the neural representations captured by $\mathrm{SCRIBE_C}$ and QAEmb (Figure 2b). While the two methods achieve comparable whole-brain alignment, $\mathrm{SCRIBE_C}$ significantly outperforms QAEmb in the AG, a region strongly implicated in semantic processing (Price et al., 2015). These differences raise the question of whether the two representations capture distinct brain-relevant information beyond their overall predictive performance. We therefore perform variance partitioning to separate the variance shared between the two representations from that uniquely explained by each method. In Figure 3b, we observe that $\mathrm{SCRIBE_C}$ explains more unique variance than QAEmb across the examined ROIs, with significant differences in the lateral temporal cortex (LTC) and the AG. The spatial distribution for a representative subject is shown in Figure 3a (see Appendix B.3.3 for all subjects). Notably, the



(a) Difference between contextual and out-of-context alignment

(b) Brain alignment scores

Figure 2: **(a)** Voxel-wise differences in brain alignment between contextual and out-of-context variants of SCRIBE ($\text{SCRIBE}_\text{C} - \text{SCRIBE}_\text{O}$) for a representative subject (subject 3). Red regions show higher predictive accuracy for the contextual model. Improvements due to contextualization are predominantly observed in language-related cortical regions. **(b)** Brain alignment for SCRIBE_O, QAEmb, and SCRIBE_C across the whole brain and language-related ROIs. Bars represent mean performance across subjects, with individual subject values overlaid and error bars indicating the standard error. Red asterisks ($*$) denote statistically significant differences (Wilcoxon signed-rank test). Contextual SCRIBE (SCRIBE_C) consistently outperforms its out-of-context (SCRIBE_O) variant, while achieving performance comparable to QAEmb.

significant differences occur in regions strongly implicated in semantic processing: in addition to the AG, the LTC has likewise been associated with semantic representation and integration (Visser et al., 2010; Pallier et al., 2011; Pylkkänen, 2020). Together, these results suggest that SCRIBE_C captures semantic-relevant neural variance that is not accounted for by QAEmb, despite their similar overall predictive performance. Additional analyses show that the representations are complementary and that SCRIBE's performance is robust after dimensionality reduction (Appendix B.3.3 and B.3.4).

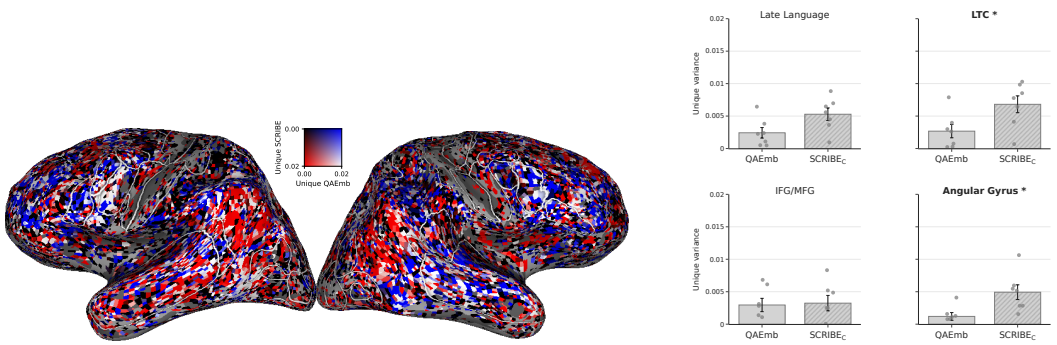
4.4 SCRIBE CAPTURES BRAIN-PREDICTIVE INFORMATION BEYOND SYNTACTIC FEATURES

Because semantic information is difficult to isolate from other linguistic information in natural language, we assess the semantic specificity of SCRIBE indirectly by asking how much of its brain-predictive variance remains after accounting for independently defined syntactic features. We thus construct a syntactic control space comprising part-of-speech tags, dependency length, and tree depth, and use variance partitioning to quantify its unique and shared contributions with SCRIBE and QAEmb (more details in Appendix B.3.5).

SCRIBE captures more unique and less syntax-related neural variance than QAEmb. After accounting for syntactic features, SCRIBE_C predicts more unique variance than QAEmb across language-related ROIs, reaching significance in late language regions overall, and specifically the LTC and AG (Figure 11b), while also sharing significantly less variance with syntax for predicting the LTC and AG (see Figure 11a). Together, these results support the greater semantic specificity of SCRIBE: its neural predictivity is less associated with syntactic information while predicting more unique variance in the brain. We extend the analysis to Llama2-7b representations (Appendix B.3.6). SCRIBE_C outperforms non-contextual Llama2-7b embeddings, while contextual Llama2-7b embeddings achieve higher brain predictivity (Figure 13). However, contextual LLM embeddings encode a broader mixture of linguistic information. Notably, SCRIBE_C shares significantly less variance with syntactic features than Llama2-7b in IFG/MFG (Figure 14), a region consistently implicated in syntactic processing and structure building (Dapretto & Bookheimer, 1999; Heim et al., 2003; Fiebach et al., 2005; Friederici, 2011).

5 DISCUSSION

Understanding how semantic representations are constructed and modulated in the brain requires models that are both *context-aware* and *interpretable*. In this work, we introduced SCRIBE, an approach that combines a cognitively grounded semantic feature space with LLM-based annotation



(a) Whole-brain spatial distribution.

(b) ROI-level aggregation.

Figure 3: Unique neural variance explained by $\mathrm{SCRIBE_C}$ and QAEmb. **(a)** Spatial distribution of voxel-wise unique variance for a representative subject (subject 2). Red indicates greater unique variance explained by $\mathrm{SCRIBE_C}$, whereas blue indicates greater unique variance explained by QAEmb. **(b)** Region-wise comparison of uniquely explained variance across semantically relevant ROIs. Bars show mean performance across subjects, with individual values overlaid and error bars indicating the standard error. $\mathrm{SCRIBE_C}$ consistently outperforms QAEmb, with significant improvements in LTC ($p = 0.039$) and AG ($p = 0.023$).

to produce contextualized, interpretable semantic representations for natural language. By using LLMs as scalable semantic annotators, SCRIBE preserves the interpretability of theory-driven feature spaces while extending them beyond isolated concepts to arbitrary words in context.

Our human study first establishes that semantic feature judgments are context-dependent: across 500 recruited participants, controlled linguistic contexts systematically shifted ratings along the targeted semantic dimensions. To our knowledge, these data constitute the first human-annotated context-dependent semantic feature norms and will be publicly released. We show that SCRIBE closely aligns with human semantic judgments both for isolated concepts and across contextual conditions.

At the neural level, incorporating context significantly improves SCRIBE's prediction of fMRI responses relative to its out-of-context representation, particularly in language-related cortical regions. Contextual SCRIBE explains more unique neural variance than the closest existing method, with significant differences in semantic regions, such as the LTC and AG. After accounting for independently defined syntactic features, SCRIBE retains more unique brain-predictive variance than QAEmb while sharing less variance with syntax in these regions, showing that its additional brain-predictive information persists after accounting for the measured syntactic features.

More broadly, our results show that interpretability and contextuality need not be treated as opposing goals. Prior feature-based models offer interpretable dimensions but rely on static out-of-context annotations, while LLM representations are contextual and scalable but difficult to attribute to specific semantic content. SCRIBE bridges this gap by representing meaning along explicit, human-interpretable dimensions that dynamically vary with context, providing a scalable framework for studying contextual semantics in the brain.

Limitations and future work. SCRIBE scores depend on the underlying LLMs and the prompting strategy. To assess sensitivity to prompt formulation, we performed ablations, which show that performance is largely robust across the tested variations (Appendix A.2.2). To reduce model-specific variability and biases, SCRIBE aggregates scores across multiple LLMs, analogous to aggregating human judgments. Additionally, the use of a fixed semantic space (Binder et al., 2016) constrains representations to a predefined set of dimensions, potentially missing aspects of meaning. However, this design choice provides a principled and interpretable basis grounded in cognitive neuroscience, enabling direct comparison with human semantic judgments and prior work.

These limitations suggest several directions for future work, including extending SCRIBE to more diverse datasets and domains, integrating it with mechanistic interpretability methods to link semantic features to model internals, and potentially expanding the semantic space to capture a broader range of semantic dimensions.

Ethics Statement

This work includes a human behavioral study conducted through Prolific. Participants provided informed consent and were compensated for their participation. The study was approved by the Ethical Review Board of at the authors' institution. No personally identifiable information was collected or released. The resulting contextual semantic ratings will be released in anonymized and aggregated form.

Reproducibility Statement

We provide detailed descriptions of SCRIBE, the semantic annotation procedure, model and prompt specifications, contextual stimulus construction, human behavioral study, brain-encoding analyses, and statistical procedures in the main text and Appendix. We will release the contextual human annotations and code necessary to reproduce the analyses and experiments presented in this work.

AI Use Statement

Generative AI tools were used in this work to generate controlled linguistic contexts for the synthetic dataset, as described in the paper, and to assist with the editing and refinement of the manuscript. All AI-assisted outputs used in the research were reviewed and verified by the authors. The authors take full responsibility for the methods, analyses, claims, and final content of the paper.

References

Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395, 2017.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36:21895–21907, 2023.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698, 2021.

Mirella Dapretto and Susan Y Bookheimer. Form and content: dissociating syntax and semantics in sentence comprehension. *Neuron*, 24(2):427–432, 1999.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019.

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.

Christian J Fiebach, Matthias Schlesewsky, Gabriele Lohmann, D Yves Von Cramon, and Angela D Friederici. Revisiting the role of broca's area in sentence processing: syntactic integration versus syntactic working memory. *Human brain mapping*, 24(2):79–91, 2005.

Abraham Jacob Fresen, Rochelle Choenni, Micha Heilbron, Willem Zuidema, and Marianne de Heer Kloots. Language models that accurately represent syntactic structure exhibit higher representational similarity to brain activity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 2011.

Matthew F Glasser, Stephen M Smith, Daniel S Marcus, Jesper LR Andersson, Edward J Auerbach, Timothy EJ Behrens, Timothy S Coalson, Michael P Harms, Mark Jenkinson, Steen Moeller, et al. The human connectome project's neuroimaging approach. *Nature neuroscience*, 19(9): 1175–1187, 2016.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

St Heim, Bertram Opitz, and Angela D Friederici. Distributed cortical networks for syntax processing: Broca's area as the common denominator. *Brain and language*, 85(3):402–408, 2003.

Paul Hoffman, Matthew A Lambon Ralph, and Timothy T Rogers. Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, 45(3):718–730, 2013.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3651–3657, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

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

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior research methods*, 52(3):1271–1291, 2020.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.

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

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. *arXiv preprint arXiv:2410.09230*, 2024.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8528, 2024.

SubbaReddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36:18001–18014, 2023.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108 (6):2522–2527, 2011.

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

Amy R Price, Michael F Bonner, Jonathan E Peelle, and Murray Grossman. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *The Journal of Neuroscience*, 35(7):3276–3284, 2015.

Liina Pylkkänen. Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375 (1791), 2020.

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

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.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Jacob Turton, David Vinson, and Robert Smith. Extrapolating binder style word embeddings to new words. In *Proceedings of the second workshop on linguistic and neurocognitive resources*, pp. 1–8, 2020.

Jacob Turton, Robert Elliott Smith, and David Vinson. Deriving contextualised semantic features from bert (and other transformer model) embeddings. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pp. 248–262, 2021.

Maya Visser, Elizabeth Jefferies, and MA Lambon Ralph. Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of cognitive neuroscience*, 22(6):1083–1094, 2010.

A METHODS

A.1 COGNITIVELY-GROUNDED, INTERPRETABLE SEMANTIC FEATURE SPACE

Semantic feature space.

To ground our approach in interpretable and cognitively motivated representations, we build on the semantic feature space introduced by Binder et al. Binder et al. (2016), which provides a cognitively and neurobiologically informed characterization of conceptual meaning. In this framework, concepts are represented as vectors over a fixed set of 65 semantic properties spanning multiple domains, including perceptual (e.g., vision, audition, somatosensation), action-related (e.g., motor effectors and interaction), spatial, temporal, and abstract properties (e.g., social, cognitive, and affective features), as summarized in Table 3. These features are grounded in converging evidence from cognitive neuroscience, which suggests that conceptual knowledge is distributed across modality-specific and heteromodal brain systems. Accordingly, the semantic dimensions were shown Binder et al. (2016) to exhibit a structured organization, clustering into groups that align with known functional divisions of the cortex, such as sensory, motor, and affective systems. This representation offers a principled and interpretable characterization of semantic content: each dimension corresponds to a well-defined aspect of meaning, enabling direct analysis of how specific properties contribute to downstream predictions.

In this work, we build on this semantic feature space, preserving its interpretability while extending it to contextual settings using LLM-based annotation.

Human semantic ratings for out-of-context concepts. The Binder et al. Binder et al. (2016) dataset consists of 535 out-of-context concepts (434 nouns, 62 verbs, and 39 adjectives). Each concept is annotated along the 65 semantic dimensions using human ratings, where participants evaluate the degree to which a concept is associated with specific properties (e.g., *something that you can easily see*, *something large in size*, *something that makes one feel happy*, etc.). Aggregating these ratings across participants yields continuous-valued feature vectors, forming an interpretable semantic representations in the shared feature space. Figure 4 shows the distribution of semantic scores obtained from the aggregated human ratings for each property.

