MODELING THE LANGUAGE CORTEX WITH FORM-INDEPENDENT AND ENRICHED REPRESENTATIONS OF SENTENCE MEANING REVEALS REMARKABLE SEMANTIC ABSTRACTNESS

Shreya Saha[1]*, **Shurui Li**[2], **Greta Tuckute**[3], **Yuanning Li**[2], **Ru-Yuan Zhang**[4], **Leila Wehbe**[5], **Evelina Fedorenko**[6] **Meenakshi Khosla**[1]

[1]University of California San Diego, [2]ShanghaiTech University, [3]Kempner Institute, Harvard University, [4]Shanghai Jiao Tong University, [5]Carnegie Mellon University, [6]Massachusetts Institute of Technology

ABSTRACT

The human language system represents both linguistic forms and meanings, but the abstractness of the meaning representations remains debated. Here, we searched for abstract representations of meaning in the language cortex by modeling neural responses to sentences using representations from vision and language models. When we generate images corresponding to sentences and extract vision model embeddings, we find that aggregating across multiple generated images yields increasingly accurate predictions of language cortex responses—sometimes rivaling large language models. Similarly, averaging embeddings across multiple paraphrases of a sentence improves prediction accuracy compared to any single paraphrase. Enriching paraphrases with contextual details that may be implicit (e.g., augmenting "I had a pancake" to include details like "maple syrup") further increases prediction accuracy, even surpassing predictions based on the embedding of the original sentence, suggesting that the language system maintains richer and broader semantic representations than language models. Together, these results demonstrate the existence of highly abstract, form-independent meaning representations within the language cortex.

1 INTRODUCTION

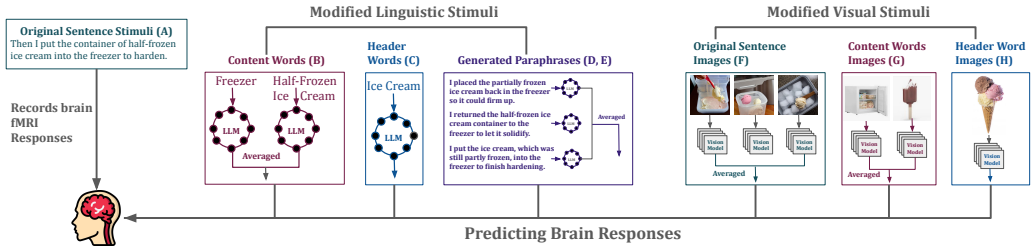

Figure 1: We probe whether neural activity in the language cortex can be explained by different representations of the original linguistic input. Starting from each original sentence, we derive and analyze eight alternative representations - (A) original sentence, (B) content words capturing the main semantic elements, (C) a header phrase summarizing a group of related sentences, (D) paraphrases of the original sentence, (E) paraphrase of the original sentence enriched with commonsense context, (F) images generated from the original sentence, (G) images generated from the content words, and (H) images generated from the header phrase. Language-based variants (A–E) are embedded with large language models, while visual variants (F–H) are embedded using vision models.

*Correspondence: ssaha@ucsd.edu

In recent years, Large Language Models (LLMs) have shown remarkable potential in modeling neural activity in the human language cortex. A central approach for studying these brain–model correspondences has been encoding models, which predict neural responses from features of linguistic stimuli. Early studies relied on corpus derived, automated and hand-constructed feature spaces including word embeddings, phonemes, syntactic structure, or narrative properties distinctions to construct encoding models of the language cortex Mitchell et al. (2008); Wehbe et al. (2014a); Huth et al. (2016); De Heer et al. (2017). These works demonstrated that word-level representations could account for meaningful variance in brain responses, but often ignored the broader sentence context in which words are embedded. Subsequent efforts showed that incorporating context, capturing relationships between words across time leads to improved modeling of language cortex activity Wehbe et al. (2014b); Jain & Huth (2018). Building on these findings, recent studies have turned to large artificial neural network–based language models, demonstrating that modern LLMs optimized for next-word prediction (e.g., GPT-style models) or trained with contextual objectives (e.g., masked language models such as BERT) provide SOTA predictions of neural responses and reveal striking convergence between artificial and biological language systems Toneva & Wehbe (2019); Schwartz et al. (2019); Schrimpf et al. (2020); Goldstein et al. (2022); Toneva et al. (2022); Hosseini et al. (2024); Tuckute et al. (2024a).

Next, researchers have sought to understand which components of language drive the similarity between model representations and brain responses. A recent study demonstrated that semantic meaning, rather than surface-level lexical or syntactic features, is the dominant factor underlying brain–model alignment Kauf et al. (2024). Research all along has also consistently upheld the importance of preserving the sentence form for activating the language cortex: original sentences reliably yield higher responses than word-level manipulations (e.g., word lists or paraphrases), highlighting the role of meaningful compositional structure. However, to our knowledge, there has been little systematic investigation of whether alternative representations (alternating either in form or even modality) that preserve semantic content can similarly predict language cortex activity.

In this paper, we examine whether the information encoded in the human language cortex can be modeled from diverse representational sources that differ substantially from the original linguistic stimulus in both form and modality. We also explore the role of commonsense knowledge, information that is implicit to humans but not explicitly present in the sentence in shaping cortical representations. Our key contributions are as follows:

1. We demonstrate that embeddings from vision foundation models, applied to visual depictions of sentences, possess non-trivial predictive power for modeling language cortex activity during sentence comprehension. Importantly, this predictivity increases when incorporating multiple diverse images per sentence, demonstrating that representational averaging provides a closer approximation of the format of linguistic meaning in the brain.
2. Next, we demonstrate that foundational large language model embeddings of paraphrased sentences also predict language cortex activity, with accuracy improving as more paraphrases are used. This mirrors the visual domain findings: greater representational diversity enhances prediction accuracy through averaging, indicating that the language cortex can be modeled even when stimulus form differs substantially from the original input.
3. Finally, we show that enriching sentences with commonsense context-information evident to humans but not explicitly present in the original text produces a substantial boost in predictivity. This finding underscores the critical role of implicit background knowledge in shaping cortical representations and suggests that effective brain-model alignment requires integrating structured, contextually relevant knowledge beyond surface-level representations.

2 Training and Datasets

We analyzed voxel responses from the 'core' language network, comprising the Inferior Frontal Gyrus, Inferior Frontal Gyrus – Orbital part, Middle Frontal Gyrus, Anterior Temporal cortex, and Posterior Temporal cortex across three datasets where participants read and processed sentences during fMRI scanning: Pereira (2018) Pereira et al. (2018), Tuckute (2024) Tuckute et al. (2024b) and Caption Scene Dataset CSD (2025) Li et al. (2025). Additional details on datasets are provided in Appendix A.1.