Table 3: Overview of the 65 semantic properties from the feature space of Binder et al. Binder et al. (2016), grouped by modality and cognitive domain. Each concept is represented as a vector over these interpretable semantic features.

Name	Type	Modality	Submodality	Description
Vision	Sensory	Vision	General	Something that you can easily see
Bright	Sensory	Vision	Intensity	Visually light or bright
Dark	Sensory	Vision	Intensity	Visually dark
Color	Sensory	Vision	Surface	Has a characteristic or defining color
Pattern	Sensory	Vision	Surface	Has a characteristic visual texture or surface pattern
Large	Sensory	Vision	Size	Large in size
Small	Sensory	Vision	Size	Small in size
Motion	Sensory	Vision	Motion	Shows visually observable movement
Biomotion	Sensory	Vision	Motion	Shows movement like that of a living thing
Fast	Sensory	Vision	Motion	Shows fast visible movement
Slow	Sensory	Vision	Motion	Shows slow visible movement
Shape	Sensory	Vision	Shape	Has a characteristic visual shape or form
Complexity	Sensory	Vision	Shape	Visually complex
Face	Sensory	Vision	Shape	Has a human or human-like face
Body	Sensory	Vision	Shape	Has human or human-like body parts
Touch	Sensory	Somatic	General	Can be easily recognized by touch
Hot	Sensory	Somatic	Surface	Hot to the touch
Cold	Sensory	Somatic	Surface	Cold to the touch
Smooth	Sensory	Somatic	Surface	Has a smooth texture to the touch
Rough	Sensory	Somatic	Surface	Has a rough texture to the touch
Light	Sensory	Somatic	Proprioception	Light in weight
Heavy	Sensory	Somatic	Proprioception	Heavy in weight
Pain	Sensory	Somatic	Nociception	Associated with pain or physical discomfort
Audition	Sensory	Audition	General	Something that you can easily hear
Loud	Sensory	Audition	Intensity	Makes a loud sound
Low	Sensory	Audition	Frequency	Has a low-pitched sound
High	Sensory	Audition	Frequency	Has a high-pitched sound
Sound	Sensory	Audition	Shape	Has a characteristic or recognizable sound
Music	Sensory	Audition	Shape	Makes a musical sound
Speech	Sensory	Audition	Phoneme	Someone or something that talks
Taste	Sensory	Gustation	Quality	Has a characteristic or defining taste
Smell	Sensory	Olfaction	Quality	Has a characteristic or defining smell
Head	Motor	Motor	Motor	Associated with actions using the face, mouth, or tongue
UpperLimb	Motor	Motor	Motor	Associated with actions using the arm, hand, or fingers
LowerLimb	Motor	Motor	Motor	Associated with actions using the leg or foot
Practice	Motor	Motor	Motor	A physical object one has personal experience using
Landmark	Spatial	Navigation	Navigation	Has a fixed location, as on a map
Path	Spatial	Navigation	Navigation	Shows changes in location along a direction or path
Scene	Spatial	Scene	Scene	Brings to mind a particular setting or physical location
Near	Spatial	Self Space	Proximity	Often physically near in everyday life
Toward	Spatial	Self Space	Toward	Associated with movement toward or into oneself
Away	Spatial	Self Space	Away	Associated with movement away from or out of oneself
Number	Number	Number	Number	Associated with a specific number or amount
Time	Event	Temporal	Time	Occurs at a typical or predictable time
Duration	Event	Temporal	Duration	Has a predictable duration, whether short or long
Long	Event	Temporal	Long	Lasts for a long period of time
Short	Event	Temporal	Short	Lasts for a short period of time
Caused	Event	Causal	Causal	Caused by a clear preceding event, action, or situation
Consequential	Event	Causal	Causal	Likely to have consequences
Social	Event	Social	Social	Involves an interaction between people
Human	Cognition	Social	Theory of Mind	Has human or human-like intentions, plans, or goals
Communication	Cognition	Social	Communication	Used by people to communicate
Self	Cognition	Self	Self	Related to one's own view of oneself
Cognition	Cognition	Cognition	Cognition	A form of mental activity or function of the mind
Benefit	Evaluation	Cognitive	Positive	Could help or benefit oneself or others
Harm	Evaluation	Cognitive	Negative	Could cause harm to oneself or others
Pleasant	Evaluation	Affective	Positive	Something one finds pleasant
Unpleasant	Evaluation	Affective	Negative	Something one finds unpleasant
Happy	Emotion	Positive	High	Makes one feel happy
Sad	Emotion	Negative	Low	Makes one feel sad
Angry	Emotion	Negative	High	Makes one feel angry
Disgusted	Emotion	Negative	High	Makes one feel disgusted
Fearful	Emotion	Negative	High	Makes one feel afraid
Surprised	Emotion	Neutral	High	Makes one feel surprised
Drive	Drive	General	General	Motivates one to do something
Needs	Drive	Basic	Basic	Would be hard to live without
Attention	Attention	General	General	Grabs one's attention
Arousal	Attention	General	Arousal	Makes one feel alert, activated, excited, or keyed up

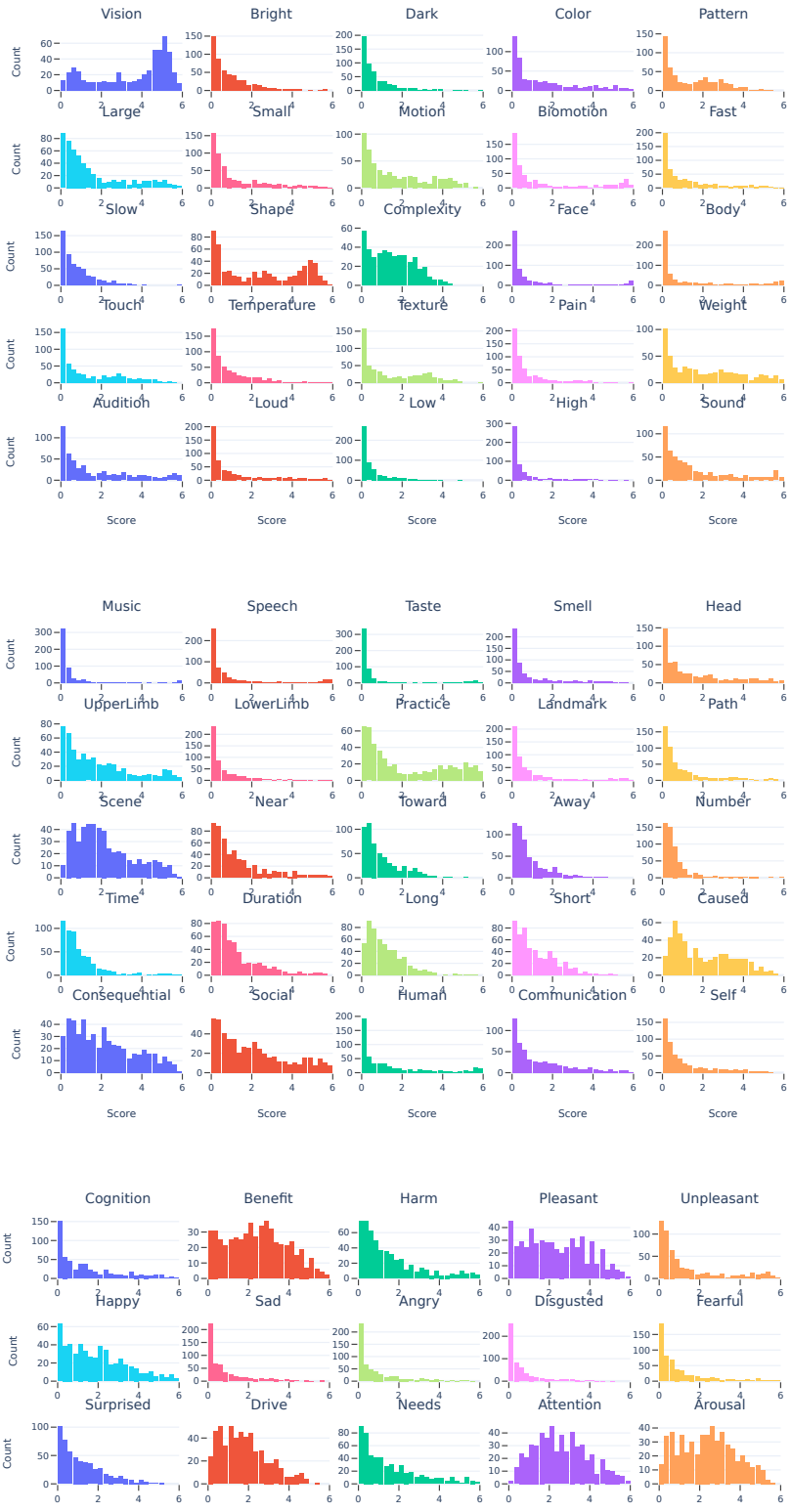


Figure 4: Distribution of semantic feature scores derived from human ratings in the Binder et al. Binder et al. (2016) dataset (composed of 535 words), shown for each of the 65 semantic properties.

A.2 LLM-BASED SEMANTIC ANNOTATIONS

SCRIBE uses LLMs to estimate semantic feature scores by prompting them to assess how strongly a target concept is associated with each semantic property. This procedure is applied both in a *out-of-context* setting, where the word is evaluated in isolation without providing any contextual sentence, and in a *contextual* setting, where the word is evaluated within a specific contextual sentence.

A.2.1 OUT-OF-CONTEXT SEMANTIC ANNOTATION

In the out-of-context setting, SCRIBE obtains semantic representations using LLM-generated annotations computed independently for each concept-property pair. For a given target concept, the model is prompted for each semantic property in the semantic feature space (described in Appendix A.1), resulting in 65 queries. Each query asks the model to rate how strongly the word is associated with the corresponding property, returning a scalar score in the range $[0, 6]$. These scores are then aggregated to form a semantic feature vector for the concept, which is directly comparable to the human ratings of Binder et al. (2016) when evaluated on the same set of concepts.

The prompt takes the following form:

```
You are an expert in rating to what degree words are
associated with semantic properties.
To what degree is the word ``{word}'' associated with
{property relation} {property explanation}?
Provide only a single number for the rating of
the word ``{word}'' on a scale from 0 (not at all
associated or not applicable) to 6 (very strongly
associated).
Do not include any explanation or text.
Output:
```

An example prompt for the concept word “bus” and property “motion” is shown below:

> *You are an expert in rating to what degree words are associated with semantic properties.*
> *To what degree is the word “bus” associated with showing a lot of visually observable movement?*
> *Provide only a single number for the rating of the word “bus” on a scale from 0 (not at all associated or not applicable) to 6 (very strongly associated).*
> *Do not include any explanation or text, just the number.*
> *Output:*

A.2.2 ABLATION OF PROMPT COMPONENTS

LLMs are known to be sensitive to prompt wording, yet our main experiments use a single prompt formulation To assess the robustness of SCRIBE to prompt design choices, we performed a prompt ablation study in which we systematically modified individual components of the prompt while keeping all other instructions unchanged.

We split out prompt into these conceptual components: (i) a role instruction assigning the model as an expert, (ii) the task description specifying the semantic property and relation to be rated, (ii) the rating scale, (iv) output instructions constraining the response format.

> **[Role]** You are an expert in rating to what degree words are associated with semantic properties. **[Task]** To what degree is the word “{word}” associated with {property relation}
>
> {property explanation}? **[Rating scale]** Provide only a single number for the rating of the
>
> word “{word}” on a scale from 0 (not at all associated or not applicable) to 6 (very strongly associated). **[Output format]** Do not include any explanation or text. Output:

We evaluated a series of prompt variants by modifying one component at a time relative to the default prompt. The tested variants are summarized in Table 4.

Table 4: Prompt variants evaluated in the prompt ablation study. Each variant modifies a single component of the default prompt while leaving all other components unchanged.

Component	**Variant**	**Modification**
Baseline	Default prompt	Prompt used throughout the main experiments.
Role	Remove expert role	Remove the instruction: "You are an expert in rating to what degree words are associated with semantic properties."
Task	Property name only	Provide only the property name instead of its definition.
Task	Property name + definition	Provide both the property name and its definition.
Task	Relation: *as*	Replace "associated with" by "as", e.g., "To what degree is the word {word} as {property}?"
Task	Relation: *characterized by*	Replace "associated with" by "characterized by", e.g., "To what degree is the word {word} characterized by {property}?"
Scale	Binary (0–1)	Replace the 0–6 rating scale with a binary scale ranging from 0 (not at all associated or not applicable) to 1 (very strongly associated).
Scale	Full scale descriptions	Keep the 0–6 scale but provide a textual description for every rating (e.g., 1 = very weakly associated, 2 = weakly associated, etc.).
Output	Semantic reasoning	Add the instruction: "Consider the meaning of the word itself when assigning the rating."

Tables 5 summarize the prompt ablation results across the individual models and their ensemble. Removing the expert role consistently led to the largest degradation in performance, indicating that role prompting is important for this task. Among the remaining ablations, the effects were less consistent across models. Variants that provided a longer rating scale, encouraged semantic reasoning, or rephrased the relation as "characterized by" occasionally achieved improvements over the baseline prompt, whereas replacing the property definition with only the property name or changing the relation to "as" generally reduced performance. However, despite these differences, the absolute improvements over the baseline were generally small, suggesting limited practical impact of these prompt modifications.

Model	Ablation	MSE	Pearson Corr.
llama3.1-8b-instruct	Base	2.232	0.627
	No role	6.532*	0.465*
	Property name only	2.790*	0.469*
	Property name + definition	1.916	0.607
	Relation “as”	3.202	0.494
	Relation “characterized by”	2.001	0.620
	Scale 0–1	2.494	0.613
	Long scale (0–6)	1.891*	0.659
	Semantic reasoning	1.927*	0.649
mistral-7b-instruct	Base	1.969	0.624
	No role	4.002*	0.516*
	Property name only	2.794	0.452
	Property name + definition	2.105	0.578
	Relation “as”	2.850	0.480
	Relation “characterized by”	1.991	0.612
	Scale 0–1	2.088	0.612
	Long scale (0–6)	2.100	0.608
	Semantic reasoning	2.136	0.602
qwen2.5-7b-instruct	Base	1.611	0.623
	No role	3.721*	0.567*
	Property name only	2.096*	0.490*
	Property name + definition	1.812	0.563
	Relation “as”	2.123	0.516
	Relation “characterized by”	1.740	0.582
	Scale 0–1	1.604	0.633
	Long scale (0–6)	1.652	0.637
	Semantic reasoning	1.599	0.641
llama3.1-70b-instruct	Base	2.023	0.703
	No role	2.519	0.644*
	Property name only	2.883	0.569*
	Property name + definition	1.973	0.693
	Relation “as”	2.155	0.667
	Relation “characterized by”	1.962	0.702
	Scale 0–1	2.289	0.671
	Long scale (0–6)	2.319	0.696
	Semantic reasoning	2.014	0.706
Ensemble	Base	1.192	0.741
	No role	2.946*	0.654*
	Property name only	1.822	0.589*
	Property name + definition	1.294	0.707
	Relation “as”	1.568	0.655
	Relation “characterized by”	1.126	0.739
	Scale 0–1	1.176	0.738
	Long scale (0–6)	1.132	0.749
	Semantic reasoning	1.175	0.745

Table 5: Prompt ablation results across individual models and their ensemble. An asterisk indicates performance significantly worse than the base prompt ($p < 0.05$).

A.2.3 CONTEXTUAL SEMANTIC ANNOTATION

In the contextual setting, SCRIBE follows the same procedure as in the out-of-context case, using LLM-generated annotations independently for each concept-property pair. The key difference is that semantic scores are now conditioned on the sentence in which the concept appears, enabling context-dependent representations.

The prompt is modified as follows:

> `You are an expert in rating to what degree words are associated with semantic properties.`
> **Given the context "{contextual_sentence}",** `to what degree is the word ``{word}'' associated with {property relation} {property explanation}?`
> `Provide a single number from 0 (not at all associated or not applicable) to 6 (very strongly associated).`
> `Do not include any explanation or text.`
> `Output:`

This formulation allows the model to adjust semantic feature values based on how the meaning of the target word is modulated by context.

An example prompt for the concept word "bus", the property "motion", and the contextual sentence of "The bus was parked at the station" is shown below:

> *You are an expert in rating to what degree words are associated with semantic properties.*
> *Given the context "The bus was parked at the station", to what degree is the word "bus" associated with showing a lot of visually observable movement?*
> *Provide only a single number for the rating of the word "bus" on a scale from 0 (not at all associated or not applicable) to 6 (very strongly associated).*
> *Do not include any explanation or text, just the number.*
> *Output:*

Examples of context window sizes. Table 6 illustrates the different context window sizes used for contextual semantic annotation. For each target word, the context consists of the k tokens preceding and including the target, with larger values of k providing progressively broader linguistic context.

Table 6: Examples of the context provided to SCRIBE for different context window sizes. The target word is shown in bold.

Example	Context size	Context
1	$k = 3$	of those **people**
	$k = 5$	I'm one of those **people**
	$k = 10$	raised with a religion and I'm one of those **people**
	$k = 50$	so a lot of people are raised with a religion and I'm one of those **people**
2	$k = 3$	the word **pretty**
	$k = 5$	extension of the word **pretty**
	$k = 10$	word beautiful was just an extension of the word **pretty**
	$k = 50$	when I was younger the word beautiful was just an extension of the word **pretty**
3	$k = 3$	date and **i**
	$k = 5$	a christian date and **i**
	$k = 10$	time i was going on a christian date and **i**
	$k = 50$	it was the first time i was going on a christian date and **i**
4	$k = 3$	learn a **lesson**
	$k = 5$	romance not learn a **lesson**
	$k = 10$	go to morocco to escape romance not learn a **lesson**
	$k = 50$	my plan was to go to morocco to escape romance not learn a **lesson**

A.3 SYNTHETIC DATASET FOR CONTROLLED CONTEXTUAL EVALUATION

To evaluate whether contextual information induces systematic changes in semantic representations, we construct a synthetic dataset designed to isolate the effect of context on individual concept–property associations. A key requirement for this evaluation is to control for the baseline strength of each association, as it constrains how much it can change under contextual manipulation. We therefore first select a subset of concept–property pairs that span the full range of baseline association strengths (Section A.3.1), and only then generate contexts that induce opposite semantic effects: one that aims to decrease the association ("low" condition) and one that aims to increase it ("high" condition) (Section A.3.2). This design enables controlled testing of whether semantic features respond predictably to targeted contextual manipulations.

A.3.1 CONCEPT AND PROPERTY SELECTION

A key challenge in evaluating contextual effects is that the baseline out-of-context strength of a concept–property association may constrain how much it can change. For example, pairs with very high initial ratings have limited room for further increase, while pairs with very low ratings may be difficult to decrease. To control for this effect, we leverage the 535 concepts from the Binder et al. Binder et al. (2016) dataset, which provides human ratings for each concept–property pair and allows us to explicitly characterize baseline out-of-context association strength.

We discretize these ratings into three levels (*low*, *mid*, and *high*) using thresholds aligned with the original rating scale:

$$\text{low} : [0, 2), \quad \text{mid} : [2, 4), \quad \text{high} : [4, 6].$$

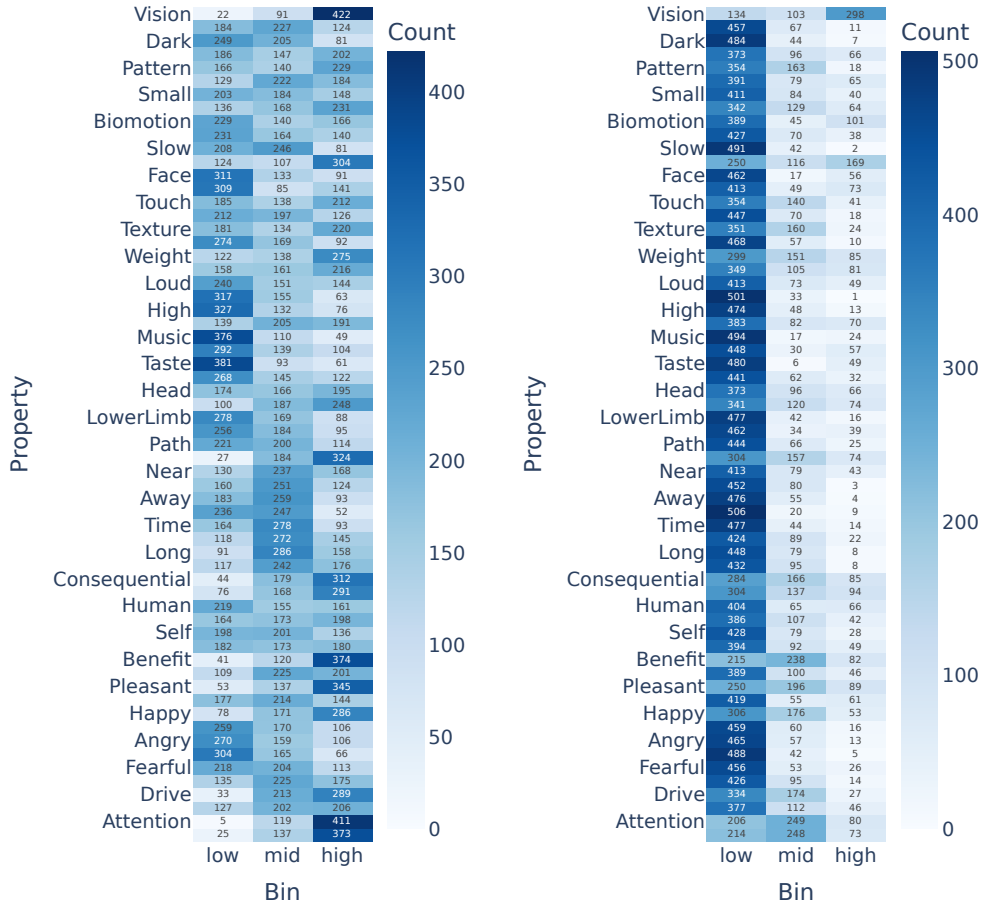
This yields a categorical matrix that assigns each concept–property pair to a baseline association bin (see Figure 5 for bin coverage).

To ensure controlled comparisons across these levels, we construct a balanced subset of concepts. Not all properties exhibit sufficient variability across concepts (Section A.1). We therefore retain only properties for which each bin contains at least 8 concepts, resulting in 57 properties after filtering, and excluding *Low, Slow, Toward, Away, Disgusted, Complexity, Taste, Dark*. For each retained property, we then select a subset of concepts that spans the full range of baseline associations using a greedy selection procedure:

1. Initialize coverage counts for each (property, bin) cell.
2. At each step, consider all remaining candidate concepts.
3. For each candidate, compute how many under-covered (property, bin) cells it would help fill.
4. Select the concept that maximizes this coverage gain.
5. Update coverage counts and repeat until all cells reach the target.

At convergence, we then select 8 concepts per property bin (low, mid, high), yielding 24 concepts per property. This balanced design ensures that contextual effects are evaluated under comparable baseline conditions, rather than being driven by saturation or floor effects.

The final set of 144 concepts is: `deceit, snub, gasp, infinity, truth, reality, advice, era, grew, evening, number, fed, terrorist, quantity, handshake, angry, jaw, street, dinner, hospital, baseball, bonfire, foot, accordion, animosity, crossed, gunshot, dropped, cough, ricochet, raspberry, egg, morning, dime, tulip, festival, apricot, sun, cloud, summer, shiny, explosion, dandelion, lightning, election, empty, ate, chime, banjo, carnival, fireworks, bus, beach, bagpipe, avalanche, toe, drew, cash, coffee, dolphin, belief, lived, protest, fiddle, leg, volcano, elm, snake, alligator, clang, girl, mouth, bicycle, sick, loud, built, apartment, ambulance, battle, college, audience, school, riot, jury, couple, thunder, pineapple, drank, finger, banana, bell, bee, cellphone, harmonica, beer, long, belch, delivered, activist, victim, man, matinee, parade, army, sum, council, lip,`



(a) Quantiles thresholds.

(b) Manual thresholds.

Figure 5: Distribution of (concept, property) pairs across discretized baseline bins under two binning strategies. **(a)** Quantile-based thresholds $(0.34, 1.5)$ yield more uniform distributions across bins; however, the threshold for the high bin is low relative to the full rating scale (0–6), leading to less interpretable groupings. **(b)** Manual thresholds aligned with the original rating scale ($[0, 2)$, $[2, 4)$, $[4, 6]$) produce more interpretable assignments of low, mid, and high association levels across properties. These observations motivate the use of fixed thresholds in our synthetic dataset construction.

`arm, walked, marched, approached, ran, hailstorm, expensive, airport, bugle, celebrated, old, tired, shouted, doctor, dead, left, hiked, friendly, insult, car, treaty, wonder, problem, funeral, injured, torment, semester, year, listened (to), screech, scream, excuse, aggressive, restaurant, worked, day, intellect.`

A.3.2 Contextual sentence generation

For each selected concept-property pair in the synthetic dataset, we then use `llama3.1-8b-instruct` to generate two sentences: one intended to decrease the target property score (low condition) and one intended to increase it (high condition). These sentences are designed to manipulate the semantic interpretation of the target concept in a controlled manner.

Prompt template to generate the contextual sentences. The prompt template used to construct the synthetic dataset is shown below:

```
You are a sentence generation assistant. Your task
is to write ONE simple and short sentence using
the target word {concept}''. The sentence should
evoke a **{level}** value for the semantic property
``{property name}'', defined as {property relation}
{property explanation}.
Constraints:
- Use the word ``{concept}'' exactly as given in the
sentence.
- Do not use the property name property name''
directly in the sentence.
- Do not include explanations or extra formatting,
only the sentence.
Output format:
Return the sentence in the following format: <ONE
sentence with ``{concept}'' >.
No additional text, keys or punctuation.
Context:
The word ``{concept}'' belongs to the super category
``{super category concept}'' and category ``{category
concept}''
Output:
```

An example to generate a contextual sentence for the concept word "bus", the property "motion", and the target level "low" is shown below:

You are a sentence generation assistant.
Your task is to write ONE simple and short sentence using the target word "bus".
The sentence should evoke a "low" value for the semantic property "Motion", defined
defined as showing a lot of visually observable movement.
Constraints:
- Use the word "bus" exactly as given in the sentence.
- Do not use the property name "Motion" as given directly in the sentence.
- Do not include explanations or extra formatting, just the sentence.
Output format:
Return the sentence in the following format: <ONE sentence with "bus" >.
No additional text, keys or punctuation.
Context:
The word "bus" belongs to the super category "artifact" and category "vehicle".
Output:

Example of contextual sentences generated. We manually inspect a subset of generated examples to ensure quality and coherence. Some examples are presented in Table 7.