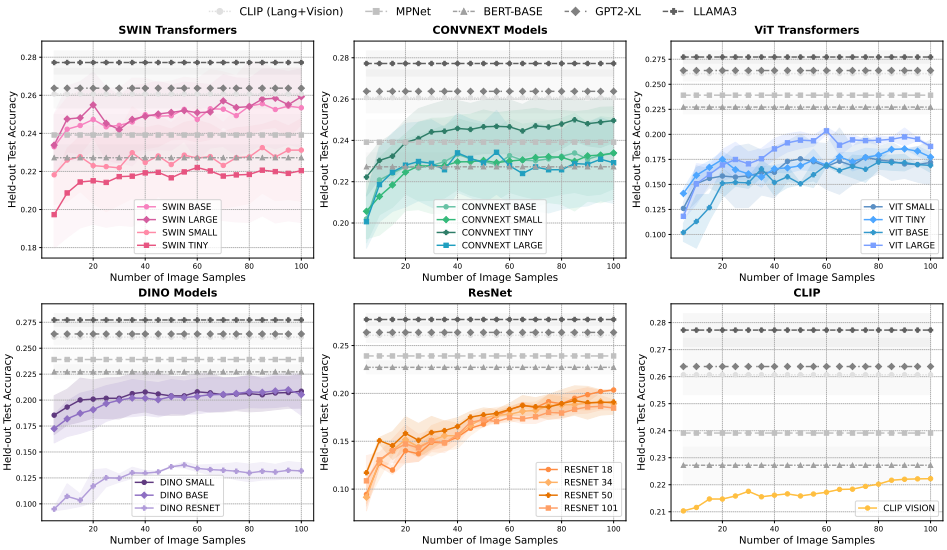

Figure 2: Comparison of performance in predicting language cortex activity between LLM embeddings of the original linguistic stimuli from the Pereira (2018) dataset (CLIP, MPNet, BERT-BASE, GPT2-XL and LLAMA3 presented in grey horizontal lines) and vision model embeddings of their corresponding visual counterparts (remaining models). Performance of the visual models increases with the number of images, sometimes surpassing some of the language models.

We first investigate whether meaning in the language cortex is captured exclusively by language model embeddings of the original sentence, or whether it can also be represented by alternative sentences that convey the same information. We also test whether representations from other modalities, such as vision, can capture this meaning (Figure 1). Finally, we assess whether predictive accuracy based on the original sentence can be further improved by enriching it with commonsense contextual information.

To evaluate predictive accuracy, we fit a ridge regression model $\boldsymbol{Y} \approx \boldsymbol{EW}$ by minimizing $\|\boldsymbol{Y} - \boldsymbol{EW}\|_F^2 + \lambda\|\boldsymbol{W}\|_F^2$, where $\boldsymbol{Y}$ denotes the voxel responses, $\boldsymbol{E}$ the embeddings from either vision or language models, and $\boldsymbol{W}$ the regression weights (examples in Appendix Figures 10, 12, 11, 14 and 15). The parameter $\lambda > 0$ is the regularization hyperparameter, chosen via cross-validation to prevent overfitting. More information on training can be found in Appendix Section A.2. We use the following featurizations:

A. **Original sentences:** We use the original sentence $\{S_i\}_{i=1}^N$ presented to the subjects, obtain the penultimate layer embeddings from an LLM as $\boldsymbol{E} = \{\boldsymbol{PL}(S_i)\}_{i=1}^N$ where $\boldsymbol{PL}(\cdot)$ denotes the penultimate layer embedding.

B. **Content words:** We extract the most concrete and semantically salient terms $\{c_j\}_{j=1}^{C^S}$ from each sentence: open-class parts of speech such as nouns and main verbs carrying the core meaning of the sentence. They are identified using a SciPy-based syntactic parser. For example, from "The boy is eating pancakes", we extract "boy" and "pancakes". Each content word is embedded separately using the penultimate layer of the LLM, then averaged to create a sentence-level representation: $\boldsymbol{E} = \frac{1}{C_S}\Sigma_{c_j}\boldsymbol{PL}(c_j)$.

C. **Header words:** For the Pereira (2018) dataset Pereira et al. (2018), sentences are grouped into paragraphs sharing a common topic. We use the paragraph header $\boldsymbol{H}$ as a high-level semantic summary. The header embedding $\boldsymbol{E} = \boldsymbol{PL}(\boldsymbol{H})$ is used to predict the averaged brain responses $\boldsymbol{Y} = \frac{1}{K}\Sigma_{j=1}^K \boldsymbol{Y}_{S_j}$ across all sentences in the paragraph, where $\boldsymbol{Y}_{S_j}$ is the brain response for sentence S_j and K is the number of sentences in the paragraph. Note that in the Pereira (2018) dataset's third experiment, several paragraphs described the same topic, we therefore merged all such paragraphs and used the shared topic as the paragraph header.

D. **Standard paraphrases:** We generate $\boldsymbol{R} = 70$ paraphrases (Appendix Table 1) for each sentence using Gemini Comanici et al. (2025) (alternative phrasings that preserve the original semantic meaning while varying in surface form). To analyze how the number of paraphrases affects prediction accuracy, we systematically sample subsets in increments of 5 (i.e., $r \in \{5, 10, 15, ., 70\}$). For each subset size r, we average the embeddings $\boldsymbol{E} = \frac{1}{r}\Sigma_{i=1}^{r}\boldsymbol{PL}(P_i^S)$, where P_i^S is the i^{th} paraphrase generated for sentence $\boldsymbol{S}$. This incremental sampling allows us to track how sentence-level representational diversity impacts brain prediction accuracy.

E. **Enriched paraphrases:** We generate $\boldsymbol{R} = 70$ paraphrases (Appendix Table 2) that embed broader contextual and inferential content: details that a human might naturally associate with a sentence, even if not explicitly stated. We process these enriched paraphrases in the same way as the standard paraphrases described in D.

F. **Sentence-Generated Images:** We use the stable diffusion model Rombach et al. (2022) to generate $M = 100$ images for each sentence $\boldsymbol{S}$, using the sentence itself as the prompt. While generating textual descriptions from images is relatively well-studied, expressing the full meaning of a sentence in a single image is far more challenging. Sentences often convey abstract concepts or temporally extended events that cannot be captured in one snapshot. To mitigate this, we create a diverse set of images per sentence, each offering a complementary visual perspective that together approximate the sentence's semantics. From these images we extract embeddings from the penultimate layer of a state-of-the-art vision model $\{\boldsymbol{VM}(I_i^S)\}_{i=1}^{M}$. We then average a subset of $m \in \{5, 10, 15, .., 100\}$ out of M embeddings to form a single representation $E = \frac{1}{m}\Sigma_{i=1}^{m}\boldsymbol{VM}(I_i^S)$, where I_i^S is the i^{th} image generated for sentence $\boldsymbol{S}$.