Table 7: Examples of controlled contextual manipulations for selected property-concept pairs. For each pair, the low-context sentence is designed to reduce the target semantic property, while the high-context sentence increases it, demonstrating how context systematically modulates semantic feature annotations.

Property	**Concept**	**Low-context sentence**	**High-context sentence**
Angry	thunder	The thunderstorm was a beautiful display of nature.	The loud thunder made me furious.
Motion	bus	The bus was parked at the station.	The bus sped down the highway.
Social	dinner	I ate dinner alone at home.	We had dinner together with friends at the restaurant.
Pleasant	fireworks	The loud fireworks startled the baby.	Watching fireworks on the 4th of July was a joyful experience.
Fast	bus	The bus moved slowly through traffic.	The bus sped down the highway.
Motion	activist	The activist sat quietly in the corner.	The activist marched through the streets.
Numer	leg	She has a broken leg.	The dog has four legs.
Small	egg	The ostrich egg is enormous.	The tiny egg was barely visible to the naked eye.

A.4 BRAIN ENCODING SETUP

A.4.1 fMRI DATASET

We use a subset of the fMRI dataset originally described and made publicly available in previous work (LeBel et al., 2023). The subset used consists of 11 stories, 8 participants, and all participants listened to and read all the stories. The speech stimuli consisted of 10 to 15 minute stories taken from The *Moth Radio Hour* podcast and used previously (Moussa et al., 2024; Antonello et al., 2023; Oota et al., 2024). The selected stories cover a wide range of topics and are highly engaging (Oota et al., 2024). The model validation dataset consisted of one 10 min story. In this work, we include 7 of the 8 subjects, excluding subject 4 due to incomplete data (a missing story). To enable comparison with QAEmb (Singh et al., 2025), an interpretability representation used as baseline, we further exclude one story ("myfirstdaywiththeyankees") that is not available in its feature set, resulting in a total of 9 stories used for training the encoding models.

A.4.2 ROIs DETAILS

The human cerebral cortex multi-modal parcellation (Glasser atlas) defines 180 regions of interest (ROIs) per hemisphere (Glasser et al., 2016). Following prior work (Moussa et al., 2024; Oota et al., 2024), we define language regions based on a subset of these ROIs, as listed in Table 8.

Table 8: Brain regions and corresponding ROI labels.

Region	**Labels**
Angular gyrus (AG)	PFm, PGs, PGi, TPOJ2, TPOJ3
Lateral temporal cortex (LTC)	STSda, STSva, STGa, TE1a, TE2a, TGv, TGd, A5, STSdp, STSvp, PSL, STV, TPOJ1
Inferior frontal gyrus (IFG)	44, 45, IFJa, IFSp
Middle frontal gyrus (MFG)	55b
Primary auditory cortex (A1)	A1
Early auditory regions	A1, PBelt, MBelt, LBelt, RI, A4

A.4.3 ESTIMATED NOISE CEILING

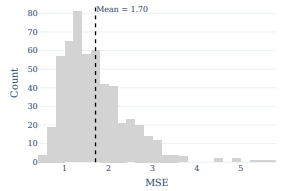
Noise in fMRI data can impair evaluations, so it is important to estimate the "noise ceiling" of each voxel in the fMRI recordings. The voxel-wise noise ceiling is estimated for all participants based on the preferred method by the fMRI dataset paper (LeBel et al., 2023), which leverages within-participant repetitions of the same story to compute the maximum explainable variance for each voxel. This noise ceiling value estimates the amount of explainable variance in the brain signal, ranging from 0 to 1. We use this estimated noise ceiling to filter noisy voxels during brain alignment evaluation. We use a filtration threshold of 0.4, in line with the findings of Antonello et al. (2023). After filtering voxels with low noise ceiling, there remain $30,000$ to $50,000$ voxels per participant.

B RESULTS

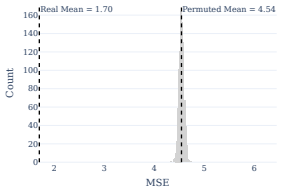
B.1 SCRIBE GENERATES SEMANTIC FEATURES THAT ALIGN WITH HUMAN JUDGMENTS

B.1.1 PERMUTATION TEST

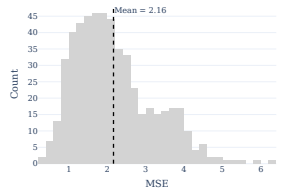
To assess whether the observed alignment between LLM-generated annotations and human semantic ratings reflects meaningful structure, we run a permutation test. Specifically, for each model, we randomly shuffle concept–property assignments and recompute evaluation metrics (MSE and Pearson correlation), thereby destroying the true semantic correspondence while preserving the marginal distributions of annotations. We repeat this procedure multiple times $(1,000)$ to obtain empirical null distributions for each metric. Figures 6 and 7 show the distributions of scores obtained from the true annotations against those from the permutation baseline. Across all models, the real distributions are clearly separated from the permuted ones, with lower MSE and higher correlation values, showing that the annotations obtained from the LLM models capture systematic relationships between concepts and semantic properties.



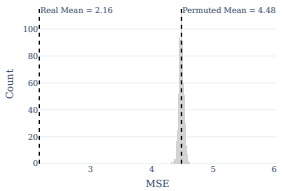
(a) MSE real (llama3.1-8b).



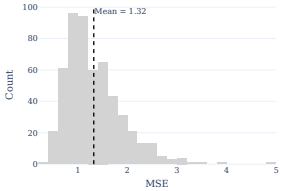
(b) MSE perm. (llama3.1-8b).



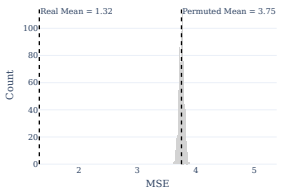
(c) MSE real (mistral-7b).



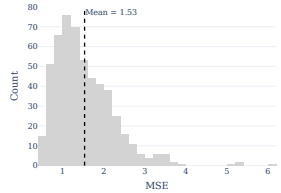
(d) MSE perm. (mistral-7b).



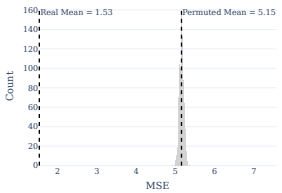
(e) MSE real (qwen2.5-7b).



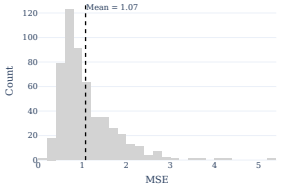
(f) MSE perm. (qwen2.5-7b).



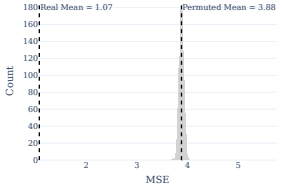
(g) MSE real (llama3.1-70b).



(h) MSE perm. (llama3.1-70b).

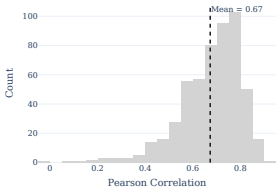


(i) MSE real (ensemble).

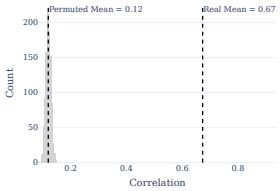


(j) MSE perm. (ensemble).

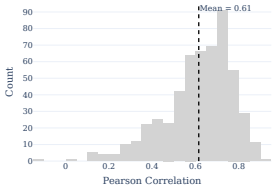
Figure 6: Distribution of mean squared error (MSE) for out-of-context semantic annotation across models. For each model, we show the distribution of the MSE values computed on the true concept–property assignments (left) and under a permutation baseline obtained by randomly shuffling concept–property pairs (right). Across all models, the true MSE distribution is substantially lower than the permutation baseline, indicating that LLM-based annotations capture meaningful semantic structure beyond chance. Results are shown for individual models and the ensemble.



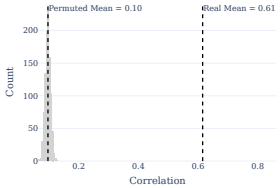
(a) Pearson real (llama3.1-8b).



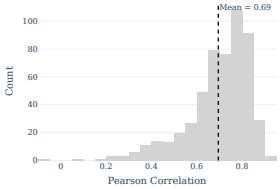
(b) Pearon perm. (llama3.1-8b).



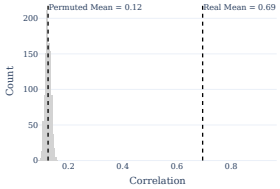
(c) Pearson real (mistral-7b).



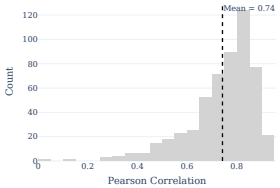
(d) Pearon perm. (mistral-7b).



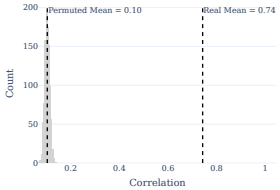
(e) Pearson real (qwen2.5-7b).



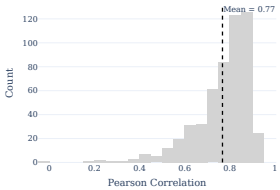
(f) Pearon perm. (qwen2.5-7b).



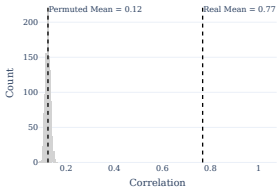
(g) Pearson real (llama3.1-70b).



(h) Pearon perm. (llama3.1-70b).



(i) Pearson real (ensemble).



(j) Pearon perm. (ensemble).

Figure 7: Distribution of Pearson correlation between LLM-generated semantic annotations and human ratings across models. For each model, we show the distribution of correlation values computed on the true concept–property assignments (left) and under a permutation baseline obtained by randomly shuffling concept–property pairs (right). Across all models, the true correlation distribution is substantially higher than the permutation baseline, indicating that LLM-based annotations capture meaningful semantic structure beyond chance. Results are shown for individual models and the ensemble.

B.2 SCRIBE GENERATES CONTEXTUALLY-AWARE SEMANTIC FEATURES

B.2.1 LARGE-SCALE HUMAN BEHAVIORAL STUDY

Semantic feature norms provide interpretable representations of conceptual meaning along dimensions such as sensory, motor, affective, social, and cognitive properties. However, existing human-annotated semantic spaces typically characterize concepts in isolation, assigning each concept a fixed set of property values regardless of the linguistic context in which it occurs. SCRIBE extends the neurocognitively grounded semantic feature space of Binder et al. (2016) to contextual settings by using LLMs to estimate semantic properties of concepts conditioned on their linguistic context. In its controlled evaluation, SCRIBE used LLM-generated contexts designed to systematically increase or decrease individual concept–property associations and showed corresponding changes in its semantic annotations (see Appendix A.3 for more information). To determine whether these contextual manipulations produce analogous changes in human semantic judgments, we conducted a human annotation study using the same controlled stimuli.

We selected 54 of the 57 properties included in the original synthetic dataset, excluding Temperature, Weight, and Texture. These three properties are derived by taking the maximum rating across pairs of opposing dimensions (Hot/Cold, Heavy/Light, and Rough/Smooth, respectively). Extending this procedure to contextual annotations would require both constituent dimensions to be evaluated within the same sentence, which is difficult to guarantee when generating contexts that independently target each property. For each of the 54 retained properties, we selected six concepts based on their original out-of-context ratings: two with low, two with mid-range, and two with high concept–property association strengths according to the original out-of-context human ratings from Binder et al. (2016). This stratification ensured coverage across the range of baseline concept–property association strengths. This resulted in a total of 324 concept–property pairs (54×6).

Each concept–property pair was evaluated under three conditions: a high-target context ($\mathrm{H_T}$) designed to increase the association with the target property, a low-target context ($\mathrm{L_T}$) designed to decrease it, and an out-of-context condition (O) in which the concept was presented without a sentence. This resulted in $324 \times 3 = 972$ unique annotation items. Participants rated the strength of each concept–property association on a $0 - 6$ scale, with contextual judgments made specifically with respect to the meaning of the concept in the provided sentence.

We recruited participants through Prolific, requiring English fluency, a $98 - 100\%$ approval rate, and at least 100 previous submissions. Each participant rated 14 semantic annotation items plus an attention check, and was exposed to only one condition for any given concept–property pair, preventing direct comparison between contextual variants. Condition assignment was balanced across participants. Participants who failed the attention check were excluded during recruitment, which continued until 500 participants had passed. As an additional quality-control criterion, we excluded participants whose mean response time was below 7 seconds per semantic annotation item. This removed 15 participants (3%), yielding a final sample of 485 participants. After quality control, every annotation item received multiple independent judgments, with an average of 6.99 ratings per item ($SD = 0.99$, median $= 7$, range $= 3$–14).

The resulting human judgments showed a clear and systematic effect of linguistic context. After averaging across annotators for each concept–property–condition combination, ratings were highest in the high-target condition ($M = 4.28$, $SD = 1.18$), intermediate in the out-of-context condition ($M = 3.32$, $SD = 1.54$), and lowest in the low-target condition ($M = 2.75$, $SD = 1.35$). One-sided Mann–Whitney U tests showed significant differences for all three directional comparisons ($H_\mathrm{T} > O, O > L_\mathrm{T}$, and $H_\mathrm{T} > L_\mathrm{T}$; all $p < .001$). The effect was also broadly distributed across the semantic feature space rather than driven by a small subset of properties: H_T ratings exceeded O ratings for all 54 properties and L_T ratings for 53 of 54 properties, while the complete expected ordering $H_\mathrm{T} > \mathrm{O} > L_\mathrm{T}$ was observed for 43 of the 54 properties.

Together, these data provide direct human validation that interpretable semantic property judgments can be systematically modulated by linguistic context and support the central assumption underlying SCRIBE's context-sensitive semantic annotations. They also provide a human-annotated resource for evaluating scalable approaches to contextual semantic annotation. Importantly, obtaining approximately seven independent judgments for only 972 controlled annotation items already required nearly 500 participants, *illustrating the practical difficulty of extending*

Table 9: Contextual modulation of semantic feature scores for the human-evaluated subset. We report mean scores for human judgments and the SCRIBE ensemble across the original (O), low-target (L_{T}), and high-target (H_{T}) conditions, together with two- and one-sided Wilcoxon signed-rank tests. Results are shown across all concept–property pairs ($\mathrm{All}_{\mathrm{O}}$) and stratified by original out-of-context association strength (L_{O}, M_{O}, H_{O}).

Annotator	**Concept**	**Average scores**			**Wilcoxon p-value (two-sided)**			**Wilcoxon p-value (one-sided)**		
	Bin	$\mathbf{L_T}$	**O**	$\mathbf{H_T}$	$\mathbf{L_T}$ vs O	$\mathbf{H_T}$ vs O	$\mathbf{H_T}$ vs $\mathbf{L_T}$	$\mathbf{L_T} < \mathbf{O}$	$\mathbf{H_T} > \mathbf{O}$	$\mathbf{H_T} > \mathbf{L_T}$
Human	$\mathrm{All}_{\mathrm{O}}$	2.746	3.308	4.286	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	L_{O}	3.367	4.161	4.822	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	M_{O}	2.439	3.035	3.981	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	H_{O}	2.851	3.499	4.454	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Ensemble	$\mathrm{All}_{\mathrm{O}}$	1.574	2.842	3.534	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	L_{O}	0.567	0.759	2.193	.003	$< .001$	$< .001$	.002	$< .001$	$< .001$
	M_{O}	1.420	2.900	3.606	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	H_{O}	2.736	4.868	4.802	$< .001$	.469	$< .001$	$< .001$	.234	$< .001$

Table 10: Comparison of low-target (L_T), original/no-context (O), and high-target (H_T) scores assigned by human annotators, individual LLMs, and LLM ensembles to the evaluated subset of the synthetic dataset. All annotators exhibit the intended ordering $L_T < O < H_T$, with significant differences in all three expected directions ($L_T < O$, $H_T > O$, and $H_T > L_T$; all $p < .001$, one-sided Wilcoxon signed-rank tests). Two-sided tests are also reported for completeness. The ensemble excluding Llama3.1-8B provides a control excluding the model used to generate the synthetic sentences.

Model	**Average scores**			**Wilcoxon p-value (two-sided)**			**Wilcoxon p-value (one-sided)**		
	$\mathbf{L_T}$	**O**	$\mathbf{H_T}$	$\mathbf{L_T}$ vs O	$\mathbf{H_T}$ vs O	$\mathbf{H_T}$ vs $\mathbf{L_T}$	$\mathbf{L_T} < \mathbf{O}$	$\mathbf{H_T} > \mathbf{O}$	$\mathbf{H_T} > \mathbf{L_T}$
Human	2.746	3.308	4.286	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Llama3.1-70B	1.704	2.840	4.080	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Llama3.1-8B	1.775	2.840	3.207	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Mistral-7B	1.386	2.840	3.494	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Qwen2.5-7B	1.433	2.840	3.354	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Ensemble	1.574	2.842	3.534	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Ensemble w/o Llama3.1-8B	1.508	2.842	3.643	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$

human annotation to arbitrary concepts and naturalistic contexts at scale. The collected data therefore serve both as a human reference for studying how context-dependent semantic representations relate to computational and neural representations, as well as a validation of SCRIBE's contextual manipulation.

B.2.2 RESULTS ON THE FULL SYNTHETIC DATASET

Table 11: Per-model results for the controlled contextual manipulation experiment (see Section 4.2 and Appendix A.3) using all the contextual generated sentences. We report average semantic scores and Wilcoxon signed-rank tests for each individual model and their ensemble. Across models, semantic feature values consistently shift in the intended direction ($H_{\mathrm{T}} \gg O \gg L_{\mathrm{T}}$), showing that contextual manipulations induce consistent and meaningful changes of SCRIBE scores. Notably, the most powerful model, `llama3.1-70b-instruct`, shows statistically significant effects across all original bins.

Model	**Concept**	**Average scores**			**Wilcoxon p-value (two-sided)**			**Wilcoxon p-value (one-sided)**		
	Bin	$\mathbf{L_T}$	**O**	$\mathbf{H_T}$	$\mathbf{L_T}$ **vs O**	$\mathbf{H_T}$ **vs O**	$\mathbf{H_T}$ **vs** $\mathbf{L_T}$	$\mathbf{L_T < O}$	$\mathbf{H_T > O}$	$\mathbf{H_T > L_T}$
Llama3.1-8B	$\mathrm{All_O}$	1.86	2.80	3.04	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{L_O}$	0.67	0.63	1.39	0.3885	< 0.001	< 0.001	0.8057	< 0.001	< 0.001
	$\mathrm{M_O}$	1.86	2.89	3.08	< 0.001	0.0436	< 0.001	< 0.001	0.0218	< 0.001
	$\mathrm{H_O}$	3.06	4.88	4.64	< 0.001	0.0048	< 0.001	< 0.001	0.9976	< 0.001
Mistral-7B	$\mathrm{All_O}$	1.37	2.80	3.11	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{L_O}$	0.35	0.63	1.59	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{M_O}$	1.23	2.89	3.30	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{H_O}$	2.52	4.88	4.43	< 0.001	< 0.001	< 0.001	< 0.001	1.0000	< 0.001
Qwen2.5-7B	$\mathrm{All_O}$	1.57	2.80	3.04	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{L_O}$	0.60	0.63	1.85	0.5730	< 0.001	< 0.001	0.2865	< 0.001	< 0.001
	$\mathrm{M_O}$	1.47	2.89	3.16	< 0.001	0.0004	< 0.001	< 0.001	0.0002	< 0.001
	$\mathrm{H_O}$	2.63	4.88	4.12	< 0.001	< 0.001	< 0.001	< 0.001	1.0000	< 0.001
Llama3.1-70B	$\mathrm{All_O}$	1.81	2.80	3.68	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{L_O}$	0.33	0.63	2.04	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{M_O}$	1.81	2.89	3.95	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{H_O}$	3.29	4.88	5.05	< 0.001	0.0236	< 0.001	< 0.001	0.0118	< 0.001
Ensemble	$\mathrm{All_O}$	1.65	2.80	3.22	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{L_O}$	0.49	0.63	1.72	< 0.001	< 0.001	< 0.001	0.0001	< 0.001	< 0.001
	$\mathrm{M_O}$	1.59	2.89	3.37	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	$\mathrm{H_O}$	2.87	4.88	4.56	< 0.001	< 0.001	< 0.001	< 0.001	1.0000	< 0.001

B.3 SCRIBE PREDICTS BRAIN ACTIVITY

B.3.1 ANNOTATION DESIGN: WORD TYPES AND CONTEXT SIZE

In this section, we report correlation values for different word type annotations and context sizes, evaluated separately for each model on voxels with a noise ceiling greater than 0.4 (see Appendix A.4.3 for details on noise ceiling estimation). Tables 12, 13, 14, and 15 present the results for the instruct-based `llama3.1-8B`, `mistral-7b`, `qwen2.5-7b`, and `llama3.1-70b` models, respectively.

Across most models, we observe a consistent pattern: performance significantly improves with contextual information, with a context window of $k = 10$ yielding the best results. This context size aligns with prior work Jain & Huth (2018); Toneva & Wehbe (2019), which shows that increasing context improves encoding performance up to a point, after which the gains saturate. Moreover, using the full set of semantic POS tags (*All Semantic POS*) leads to the highest overall correlations.

Table 12: Brain encoding correlations across context sizes using `llama3.1-8b-instruct`. We compare out-of-context annotations, fixed-size context windows ($k \in \{3, 5, 10, 50\}$), and an *up to concept* condition including all preceding stimulus context. Results are reported for three POS subsets: *Nouns*, *Nouns, Verbs, Adjectives*, and *All Semantic POS*. Values correspond to noise-ceiling-normalized correlations. Bold indicates the best context within each POS group, and underlining indicates the overall best configuration. $*$ denotes significant improvements over the out-of-context baseline (one-sided Wilcoxon signed-rank test, $p < 0.05$).

Word type	Context Size	S1	S2	S3	S5	S6	S7	S8	Avg $\pm$ STE
Nouns	out-of-context	0.1372	0.1122	0.1190	0.1066	0.0933	0.0896	0.0931	0.1073 $\pm$ 0.0065
	window/k3	0.1420	0.1237	0.1301	0.1217	0.0789	0.0881	0.0807	0.1093 $\pm$ 0.0098
	window/k5	0.1442	0.1202	0.1330	0.1129	0.1051	0.0873	0.0884	0.1130 $\pm$ 0.0081*
	window/k10	0.1621	0.1362	0.1384	0.1200	0.1101	0.0869	0.0858	0.1199 $\pm$ 0.0106*
	window/k50	0.1760	0.1485	0.1207	0.1381	0.1116	0.0831	0.0928	**0.1244 $\pm$ 0.0123**
	up to concept	0.1566	0.1191	0.1001	0.1048	0.0732	0.0777	0.0681	0.1000 $\pm$ 0.0118
Nouns, Verbs, Adj	out-of-context	0.1854	0.1479	0.1379	0.1534	0.1392	0.1222	0.1129	0.1427 $\pm$ 0.0089
	window/k3	0.1841	0.1713	0.1446	0.1574	0.1500	0.1258	0.1062	0.1485 $\pm$ 0.0100
	window/k5	0.1985	0.1741	0.1483	0.1666	0.1704	0.1293	0.1124	0.1571 $\pm$ 0.0110*
	window/k10	0.2030	0.1795	0.1463	0.1654	0.1710	0.1189	0.1032	0.1553 $\pm$ 0.0132
	window/k50	0.2228	0.1822	0.1330	0.1669	0.1854	0.1236	0.1051	**0.1599 $\pm$ 0.0156**
	up to concept	0.2137	0.1749	0.1105	0.1465	0.1516	0.1186	0.1020	0.1454 $\pm$ 0.0150
All Semantic POS	out-of-context	0.1869	0.1516	0.1515	0.1665	0.1142	0.1251	0.1099	0.1437 $\pm$ 0.0108
	window/k3	0.1972	0.1800	0.1561	0.1723	0.1292	0.1359	0.1085	0.1542 $\pm$ 0.0118*
	window/k5	0.2042	0.1909	0.1519	0.1723	0.1458	0.1346	0.1154	0.1593 $\pm$ 0.0119*
	window/k10	0.2107	0.1996	0.1546	0.1849	0.1543	0.1300	0.1036	**0.1625 $\pm$ 0.0147***
	window/k50	0.2215	0.1881	0.1483	0.1634	0.1666	0.1181	0.0912	0.1567 $\pm$ 0.0163
	up to concept	0.2153	0.1633	0.1259	0.1546	0.1072	0.1146	0.0957	0.1395 $\pm$ 0.0157

Table 13: Brain encoding correlations across context sizes using `mistral-7b-instruct`. We compare out-of-context annotations, fixed-size context windows ($k \in \{3, 5, 10, 50\}$), and an *up to concept* condition including all preceding stimulus context. Results are reported for three POS subsets: *Nouns*, *Nouns, Verbs, Adjectives*, and *All Semantic POS*. Values correspond to noise-ceiling-normalized correlations. Bold indicates the best context within each POS group, and underlining indicates the overall best configuration. $*$ denotes significant improvements over the out-of-context baseline (one-sided Wilcoxon signed-rank test, $p < 0.05$).