G. **Content-word images:** For each content word $\{c_j\}_{j=1}^{C^S}$ in sentence $\boldsymbol{S}$, we use the stable diffusion model to generate M images. From these images we extract embeddings from the penultimate layer of a state-of-the-art vision model $\{\boldsymbol{VM}(I_k^{c_j})\}_{k=1}^{M}$. We first average a selected subset of m embeddings to obtain a single representation for each content word, and then average across all content word representations to produce a sentence-level embedding $\boldsymbol{E} = \frac{1}{C^S}\Sigma_{c_j}\frac{1}{m}\Sigma_{k=1}^{m}\boldsymbol{VM}(I_k^{c_j})$, where $I_k^{c_j}$ is the k^{th} image generated for content word c_j.

H **Header images:** We generated M images of the same header word, get vision model embeddings of them, and average a subset m of these images to get an input representation $\boldsymbol{E} = \frac{1}{m}\Sigma_{i=1}^{m}\boldsymbol{VM}(I_i^H)$ to predict $\boldsymbol{Y} = \frac{1}{K}\Sigma_{j=1}^{K}\boldsymbol{Y}_{S_j}$ as described above, where I_i^H is the i^{th} image generated for header word $\boldsymbol{H}$.

We used a wide range of vision models spanning diverse architectures and training objectives He et al. (2016); Liu et al. (2022; 2021); Dosovitskiy et al. (2020); Caron et al. (2021); Radford et al. (2021), and language models spanning encoder–decoder and decoder-only architectures, with causal and non-causal variants Song et al. (2020); Devlin et al. (2019); Radford et al. (2019); Team et al. (2025); Dubey et al. (2024) (details are provided in Appendix Sections A.3 and A.4).

3 RESULTS

3.1 VISUAL MODELS CAPTURE MEANING IN THE LANGUAGE CORTEX

3.1.1 SENTENCE-LEVEL COMPARISON

First, we compare how vision and language models predict brain responses in the language cortex. Using LLM embeddings of each full sentence and vision-model embeddings of their corresponding generated images (A vs F), we train encoding models to predict cortical activity. With only a single image per sentence, language-based models outperform vision-based ones, though vision embeddings also demonstrate meaningful predictive power for language cortex activity, with some models (such as SWIN transformers) achieving competitive performance. This pattern emerges both in the Pereira (2018) dataset using images generated from sentence prompts (Figure 2) and in the CSD (2025) dataset using original COCO images that correspond to the caption stimuli (Figure 3). As we increase the number of generated images per sentence and average their embeddings, the performance of vision models improves substantially—sometimes even surpassing certain language models (see Pereira (2018) results in Figure 2, Tuckute (2024) results in Appendix Figures 17, and CSD (2025) in Appendix Figure 18).

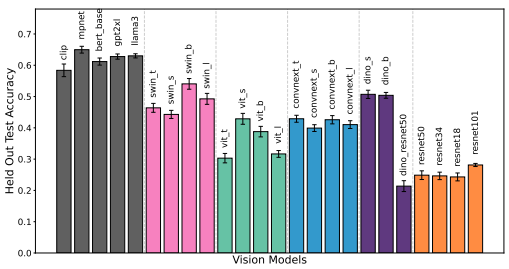

Figure 3: Comparison of performance in predicting language cortex activity between LLM embeddings of the original linguistic stimuli from CSD (2025) dataset (first 5 bars) and single image vision model embedding of their original COCO visual counterparts (remaining bars).

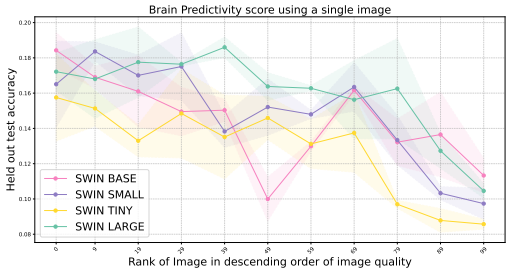

Figure 4: Performance of SWIN vision models in predicting language cortex activity using a single image, with experiments repeated across images sorted in order of decreasing quality (defined as the cosine similarity between the sentence and the image's CLIP embeddings).

This cross-modal success demonstrates that language cortex responses can be predicted using representations from visual content that preserves the semantic meaning. While individual images may miss certain semantic nuances, aggregating multiple visual perspectives systematically improves neural prediction accuracy. The robustness of this representational averaging effect, where performance gains from multiple images occur consistently across diverse vision architectures and training objectives highlights that this is a fundamental property of how semantic content can be distilled from varied visual exemplars, even though baseline performance levels differ across model families.

A critical nuance in our analysis is in the quality of the generated images, which is computed as the cosine similarity between CLIP-based embeddings of each image and its corresponding sentence. We use this similarity as a proxy for semantic alignment. Despite being generated from the same sentence, images vary widely in how accurately they capture the intended meaning, as diffusion models are not perfect and often produce off-topic or visually noisy samples (Appendix Figure 13).

Currently, all generated images are treated equally during the averaging process. However, only a subset of these images truly reflect the core content of the sentence, while others may diverge significantly. Our analysis shows that this variance directly impacts model performance: the predictive performance of models using single image embeddings declines systematically with decreasing image quality (Figure 4), demonstrating that low quality images fail to encode the critical semantic information present in the original sentence, leading to poorer alignment with brain responses.

We tested whether sorting images by semantic quality yielded more interpretable performance trajectories. When we incrementally include images in descending order of CLIP-based semantic similarity, we observe a sharp rise in accuracy as high-quality images are added, followed by a plateau, and eventually a decline as lower-quality images introduce noise (Figure 5). This contrasts with random ordering, where prediction accuracy is consistently lower and continually increases with exemplar averaging, confirming that image quality systematically affects neural prediction accuracy.