Word type	Context Size	S1	S2	S3	S5	S6	S7	S8	Avg $\pm$ STE
Nouns	out-of-context	0.1368	0.0917	0.1012	0.0848	0.0821	0.0871	0.0693	0.0933 $\pm$ 0.0081
	window/k3	0.1449	0.1248	0.1000	0.0938	0.0445	0.0939	0.0721	0.0963 $\pm$ 0.0124
	window/k5	0.1506	0.1337	0.1156	0.0975	0.0794	0.1048	0.0898	0.1102 $\pm$ 0.0095*
	window/k10	0.1729	0.1521	0.1340	0.1300	0.0897	0.1103	0.0944	**0.1262 $\pm$ 0.0115***
	window/k50	0.1622	0.1441	0.1248	0.1206	0.0871	0.0975	0.0880	0.1177 $\pm$ 0.0109*
	up to concept	0.1318	0.1171	0.1016	0.1051	0.0790	0.0832	0.0720	0.0985 $\pm$ 0.0082
Nouns, Verbs, Adj	out-of-context	0.1828	0.1414	0.1138	0.1405	0.1124	0.1306	0.0918	0.1305 $\pm$ 0.0110
	window/k3	0.2036	0.1859	0.1055	0.1539	0.1180	0.1350	0.0984	0.1429 $\pm$ 0.0152
	window/k5	0.2041	0.1829	0.1279	0.1650	0.1195	0.1476	0.1154	0.1518 $\pm$ 0.0128*
	window/k10	0.2184	0.2023	0.1299	0.1836	0.1452	0.1511	0.1309	**0.1659 $\pm$ 0.0134***
	window/k50	0.2136	0.1757	0.1294	0.1661	0.1409	0.1532	0.1032	0.1546 $\pm$ 0.0134*
	up to concept	0.1809	0.1781	0.0948	0.1475	0.1312	0.1417	0.0982	0.1389 $\pm$ 0.0130
All Semantic POS	out-of-context	0.1892	0.1623	0.1463	0.1600	0.0923	0.1382	0.0958	0.1406 $\pm$ 0.0134
	window/k3	0.2104	0.1972	0.1322	0.1787	0.1184	0.1289	0.1034	0.1527 $\pm$ 0.0159
	window/k5	0.2141	0.2016	0.1506	0.1766	0.1117	0.1438	0.1171	0.1594 $\pm$ 0.0150*
	window/k10	0.2240	0.2159	0.1638	0.1897	0.1463	0.1493	0.1321	**0.1744 $\pm$ 0.0136***
	window/k50	0.2175	0.1879	0.1636	0.1685	0.1421	0.1562	0.1102	0.1637 $\pm$ 0.0128*
	up to concept	0.1792	0.1810	0.1141	0.1565	0.0933	0.1314	0.0865	0.1346 $\pm$ 0.0147

Table 14: Brain encoding correlations across context sizes using `qwen2.5-7b-instruct`. We compare out-of-context annotations, fixed-size context windows ($k \in \{3, 5, 10, 50\}$), and an *up to concept* condition including all preceding stimulus context. Results are reported for three POS subsets: *Nouns*, *Nouns, Verbs, Adjectives*, and *All Semantic POS*. Values correspond to noise-ceiling-normalized correlations. Bold indicates the best context within each POS group, and underlining indicates the overall best configuration. $*$ denotes significant improvements over the out-of-context baseline (one-sided Wilcoxon signed-rank test, $p < 0.05$).

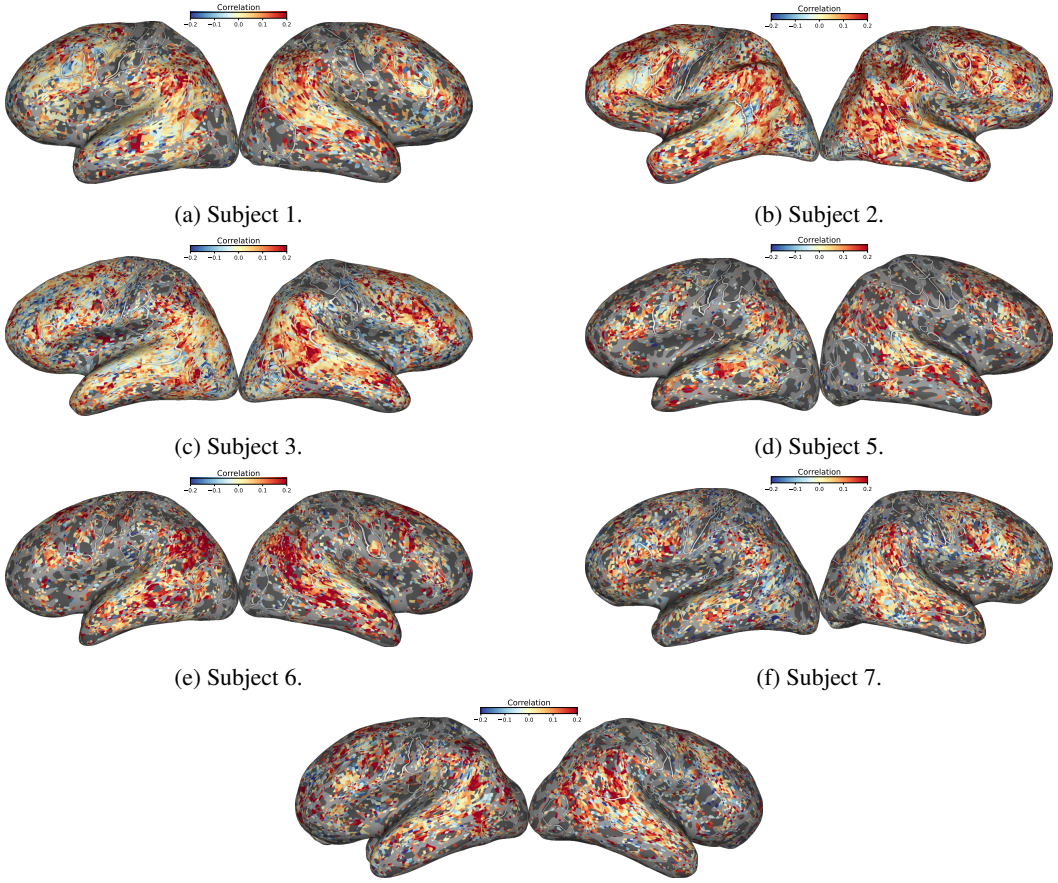
Word type	Context Size	S1	S2	S3	S5	S6	S7	S8	Avg $\pm$ STE
Nouns	out-of-context	0.1420	0.1083	0.1209	0.1181	0.0919	0.0883	0.0811	0.1072 $\pm$ 0.0082
	window/k3	0.1537	0.1455	0.1360	0.1236	0.0880	0.0857	0.0906	0.1176 $\pm$ 0.0110*
	window/k5	0.1618	0.1482	0.1321	0.1304	0.1061	0.0927	0.0917	0.1233 $\pm$ 0.0103*
	window/k10	0.1809	0.1581	0.1449	0.1396	0.1097	0.0958	0.1036	**0.1332 $\pm$ 0.0119***
	window/k50	0.1823	0.1363	0.1358	0.1358	0.0987	0.0878	0.0883	0.1236 $\pm$ 0.0130*
	up to concept	0.1606	0.1162	0.1415	0.1211	0.0917	0.0856	0.0876	0.1149 $\pm$ 0.0109*
Nouns, Verbs, Adj	out-of-context	0.1842	0.1583	0.1170	0.1543	0.1338	0.1363	0.0960	0.1400 $\pm$ 0.0109
	window/k3	0.2016	0.1888	0.1362	0.1716	0.1535	0.1292	0.1142	0.1564 $\pm$ 0.0122*
	window/k5	0.2084	0.1880	0.1335	0.1839	0.1647	0.1357	0.1166	0.1616 $\pm$ 0.0128*
	window/k10	0.2180	0.2032	0.1433	0.1922	0.1806	0.1404	0.1324	**0.1729 $\pm$ 0.0129***
	window/k50	0.2256	0.1821	0.1362	0.1891	0.1696	0.1284	0.1158	0.1638 $\pm$ 0.0148*
	up to concept	0.2045	0.1595	0.1434	0.1582	0.1318	0.1412	0.1117	0.1500 $\pm$ 0.0110*
All Semantic POS	out-of-context	0.1934	0.1655	0.1324	0.1619	0.1199	0.1583	0.0963	0.1468 $\pm$ 0.0123
	window/k3	0.2073	0.2022	0.1585	0.1821	0.1585	0.1420	0.1078	0.1655 $\pm$ 0.0132*
	window/k5	0.2164	0.2025	0.1593	0.1995	0.1660	0.1432	0.1196	0.1724 $\pm$ 0.0133*
	window/k10	0.2242	0.2227	0.1701	0.1967	0.1740	0.1516	0.1318	**0.1816 $\pm$ 0.0132***
	window/k50	0.2342	0.1888	0.1423	0.1930	0.1759	0.1383	0.1249	0.1711 $\pm$ 0.0145*
	up to concept	0.1964	0.1721	0.1416	0.1634	0.1028	0.1363	0.1015	0.1449 $\pm$ 0.0133

Table 15: Brain encoding correlations across context sizes using `llama3.1-70b-instruct`. We compare out-of-context annotations, fixed-size context windows ($k \in \{3, 5, 10, 50\}$), and an *up to concept* condition including all preceding stimulus context. Results are reported for three POS subsets: *Nouns*, *Nouns, Verbs, Adjectives*, and *All Semantic POS*. Values correspond to noise-ceiling-normalized correlations. Bold indicates the best context within each POS group, and underlining indicates the overall best configuration. $*$ denotes significant improvements over the out-of-context baseline (one-sided Wilcoxon signed-rank test, $p < 0.05$).

Word type	Context Size	S1	S2	S3	S5	S6	S7	S8	Avg $\pm$ STE
Nouns	out-of-context	0.1347	0.1239	0.1135	0.1104	0.0942	0.0943	0.0872	0.1083 $\pm$ 0.0066
	window/k3	0.1454	0.1430	0.1136	0.1143	0.0832	0.0971	0.0838	0.1115 $\pm$ 0.0097
	window/k5	0.1502	0.1382	0.1241	0.1247	0.0791	0.1026	0.0926	0.1160 $\pm$ 0.0096
	window/k10	0.1711	0.1519	0.1295	0.1300	0.0927	0.1094	0.0961	**0.1258 $\pm$ 0.0109***
	window/k50	0.1789	0.1372	0.1069	0.1370	0.0686	0.0948	0.0947	0.1169 $\pm$ 0.0139
	up to concept	0.1509	0.1383	0.1080	0.1200	0.0524	0.0905	0.1028	0.1090 $\pm$ 0.0123
Nouns, Verbs, Adj	out-of-context	0.1710	0.1649	0.1275	0.1379	0.1420	0.1194	0.1020	0.1378 $\pm$ 0.0092
	window/k3	0.1952	0.2025	0.1409	0.1590	0.1495	0.1286	0.1015	0.1539 $\pm$ 0.0135*
	window/k5	0.2048	0.2045	0.1474	0.1772	0.1600	0.1381	0.1202	0.1646 $\pm$ 0.0123*
	window/k10	0.2174	0.2152	0.1570	0.1885	0.1821	0.1418	0.1357	**0.1768 $\pm$ 0.0125***
	window/k50	0.2196	0.1881	0.1415	0.1802	0.1455	0.1329	0.1184	0.1609 $\pm$ 0.0136*
	up to concept	0.2083	0.1806	0.1177	0.1542	0.1434	0.1364	0.1449	0.1551 $\pm$ 0.0114*
All Semantic POS	out-of-context	0.1783	0.1628	0.1482	0.1525	0.1179	0.1306	0.0874	0.1397 $\pm$ 0.0115
	window/k3	0.1992	0.1972	0.1569	0.1686	0.1422	0.1413	0.0950	0.1572 $\pm$ 0.0137*
	window/k5	0.2005	0.2197	0.1693	0.1880	0.1642	0.1432	0.1171	0.1693 $\pm$ 0.0143*
	window/k10	0.2182	0.2181	0.1731	0.1993	0.1758	0.1522	0.1235	**0.1800 $\pm$ 0.0132***
	window/k50	0.2282	0.2017	0.1696	0.1747	0.1459	0.1381	0.1237	0.1688 $\pm$ 0.0139*
	up to concept	0.2141	0.1887	0.1324	0.1669	0.1197	0.1386	0.1402	0.1572 $\pm$ 0.0129

B.3.2 VOXEL-WISE DIFFERENCE BRAIN MAPS OF OUT-OF-CONTEXT AND CONTEXTUAL SCRIBE

To visualize the spatial effect of contextualization, we computed voxel-wise differences in brain-prediction accuracy between SCRIBE_C and SCRIBE_O. Figure 8 shows these difference maps for each subject, with positive values indicating greater prediction accuracy for the contextual representation.



(a) Subject 1.

(b) Subject 2.

(c) Subject 3.

(d) Subject 5.

(e) Subject 6.

(f) Subject 7.

(g) Subject 8.

Figure 8: Voxel-wise differences in brain-prediction accuracy between contextual and out-of-context SCRIBE representations ($\mathrm{SCRIBE}_C - \mathrm{SCRIBE}_O$) across subjects. Red regions indicate greater prediction accuracy for SCRIBE_C, whereas blue regions indicate greater accuracy for SCRIBE_O. Contextualization-related improvements are predominantly observed in language-related cortical regions, with stronger effects in the right hemisphere.

B.3.3 COMPLEMENTARITY BETWEEN QAEMB AND SCRIBE

Although $\mathrm{SCRIBE}_{\mathrm{C}}$ and QAEmb achieve comparable brain-encoding performance individually, an important question is whether the two representations capture redundant semantic information or provide complementary information about neural responses. We investigated this in two ways: first, by quantifying the unique variance explained by each representation through variance partitioning; and second, by testing whether combining their feature spaces improves held-out brain prediction.

Variance partitioning showed that both $\mathrm{SCRIBE}_{\mathrm{C}}$ and QAEmb explain unique variance in neural responses beyond their shared contribution (Figure 9). Despite their similar overall encoding performance, $\mathrm{SCRIBE}_{\mathrm{C}}$ explained more unique variance than QAEmb across several higher-level language and semantic ROIs, with significant differences in the lateral temporal cortex (LTC; $p = 0.039$) and angular gyrus (AG; $p = 0.023$), and a numerical difference in the late-language network. These results indicate that the two representations are not fully redundant and that $\mathrm{SCRIBE}_{\mathrm{C}}$ captures brain-predictive information not accounted for by QAEmb. To examine the spatial distribution of these unique contributions, Figure 10 shows the voxel-wise unique-variance contrast between SCRIBE and QAEmb for each subject.