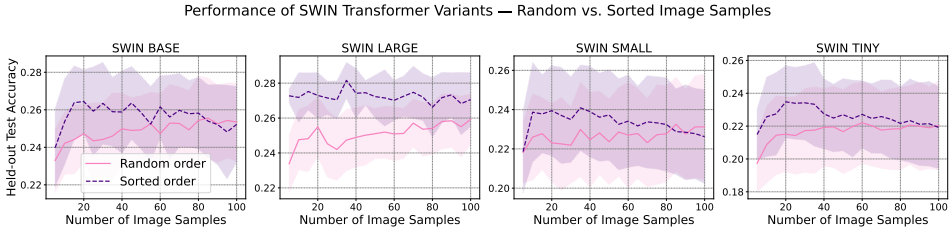

Figure 5: Pereira (2018) Dataset - Comparison of vision model performance in predicting language cortex activity using multiple images, with images ordered randomly versus in order of decreasing quality. Averaging more images helps in the case of random ordering, consistent with averaging the noise from non-ideal images. Adding more images eventually hurts in the case of ordered images, where eventually, less useful images are being incorporated.

3.1.2 CONTENT-WORD COMPARISON

Next, we tested whether the core semantic elements of sentences could predict language cortex responses when stripped of their full compositional structure. We evaluated the capacity of vision and language models to predict language-cortex responses by focusing on content words within each sentence—primarily nouns and main verbs that carry the essential semantic content (examples in Appendix Figure 12). This approach reduces sentences to their key conceptual building blocks while removing grammatical structure, function words, and compositional relationships. For each content word, we extracted embeddings from language models, and for the corresponding generated images produced by Stable Diffusion, we obtained visual embeddings using vision networks (featurization B vs G) (Figure 6). Even with a small number of content-word-based images, vision models achieved performance levels comparable to, and in most cases surpassing, those of certain language models such as BERT-Base and MPNet. We also observed a consistent upward trend in accuracy with increasing numbers of images, suggesting that aggregating multiple visual representations of individual concepts captures semantic nuances that contribute to language cortex responses.

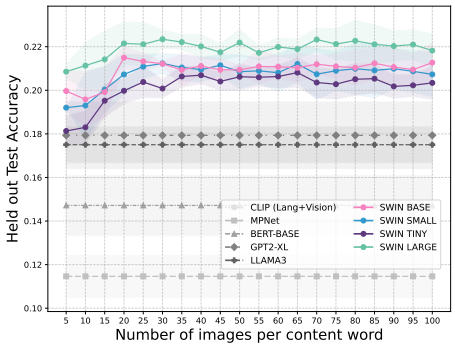

Figure 6: Comparison of performance in predicting language cortex activity between LLM embeddings of the content-word linguistic stimuli and SWIN model embeddings of their visual counterparts. Performance of SWIN vision models (which are trained without language supervision) increases with the number of images, surpassing many of the language models.

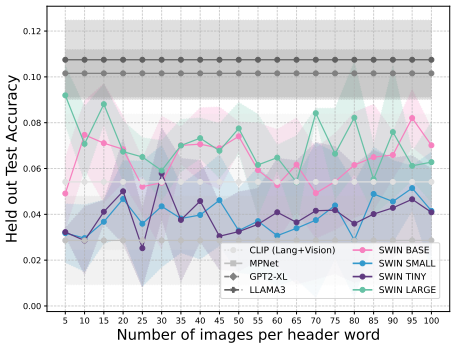

Figure 7: Comparison of performance at predicting language cortex activity between LLM embeddings of the header word groups from the Pereira (2018) dataset and SWIN model embeddings of their visual counterparts. For both language and vision models, the performance is worse than when looking at content-words or the original sentence.

These experiments reveal that when we reduce sentences to their core conceptual content, vision and language models achieve comparable predictive power for language cortex responses. The fact that vision model embeddings of images of individual words like 'boy' and 'pancakes' can predict brain responses nearly as well as linguistic representations is particularly striking given that these vision models were trained purely on natural image statistics without language supervision. This suggests that the meaning component of language cortex responses taps into conceptual representations that can be accessed through the statistical structure of visual experience alone, indicating that semantic processing in the language cortex may be grounded in modality-independent principles that both linguistic and visual systems can converge upon. However, the remaining performance differences between content words and full sentences (compare Figures 6 and 2) highlight that compositional structure, grammatical relationships, and contextual integration also contribute meaningfully to language cortex responses, though perhaps to a lesser degree than the core conceptual content.

3.1.3 HEADER-WORD COMPARISON

We now compare LLM embeddings of header words with vision model embeddings of corresponding images to predict averaged brain responses to sentence groups sharing the same topic (C vs H). Vision and language models achieve remarkably comparable performance levels, though both perform substantially lower than when using original sentences or content words (Figure 7). The averaging effect with increasing numbers of images is consistent in smaller Swin models but modest, possibly because individual header words are already concrete and semantically focused enough

that a few images capture their core meaning. This experiment represents the most abstract test of cross-modal prediction: using single thematic words (e.g., 'toaster', 'beekeeping') to predict language cortex responses to entire passages on those topics. The near-equivalence between vision and language models at this level suggests that when semantic content is highly abstracted, modality differences become minimal. However, the overall lower accuracy indicates that much of the predictive power for language cortex responses comes from the richer semantic and compositional information present in full sentences and individual concepts, rather than abstract thematic categories alone.

3.2 PARAPHRASES

3.2.1 COMPARING ORIGINAL SENTENCES WITH PARAPHRASES

So far, we examined the potential of cross-modal representations for modeling the language cortex. Here, we turn to alternative representations within the same modality, comparing the modeling performance of LLM embeddings of the original sentence with the mean embeddings of its paraphrases (A vs D). In the Pereira (2018) dataset, paraphrase embeddings alone yielded reasonable prediction accuracy. While averaging over a small number of paraphrases (≤ 5) underperformed relative to the original sentence embeddings (Figure 8), performance improved steadily as more paraphrases were incorporated (pink curve), at times surpassing the original sentence baseline (grey curve). This benefit was weaker in the Tuckute (2024) dataset (Appendix Figure 21), likely for two reasons: (i) responses in this dataset were averaged across voxels, which eliminated voxel-specific information that may carry important signal, and (ii) Pereira (2018) paragraphs are longer, richer and more content-diverse, making paraphrases more likely to be diverse, whereas Tuckute (2024) sentences are simpler and more abstract, yielding paraphrases with limited variation (Appendix Tables 14, 15).

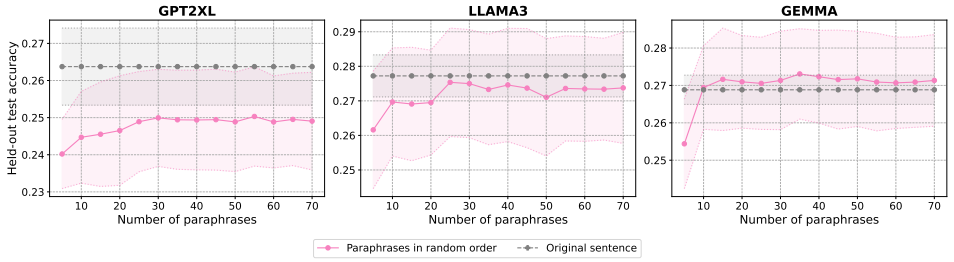

Figure 8: Pereira (2018) dataset - Comparison of the performance in predicting language cortex activity by LLM embeddings of the original linguistic stimuli presented to subjects with averaged embeddings of generated paraphrases. Averaging embeddings across more paraphrases steadily improves prediction accuracy, in some cases matching or surpassing the original sentence.

This pattern parallels our vision-based experiments, where a single image was less predictive than the original sentence but prediction accuracy improved as multiple visual samples were averaged. As in the vision case, alternative modalities or representational forms show greater potential for modeling the language cortex when the original stimuli presented to subjects are information-rich and concrete. Similarly, when we sort the paraphrases based on their semantic similarity to the original sentence (measured using the cosine similarity in the CLIP language encoder space) and incrementally add them in descending order of similarity (Appendix Figures 22 and 23), we observe a similar trend: accuracy improves rapidly at first and then saturates (green plot). This suggests diminishing returns as lower quality or less semantically similar paraphrases are added, highlighting that not all paraphrases contribute equally, and the quality of information matters as much as quantity.

Averaging paraphrase embeddings narrows, but typically does not erase the gap to the original sentence; it exceeds the original mainly with stronger LLMs (Gemma3) and for semantically rich stimuli (e.g., Pereira (2018)). This pattern suggests two components in the language cortex signal: (i) a content-dominant component recoverable from many surface realizations, and (ii) a form-sensitive residual captured best by the original sentence.

We also tested whether paraphrases add complementary information to the original by concatenating the original embedding with the paraphrase average (orange curve in Appendix Figs. 24, 25). In Pereira (2018), concatenation surpasses the original, indicating that paraphrases contribute ad-

ditional, brain-relevant content beyond surface form. In Tuckute (2024), concatenation improves over paraphrases alone but does not surpass the original, consistent with paraphrases adding little new information for short/abstract stimuli. These findings suggest that paraphrase-based representations contribute most when they introduce substantive semantic variation, whereas in the absence of such variation the precise linguistic form of the original sentence remains the strongest predictor of language-cortex responses.

3.2.2 Comparing Original Sentences with Commonsense-Enriched Paraphrases

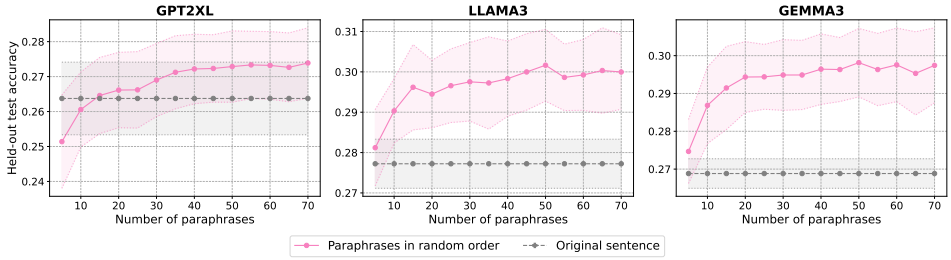

Figure 9: Pereira (2018) Dataset - Comparison of LLM embeddings of the original linguistic stimuli presented to subjects with averaged embeddings of generated paraphrases with additional context. Averaging embeddings of paraphrases enriched with contextual information yields higher prediction accuracy than the original sentence embeddings.

Our earlier experiments hinted at a key idea: adding supplemental information to a sentence, beyond its original form can often enhance our ability to model brain responses. In the above analyses, we preserved the original sentence embedding and concatenated it with embeddings from paraphrases. However, those paraphrases were semantically near-identical to the original sentence, merely differing in surface structure. As a result, the added information was limited, more a stylistic variation than a substantive expansion of meaning.

This led us to ask a deeper question: can paraphrases enriched with new commonsense context, while still preserving the core semantics of the original sentence further improve brain predictivity (featurization A vs E)? In this next set of experiments (illustrated in Figure 9, Appendix figure 19), we intentionally moved beyond superficial rewording and generated paraphrases that embedded broader contextual and inferential content: details that a human might naturally associate with the sentence, even if not explicitly stated. These enriched paraphrases preserve the original intent, but vary more significantly in both form and informational density. More broadly, we were interested in whether adding such rich context could compensate for or even outweigh the importance of original form or modality when modeling responses in the language cortex (Appendix Table 14).

We found that the effectiveness of this approach depended strongly on the nature of the dataset. In the Pereira (2018) dataset, whose sentences are typically rich, concrete, and highly visualizable, it was relatively easy to generate extended paraphrases with coherent, plausible context. In these cases, even a small number of enriched paraphrases substantially outperformed the original sentences in modeling brain activity (pink curve in Figure 9), and performance continued to improve without plateauing. This suggests that additional meaningful information, particularly information the brain may implicitly infer can have a strong additive effect on neural predictivity. We next asked whether combining the original sentence with these enriched paraphrases could further improve model performance, reasoning that the specific linguistic form of the original sentence might add information beyond paraphrased content (Appendix Figure 27). When we concatenated the original sentence embedding with the averaged paraphrase embedding, we did not see as major an improvement as we had seen when we used paraphrases without additional context. For older or less powerful models, such as GPT-2 XL, concatenation yielded a modest boost relative to paraphrases alone, likely because these models benefit from the precise structure of the original sentence combined with the additional semantic variation from the paraphrases. However, for more powerful LLMs, such as Llama 3 and Gemma 3, the improvement was negligible, and in some cases the paraphrases alone performed slightly better than the concatenated representation. These stronger models already en-

code rich, contextually robust representations of the original sentence, so the explicit addition of the original embedding contributes little new information.

By contrast, for the Tuckute (2024) dataset this strategy was less effective. These sentences are short (6 words long) and some of them abstract, making it difficult to generate paraphrases that are both semantically coherent and genuinely informative. Many paraphrases introduced noise or drifted from the original meaning (Appendix Table 15), so the enriched paraphrases alone failed to outperform the original sentences (Appendix Figure 26). In this setting, the exact linguistic form of the stimulus carried more predictive power than additional semantic content. As a result, appending the original sentence by concatenation led to performance gains, although it does not always beat the prediction accuracy when the embedding of the original sentence is used. (Appendix Figure 28).