To further test whether this complementary information translates into improved brain prediction, we trained an additional encoding model using the concatenation of the $\mathrm{SCRIBE}_{\mathrm{C}}$ and QAEmb feature spaces and compared its prediction performance with models using either representation alone (Table 16). At the whole-brain level, combining QAEmb with $\mathrm{SCRIBE}_{\mathrm{C}}$ significantly improved prediction performance over both $\mathrm{SCRIBE}_{\mathrm{C}}$ and QAEmb individually. The combined representation achieved a mean noise-ceiling-normalized correlation of 0.2022 ± 0.0146, compared with 0.1837 ± 0.0144 for $\mathrm{SCRIBE}_{\mathrm{C}}$ and 0.1778 ± 0.0125 for QAEmb (one-sided Wilcoxon signed-rank tests, $p = 0.0078$ for both comparisons). This corresponds to a 13.7% improvement in mean correlation over QAEmb alone.

The same pattern was observed within cortical regions associated with higher-level language and semantic processing. The combined representation yielded higher mean prediction performance than either individual representation in the late-language network (0.1577 ± 0.0213), LTC (0.1644 ± 0.0236), IFG/MFG (0.1458 ± 0.0196), and AG (0.1497 ± 0.0290). The combined representation significantly outperformed both individual representations in all four ROIs (one-sided Wilcoxon signed-rank tests, $p < 0.05$; Table 16).

Together, the variance-partitioning and joint-model analyses provide converging evidence that $\mathrm{SCRIBE}_{\mathrm{C}}$ and QAEmb capture complementary brain-relevant information. Both representations account for unique variance in neural responses, while $\mathrm{SCRIBE}_{\mathrm{C}}$ shows greater unique contributions in several higher-level language and semantic regions. Moreover, combining the two representations consistently improves held-out prediction performance over either representation alone, indicating that the information captured by SCRIBE is not subsumed by QAEmb.

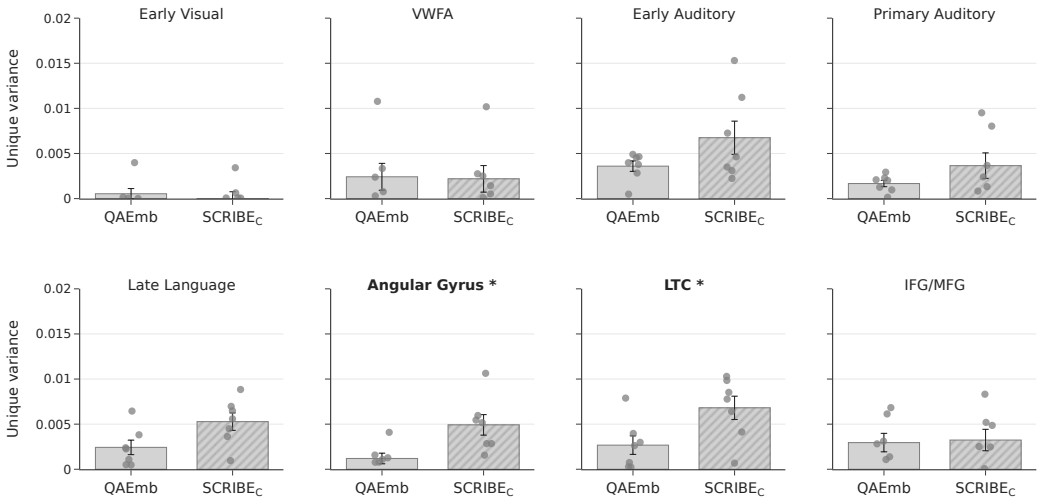
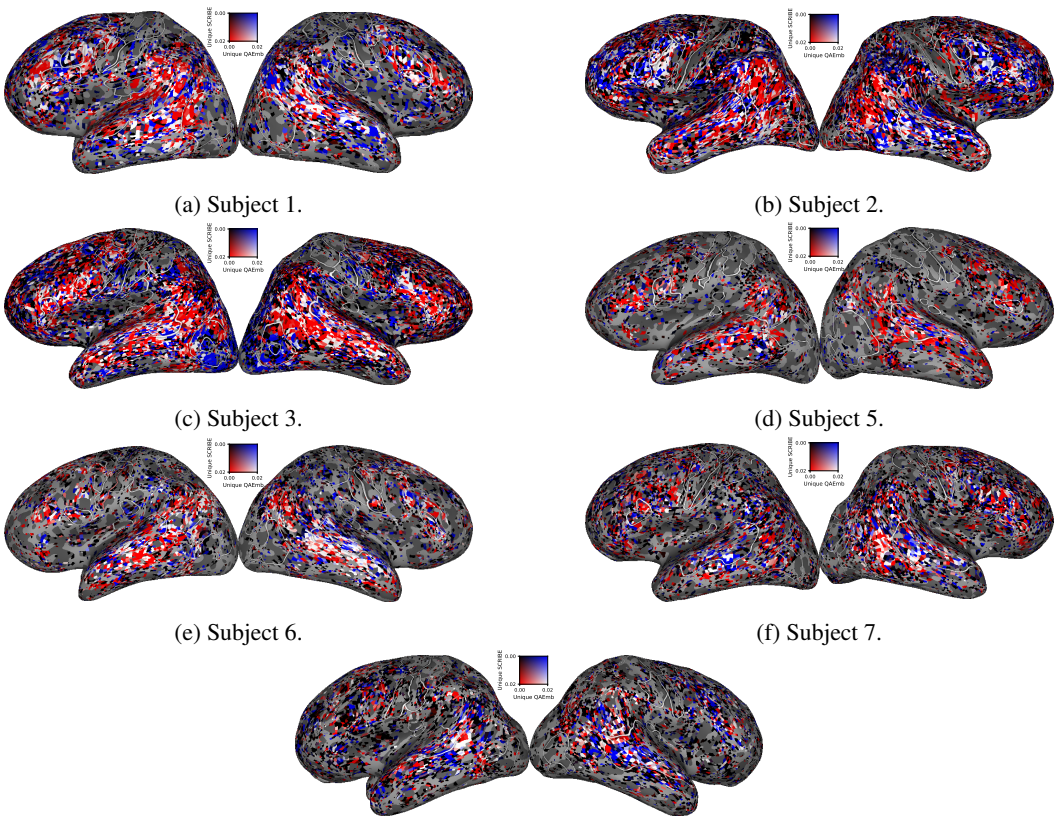


Figure 9: Unique variance explained by SCRIBE_C and QAEmb across cortical regions. Bars show the mean unique explained variance across subjects, with error bars indicating the standard error of the mean; individual dots represent subjects. p-values correspond to one-sided paired Wilcoxon signed-rank tests testing whether the unique variance explained by SCRIBE_C is greater than that explained by QAEmb. Asterisks indicate significant differences ($p < 0.05$).

Table 16: Noise-ceiling-normalized brain encoding correlations for SCRIBE (SCRIBE_C), QAEmb, and their combined representation, evaluated across the whole brain and semantic language ROIs. Values are reported for individual subjects and as mean $\pm$ standard error across subjects. $*$ denotes that QAEmb + SCRIBE_C significantly outperforms both individual representations (one-sided Wilcoxon signed-rank tests, $p < 0.05$).

Region	Method	S1	S2	S3	S5	S6	S7	S8	Avg $\pm$ STE
Whole Brain	SCRIBE_C	0.2284	0.2263	0.1714	0.2023	0.1800	0.1519	0.1258	0.1837 $\pm$ 0.0144
	QAEmb	0.2343	0.2115	0.1626	0.1788	0.1554	0.1544	0.1477	0.1778 $\pm$ 0.0125
	QAEmb + SCRIBE_C	0.2579	0.2426	0.1895	0.2140	0.1926	0.1684	0.1504	**0.2022 $\pm$ 0.0146***
Late Language	SCRIBE_C	0.1673	0.1990	0.2112	0.1348	0.1371	0.1191	0.0536	0.1460 $\pm$ 0.0201
	QAEmb	0.1715	0.1943	0.1932	0.1136	0.1214	0.1153	0.0615	0.1387 $\pm$ 0.0186
	QAEmb + SCRIBE_C	0.1880	0.2145	0.2253	0.1395	0.1450	0.1285	0.0632	**0.1577 $\pm$ 0.0213***
LTC	SCRIBE_C	0.1991	0.2046	0.1770	0.1609	0.1819	0.1155	0.0377	0.1538 $\pm$ 0.0223
	QAEmb	0.2026	0.1963	0.1569	0.1405	0.1571	0.1147	0.0415	0.1442 $\pm$ 0.0206
	QAEmb + SCRIBE_C	0.2218	0.2174	0.1870	0.1665	0.1884	0.1258	0.0435	**0.1644 $\pm$ 0.0236***
IFG/MFG	SCRIBE_C	0.1819	0.1424	0.1948	0.1200	0.0769	0.1350	0.0673	0.1312 $\pm$ 0.0182
	QAEmb	0.1901	0.1651	0.1853	0.0989	0.0726	0.1251	0.0948	0.1331 $\pm$ 0.0178
	QAEmb + SCRIBE_C	0.2060	0.1690	0.2104	0.1227	0.0837	0.1422	0.0870	**0.1458 $\pm$ 0.0196***
Angular Gyrus	SCRIBE_C	0.0926	0.2294	0.2583	0.1236	0.1099	0.1011	0.0648	0.1399 $\pm$ 0.0279
	QAEmb	0.0922	0.2058	0.2365	0.1006	0.0905	0.1016	0.0602	0.1268 $\pm$ 0.0251
	QAEmb + SCRIBE_C	0.1063	0.2378	0.2771	0.1296	0.1156	0.1118	0.0698	**0.1497 $\pm$ 0.0290***



(a) Subject 1.

(b) Subject 2.

(c) Subject 3.

(d) Subject 5.

(e) Subject 6.

(f) Subject 7.

(g) Subject 8.

Figure 10: Voxel-wise unique variance explained by SCRIBE and QAEmb across subjects. Variance partitioning was performed directly between the two representations, without including the syntactic control features. Red regions indicate greater unique variance explained by SCRIBE, whereas blue regions indicate greater unique variance explained by QAEmb. Only voxels with a noise ceiling above 0.4 are shown. Across subjects, the maps predominantly show greater unique variance for SCRIBE, particularly in language-related cortical areas, although this pattern is less pronounced in Subject 8.

B.3.4 CONTROLLING FOR FEATURE-SPACE DIMENSIONALITY

$\mathrm{SCRIBE_C}$ and QAEmb differ in the dimensionality of their final feature spaces: $\mathrm{SCRIBE_C}$ contains 65 semantic properties, whereas QAEmb uses 35 features selected from an initial set of 606 questions. We therefore investigated whether differences in brain-encoding performance could be explained simply by the larger dimensionality of $\mathrm{SCRIBE_C}$.

To control for this possibility, we applied sparse feature selection to our contextual variant of $\mathrm{SCRIBE_C}$, following the feature-selection approach used by QAEmb. Feature selection retained 33 of the original 65 semantic dimensions, resulting in a representation with approximately half the dimensionality of the original $\mathrm{SCRIBE_C}$ space and a dimensionality comparable to QAEmb. The selected dimensions were: *Bright, Dark, Large, Small, Fast, Slow, Face, Temperature, Pain, Weight, Audition, High, Music, Taste, Smell, Head, UpperLimb, LowerLimb, Landmark, Toward, Number, Time, Long, Self, Cognition, Happy, Sad, Angry, Disgusted, Fearful, Surprised, Needs*, and *Arousal*.

Despite reducing the representation from 65 to 33 dimensions, whole-brain encoding performance was preserved (Table 17). The full $\mathrm{SCRIBE_C}$ representation achieved a mean noise-ceiling-normalized correlation of 0.1837 ± 0.0144, compared with 0.1859 ± 0.0126 after feature selection. The difference was not statistically significant (two-sided Wilcoxon signed-rank test, $p = 0.5781$).

This preservation of performance was also observed across the examined language and semantic ROIs. The full and feature-selected representations achieved highly similar correlations in the late-language network (0.1460 ± 0.0201 vs. 0.1452 ± 0.0197), lateral temporal cortex (LTC; 0.1538 ± 0.0223 vs. 0.1522 ± 0.0216), inferior/middle frontal gyrus (IFG/MFG; 0.1312 ± 0.0182 vs. 0.1344 ± 0.0181), and angular gyrus (AG; 0.1399 ± 0.0279 vs. 0.1380 ± 0.0268; Table 17).

Together, these results indicate that $\mathrm{SCRIBE_C}$ does not depend on its larger number of semantic dimensions to achieve its brain-encoding performance. Reducing the representation from 65 to 33 dimensions preserves both whole-brain prediction accuracy and performance across the examined semantic-language ROIs, suggesting that the observed results are not simply attributable to differences in feature-space dimensionality between $\mathrm{SCRIBE_C}$ and QAEmb.

Table 17: Noise-ceiling-normalized brain encoding correlations for $\mathrm{SCRIBE_C}$ before and after feature selection, evaluated across the whole brain and semantic language ROIs. Feature selection reduces the representation from 65 to 33 semantic dimensions. Values are reported for individual subjects and as mean ± standard error across subjects. At the whole-brain level, performance does not significantly differ after feature selection (two-sided Wilcoxon signed-rank test).