These results suggest that predictive accuracy in the language cortex is shaped strongly by the informativeness of semantic content. When added information is meaningful and consistent with the brain's associative priors, it enhances predictivity. Thus, enriching inputs with inferential context can provide substantial gains beyond what is available from the original linguistic form alone.

4 DISCUSSION

In this work, we investigated whether the human language cortex can be modeled by stimuli that differ from the original linguistic input in modality or form while preserving semantic content. Our central aim was to examine the limits of representational flexibility, that is, the extent to which neural responses to linguistic stimuli can be accurately predicted using representations derived from alternative inputs that vary dramatically in surface form (paraphrases), modality (visual vs. linguistic), and information density (enriched vs. original content).

These analyses provide converging evidence that the human language cortex maintains highly abstract, form-independent representations of semantic meaning. This conclusion emerges from three key observations: vision model embeddings can predict language cortex responses when aggregated across multiple images, paraphrase embeddings show similar predictive power when averaged, and semantically enriched paraphrases can exceed the predictive accuracy of original sentences.

The success of cross-modal prediction using visual representations to model language cortex responses is particularly striking given that the language cortex was not exposed to any visual input during the original experiments. This suggests that meaning in the language cortex is encoded in a format that transcends specific sensory modalities. Further, the effectiveness of representational averaging across examples sharing semantic content: whether multiple images or paraphrases suggests that this process amplifies shared meaning while reducing noise from surface-level variations. This finding demonstrates that the meaning component of language cortex responses can be captured even when input content is embedded in highly variable surface forms or accessed through entirely different sensory modalities. Additionally, the finding that enriched paraphrases can exceed original sentence predictions suggests that the language cortex constructs enriched semantic representations that extend far beyond the information literally present in the text, incorporating the contextual associations and commonsense inferences that humans naturally bring to comprehension. This finding suggests a core difference in the breadth of semantic associations that brains and language models maintain, with the brain capturing richer and more encompassing representations, perhaps due to the broader range of real-world tasks it learns to perform that go beyond next word prediction.

Several limitations should be acknowledged. First, our visual generation relied on current diffusion models, which may not optimally capture the visual content most relevant to language processing. Future work could explore whether more sophisticated multimodal models or human-generated images yield different patterns. Second, our commonsense enrichment was generated algorithmically and may not reflect the specific knowledge integration processes that occur during natural reading or listening, where representations evolve on a word-by-word basis. Lastly, the fMRI datasets used in this study have inherently slow temporal resolution, providing only coarse snapshots of neural activity averaged over several seconds rather than millisecond-level dynamics. The dataset-dependent effects we observed highlight the need for more diverse neuroimaging corpora that span the full range of linguistic content, from concrete to abstract.

Despite these limitations, our work demonstrates how powerful generative models can be leveraged to create alternative input types that preserve semantic content while varying in modality (visual vs.

linguistic) and surface form (paraphrases), enabling cross-modal and cross-form neural prediction. Through averaging model representations of multiple generated variants, we can isolate the shared meaning components that drive neural responses. This methodological approach opens new avenues for understanding how intelligent systems, both biological and artificial, extract and represent meaning from experience that can be captured across diverse sensory modalities and surface forms.

REFERENCES

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wendy A De Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A Blank, and Jonathan C Kao. What are large language models mapping to in the brain? a case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*, 2024.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63, 2024.

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

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

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

Yuanning Li, Shurui Li, Zheyu Jin, Shi Gu, and Ru-Yuan Zhang. A large-scale vision-language fmri dataset for multi-modal semantic processing. *Journal of Vision*, 25(9):2445–2445, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

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.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*, pp. 2020–06, 2020.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *Advances in neural information processing systems*, 32, 2019.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 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.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science*, 2(11):745–757, 2022.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47(2024):277–301, 2024a.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024b.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014a.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 233–243, 2014b.

A Appendix

A.1 Datasets

The following datasets were used in our study:

1. Pereira et al. (2018), who recorded brain responses from 16 native English speakers as they silently read short, syntactically and semantically diverse passages. The study comprised three experiments: (1) isolated words and simple phrases to probe basic lexical–semantic representations, (2) 384 full sentences grouped into 96 short narratives on varied concrete topics such as professions, clothing, birds, musical instruments, natural disasters, and crimes, and (3) an additional 243 full sentences organized into 72 narratives of similar thematic breadth. For our analyses, we focus on the five participants who completed both Experiments 2 and 3, using only these two experiments because they provide full-length sentence stimuli: 627 unique sentences in total, suitable for modeling rich compositional semantics.
2. Tuckute et al. (2024b), who recorded brain responses from five native English speakers as they read 1,000 six-word sentences in an event-related fMRI design. The sentences were sampled from text corpora to maximize both semantic coverage and stylistic diversity. Each sentence was presented individually (2 s on screen, 4 s inter-stimulus interval), with runs of 50 sentences and short fixation blocks interleaved. Participants were instructed to read attentively and think about the meaning of each sentence, and to encourage engagement, they were informed that a brief memory task would follow the scan. Because each stimulus was presented to participants only once, we trained the encoding models to predict the average response across all voxels in the functionally-defined five language fROIs. This dataset provides brain responses to a semantically and stylistically diverse set of decontextualized sentences, well suited for modeling sentence-level representations.

   Note that the sentences in this dataset are relatively abstract and simpler compared to those in the Pereira (2018) dataset. As a result, many sentences lack clearly identifiable content words (see Section 2). For this reason, we did not perform the content-word experiments on this dataset.
3. Li et al. (2025) (Caption Scene Dataset) comprises fMRI data from eight participants performing a semantic matching task: each participant first read a Chinese caption and then viewed a corresponding MS-COCO image Lin et al. (2014) to decide whether the text and image matched semantically. For our analyses, we focused on the four subjects with the highest signal-to-noise ratios. Because the image was shown only after the caption, the neural responses during caption reading are uncontaminated by visual input. We therefore analyze only these "pure language" responses from the caption-reading phase. The full dataset contains 9,375 unique captions and 9,494 images (18,893 total stimuli). From these, we use the 983 caption stimuli that were presented to all four selected subjects. The accuracies reported on this dataset are noise-normalized, with noise ceiling computed following the NSD procedure Allen et al. (2022). In our case, given only two repetitions for each stimulus, the noise ceiling is equivalent to split-half reliability. Note that since the