Region	Method	S1	S2	S3	S5	S6	S7	S8	Avg ± STE
Whole Brain	$\mathrm{SCRIBE_C}$ (65)	0.2284	0.2263	0.1714	0.2023	0.1800	0.1519	0.1258	0.1837 ± 0.0144
	$\mathrm{SCRIBE_C}$ (33)	0.2275	0.2101	0.1848	0.2046	0.1897	0.1504	0.1340	0.1859 ± 0.0126
Late Language	$\mathrm{SCRIBE_C}$ (65)	0.1673	0.1990	0.2112	0.1348	0.1371	0.1191	0.0536	0.1460 ± 0.0201
	$\mathrm{SCRIBE_C}$ (33)	0.1660	0.1897	0.2124	0.1351	0.1396	0.1178	0.0561	0.1452 ± 0.0197
LTC	$\mathrm{SCRIBE_C}$ (65)	0.1991	0.2046	0.1770	0.1609	0.1819	0.1155	0.0377	0.1538 ± 0.0223
	$\mathrm{SCRIBE_C}$ (33)	0.1938	0.1974	0.1762	0.1615	0.1814	0.1146	0.0402	0.1522 ± 0.0216
IFG/MFG	$\mathrm{SCRIBE_C}$ (65)	0.1819	0.1424	0.1948	0.1200	0.0769	0.1350	0.0673	0.1312 ± 0.0182
	$\mathrm{SCRIBE_C}$ (33)	0.1835	0.1396	0.2056	0.1215	0.0876	0.1356	0.0673	0.1344 ± 0.0181
Angular Gyrus	$\mathrm{SCRIBE_C}$ (65)	0.0926	0.2294	0.2583	0.1236	0.1099	0.1011	0.0648	0.1399 ± 0.0279
	$\mathrm{SCRIBE_C}$ (33)	0.0954	0.2148	0.2567	0.1224	0.1095	0.0981	0.0690	0.1380 ± 0.0268

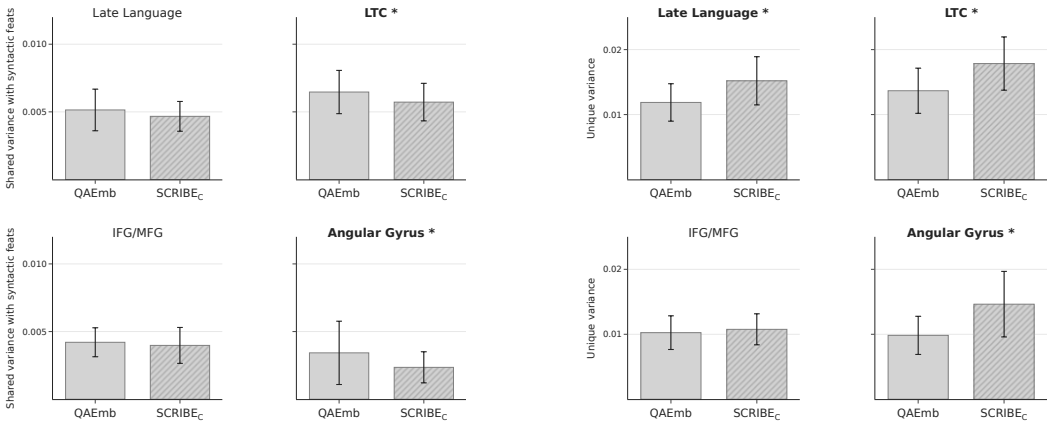
B.3.5 VARIANCE PARTITIONING WITH SYNTACTIC FEATURES

To further characterize the information underlying the brain-encoding performance of SCRIBE, we performed variance partitioning with an explicit syntactic feature space. Syntactic features comprised one-hot encoded part-of-speech tags and dependency parses obtained with spaCy, together with constituency tree depth obtained using the Benepar parser. We compared $\mathrm{SCRIBE_C}$ with QAEmb to assess both how much of their brain-predictive variance overlaps with these syntactic features and how much remains uniquely explained after accounting for them.

Figure 11a shows the variance that $\mathrm{SCRIBE_C}$ and QAEmb share with the syntactic feature space. $\mathrm{SCRIBE_C}$ shares significantly less variance with syntactic features than QAEmb in LTC and AG, with the same numerical direction in Late Language and IFG/MFG. Thus, the greater brain-prediction accuracy of $\mathrm{SCRIBE_C}$ relative to QAEmb is not accompanied by greater overlap with the measured syntactic information.

Complementing this result, Figure 11b shows the variance uniquely explained by each representation after accounting for the syntactic feature space. $\mathrm{SCRIBE_C}$ retains significantly more unique brain-predictive variance than QAEmb in Late Language, LTC, and AG, with no significant difference in IFG/MFG.

Together, these analyses show that, relative to QAEmb, $\mathrm{SCRIBE_C}$ retains more unique brain-predictive information while exhibiting less overlap with the measured syntactic features, supporting a more semantically selective interpretation of its neural predictivity.



(a) Shared variance with syntactic features.

(b) Unique variance beyond syntactic features.

Figure 11: Variance partitioning of SCRIBE_C and QAEmb with syntactic features across language-related ROIs. (a) Brain-predictive variance shared with the syntactic feature space. (b) Unique brain-predictive variance retained by each representation after accounting for the syntactic feature space. Bars show means across subjects and error bars indicate the standard error of the mean. Bold ROI labels and asterisks indicate significant paired differences between SCRIBE_C and QAEmb (one-sided paired Wilcoxon signed-rank test, $p < 0.05$).

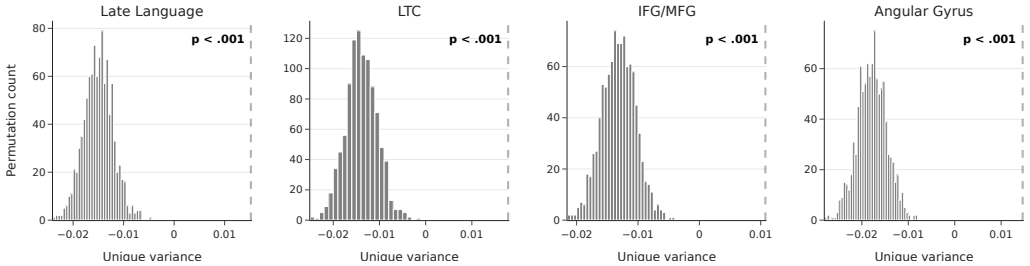


Figure 12: Permutation-based significance testing of the unique neural variance explained by $\mathrm{SCRIBE_C}$ after accounting for syntactic features. Histograms show permutation-based null distributions, and dashed lines indicate the observed unique variance.

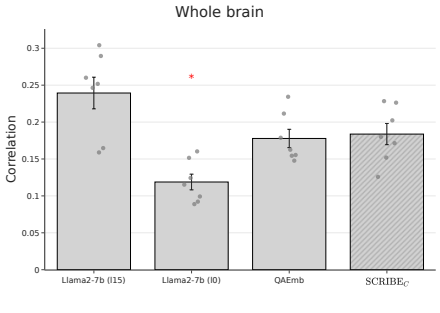
B.3.6 Comparison with LLM Embeddings

To compare SCRIBE with representations extracted directly from an LLM, we evaluated brain-encoding performance using Llama2-7B embeddings. We consider both non-contextual embeddings extracted from layer 0 and contextual embeddings extracted from layer 15.

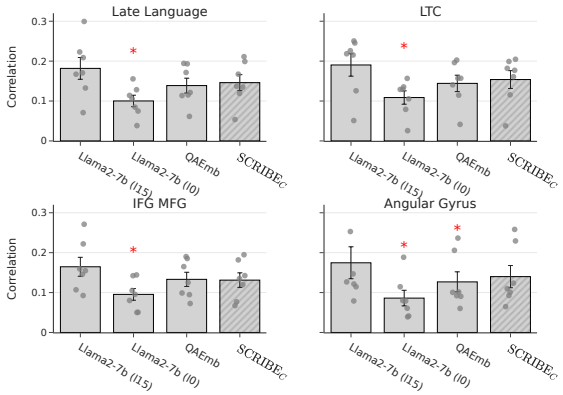
Figure 13 shows the whole-brain and region-wise comparison between the contextual Llama2-7B embeddings, QAEmb, and SCRIBE_C. Compared with non-contextual Llama2-7B embeddings, SCRIBE_C achieved substantially higher whole-brain prediction accuracy (0.1837 ± 0.0138 vs. 0.1185 ± 0.0106), corresponding to a 55% relative improvement. SCRIBE_C also significantly outperformed the non-contextual embeddings in the Late Language ROIs, AG, LTC, and IFG/MFG. In contrast, contextual Llama2-7B embeddings achieved higher overall brain-encoding performance than SCRIBE_C (0.2387 ± 0.0215 vs. 0.1837 ± 0.0138).

The higher predictive performance of contextual Llama2-7B embeddings does not, however, establish which linguistic information drives this advantage. Because contextual LLM embeddings may encode multiple sources of linguistic information, we used variance partitioning with explicit syntactic features to further characterize the information underlying their brain predictivity. Figure 14 compares the variance that contextual Llama2-7B and SCRIBE share with the syntactic feature space. SCRIBE exhibits significantly less shared variance than Llama2-7B in IFG/MFG, with the same numerical direction in Late Language and LTC. This difference is particularly notable given the established involvement of left inferior frontal regions in syntactic processing (Dapretto & Bookheimer, 1999; Heim et al., 2003; Friederici, 2011).

Finally, we tested whether the variance uniquely explained by SCRIBE exceeds chance. Permutation tests confirmed significant unique variance across all four language-related ROIs (Late Language, LTC, IFG/MFG, and AG; all $p \leq .002$; Figure 15). Thus, despite the higher overall brain-prediction accuracy of contextual Llama2-7B embeddings, SCRIBE captures significant unique brain-predictive information and shows reduced overlap with syntactic features in IFG/MFG.

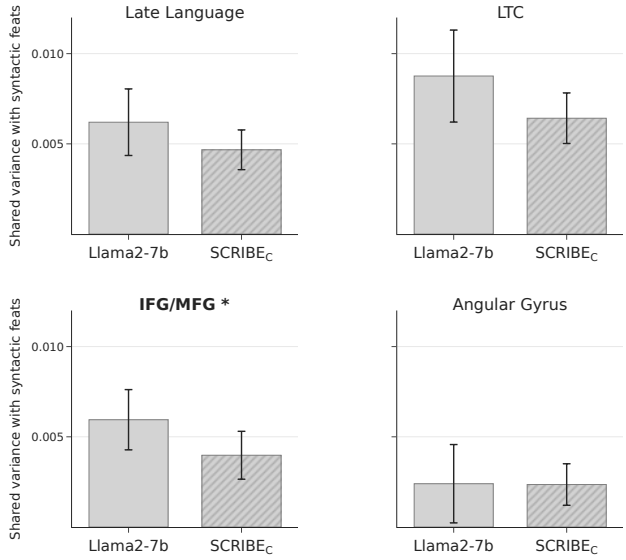


(a) Whole-brain encoding performance.



(b) Encoding performance across language-related ROIs.

Figure 13: Brain-encoding performance of Llama2-7b, QAEmb, and SCRIBE_C. **(a)** Whole-brain encoding performance. **(b)** Encoding performance across the Late Language, lateral temporal cortex (LTC), IFG/MFG, and Angular Gyrus ROIs. Bars show mean noise-ceiling-normalized correlation across subjects, individual points represent subjects, and error bars indicate the standard error of the mean. Asterisks indicate significant differences between SCRIBE_C and the corresponding baseline ($p < 0.05$; one-sided Wilcoxon signed-rank test).



(a) Shared variance with syntax.

Figure 14: Shared brain-predictive variance with syntactic features for contextual Llama2-7B and SCRIBE ($\mathrm{SCRIBE_C}$). Bars show the mean shared variance across subjects and error bars indicate the standard error of the mean. SCRIBE shares significantly less variance with the syntactic feature space than Llama2-7B in IFG/MFG (one-sided paired Wilcoxon signed-rank test, $p < 0.05$), indicated by the bold ROI label and asterisk.

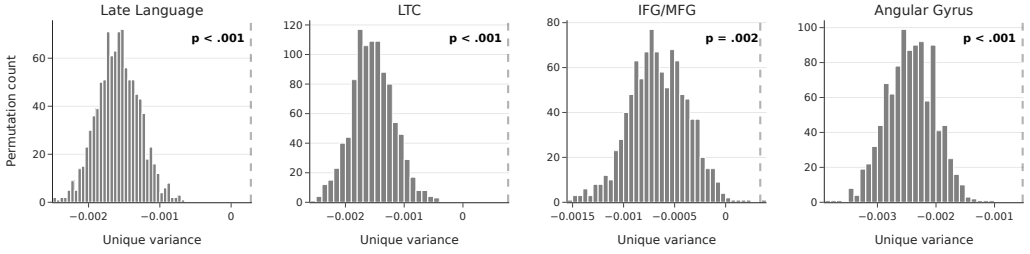


Figure 15: Permutation-based significance testing of the unique neural variance explained by $\mathrm{SCRIBE_C}$ in variance partitioning with Llama2-7b. Histograms show permutation-based null distributions, and dashed lines indicate the observed unique variance.