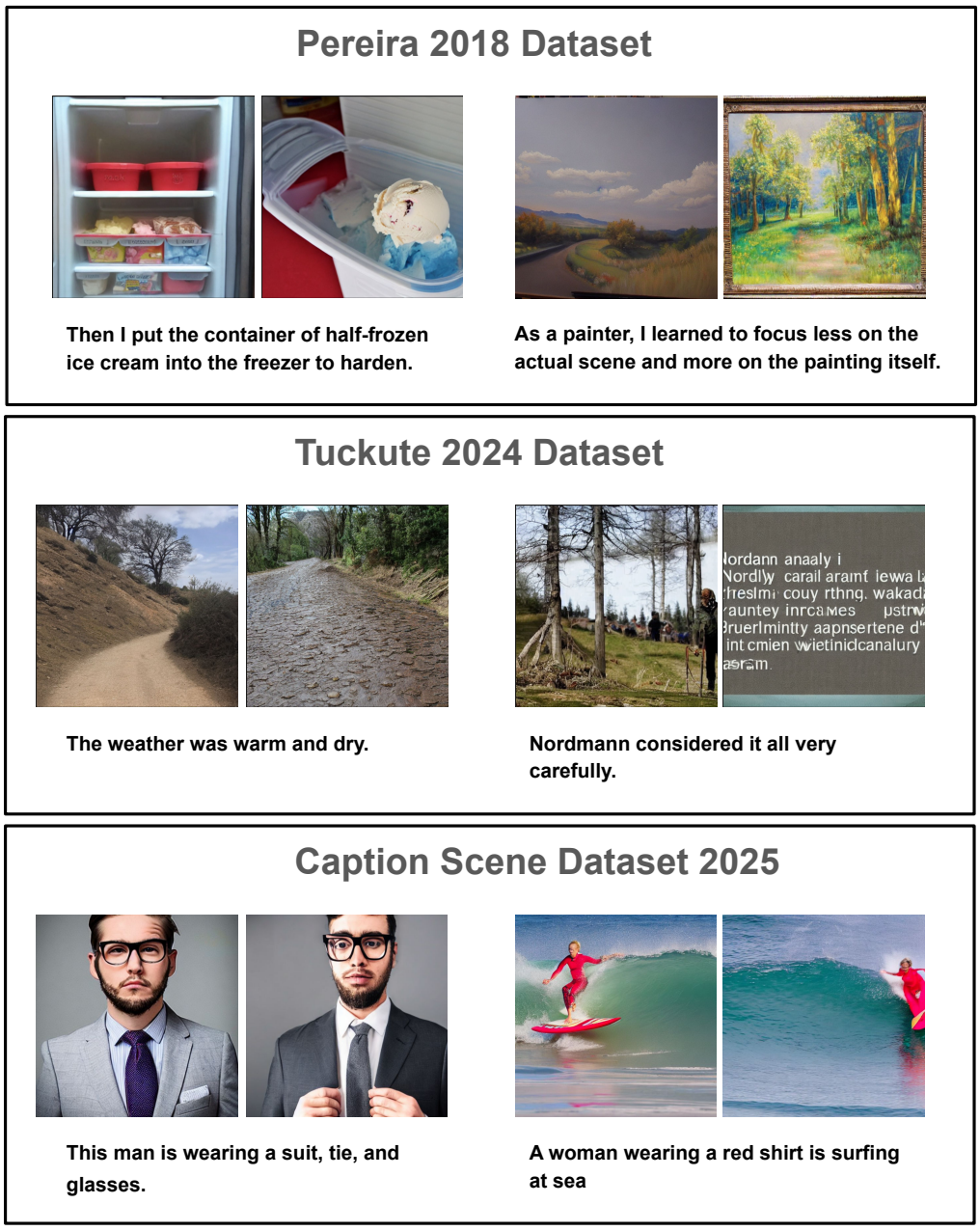

Figure 10: Examples of sentences used, and images generated for these sentences using Stable Diffusion

captions in the Captions Scene Dataset are written in Chinese, we do not apply our linguistic modulation analyses, such as isolating content words or generating paraphrases to this dataset.

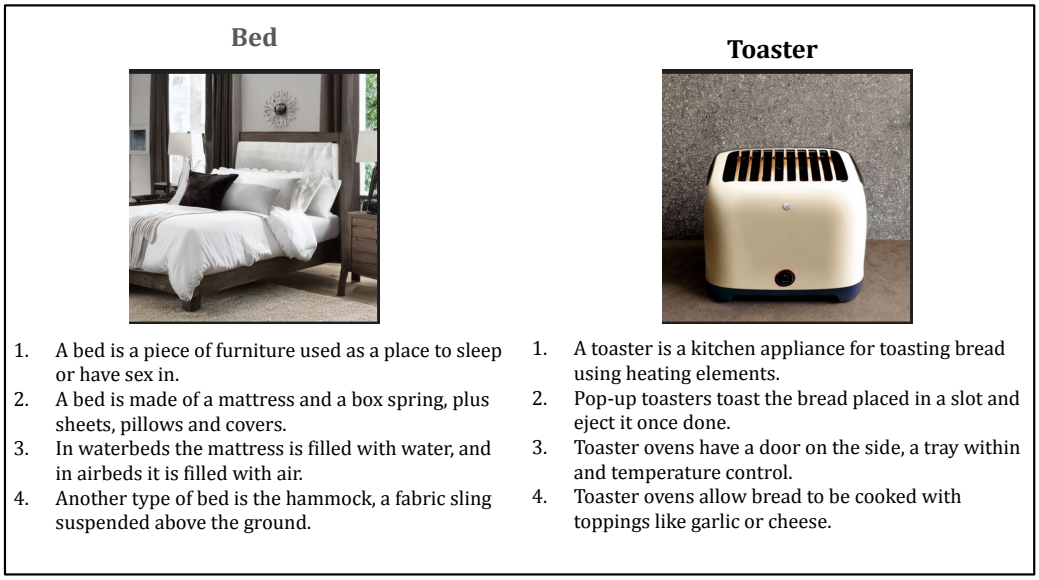

Figure 11: Examples of header words for a group of sentences in Pereira (2018) dataset, and images generated for these header words using Stable Diffusion

Figure 12: Examples of content words/phrases in a sentence, and images generated for these content words/phrases using Stable Diffusion

A.2 Training Details

For the Pereira (2018) dataset, we trained the model using backpropagation with a combined mean squared error and correlation loss between the true and predicted voxel responses. This is because a few voxel measurements contained `NaN` values, we retained all columns and avoided discarding affected columns. Brain responses for the remaining datasets were estimated using the closed-form ridge solution, with λ chosen by k-fold cross-validation ($k = 1$ for the Tuckute (2024) dataset and $k = 10$ for the CSD (2025) dataset). Test accuracy is quantified as the Pearson correlation coefficient computed between the predicted and observed voxel responses across the held-out test set.

When training via backpropagation, we cross-validated across a grid of λ values $(0.0, 0.1, 0.01, 0.001, 0.0001)$. Further, for each dataset we generated multiple train–validation–test splits (3 for Pereira (2018), and 5 for the rest) and reported the averaged results across them. For the Pereira (2018) dataset, we designed these splits to ensure a fair evaluation, since having sentences from the same paragraph appear in both training and test sets could allow the model to exploit shared context Kauf et al. (2024); Feghhi et al. (2024). To assess whether the observed trends hold independently of contextual overlap and temporal autocorrelation, we created two random splits and one targeted variants: a split in which the test set contained only the first sentence from each of 63 paragraphs.

Bees crawl across his bare arms and hands, but they don't sting, because they're gentle

In developed countries, blindness is more likely to be caused by genetic problems.

Figure 13: Examples of images generated for a given sentence, ranked in descending order of quality. Image quality is measured by the cosine similarity between the CLIP embeddings of the original sentence and the generated images.

Original Sentence	Paraphrases without any additional context	Paraphrases with commonsense context
Theft is a crime and can be punishable with jail or fines.	Illegally taking items is a criminal offense, resulting in prison or fines.	The unlawful appropriation of another's property is theft, a criminal offense that is subject to penalties including jail or financial sanctions imposed by the legal system.
	Committing theft incurs punishment such as jail or a fine.	The unauthorized taking of another's property, known as theft, is against the law and can result in penalties such as serving time in prison or paying monetary fines.
	Taking items illegally results in legal consequences: jail or fines.	The unlawful appropriation of property is theft, a criminal offense that is subject to penalties including jail or financial sanctions.
Many bees have stingers and can attack if they feel threatened.	Bees will sting to protect their colony when threatened.	The stinging apparatus of a bee is its main deterrent against predators or disturbance when it feels threatened.
	Bees can deliver a sting when they are feeling threatened.	The primary purpose of a bee's stinger is self-protection and the defense of its colony from harm.
	Stinging is how bees react to feeling unsafe or attacked.	Be careful around bees, especially near flowers or potential hive sites, as they can sting if they feel their space is invaded.

Figure 14: Examples of paraphrases generated for sentences in the Pereira (2018) dataset

A.3 Language Models used for Evaluation

1. **CLIP Language Encoder** Radford et al. (2021): The text branch of CLIP, trained jointly with its vision counterpart using a contrastive image–text objective. It produces rich sentence-level embeddings aligned to visual concepts, enabling zero-shot cross-modal tasks.

2. **MPNET** Song et al. (2020): A Transformer-based language model that combines masked language modeling with permutation-based training, allowing it to capture bidirectional context and sequential dependency simultaneously.

3. **BERT Base** Devlin et al. (2019): A bidirectional Transformer pretrained with masked language modeling, providing contextual word and sentence embeddings.

Original Sentence	Paraphrases without any additional context	Paraphrases with commonsense context
Do any stand out in particular?	Which feature has the most impact?	Which features in the photograph do you find most remarkable or noteworthy?
	What's your primary focus in this shot?	What features in the scene are most distinguishable or visually prominent?
	See anything that sparks curiosity?	Does anything seem particularly sharp or in focus compared to the rest?
This truly was a depressing place.	The band's rhythm feels reluctant.	Playing live, the band exhibited a hesitant rhythm, possibly due to opening night jitters.
	Uncertainty marks the band's rhythm.	The band played with a hesitant rhythm, unsure of the upcoming key change.
	The band's rhythm is marked by hesitation.	The rhythm played by the band was hesitant, making it hard to tap your foot along.

Figure 15: Examples of paraphrases generated for sentences in the Tuckute (2024) dataset

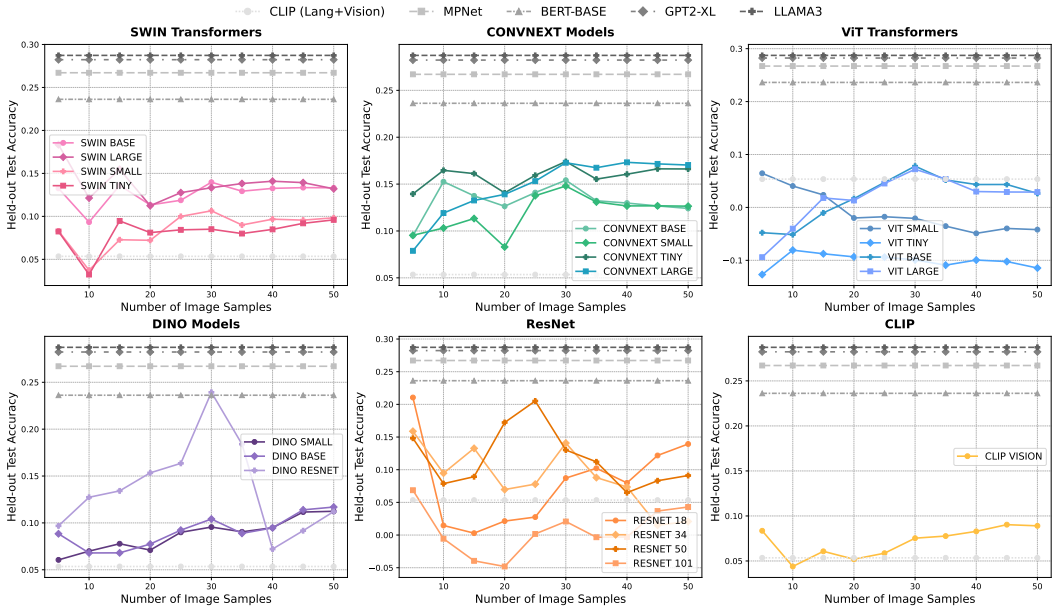

Figure 16: Performance comparison between LLM embeddings of the original linguistic stimuli from the Tuckute (2024) dataset and vision model embeddings of their corresponding visual counterparts.

4. **GPT-2 XL** Radford et al. (2019): An autoregressive Transformer with 1.5 B parameters, trained to predict the next token in large-scale web text.

5. **Gemma-3** Team et al. (2025): Google's latest open-weight decoder-only Transformer with grouped-query attention and alternating local/global layers for efficient long-context reasoning (up to 128K tokens). Larger variants add a SigLIP vision encoder for multimodal inputs and support quantization for lightweight deployment.

6. **LLaMA-3** Dubey et al. (2024): Meta's third-generation decoder-only Transformer featuring grouped-query attention and efficient scaling.

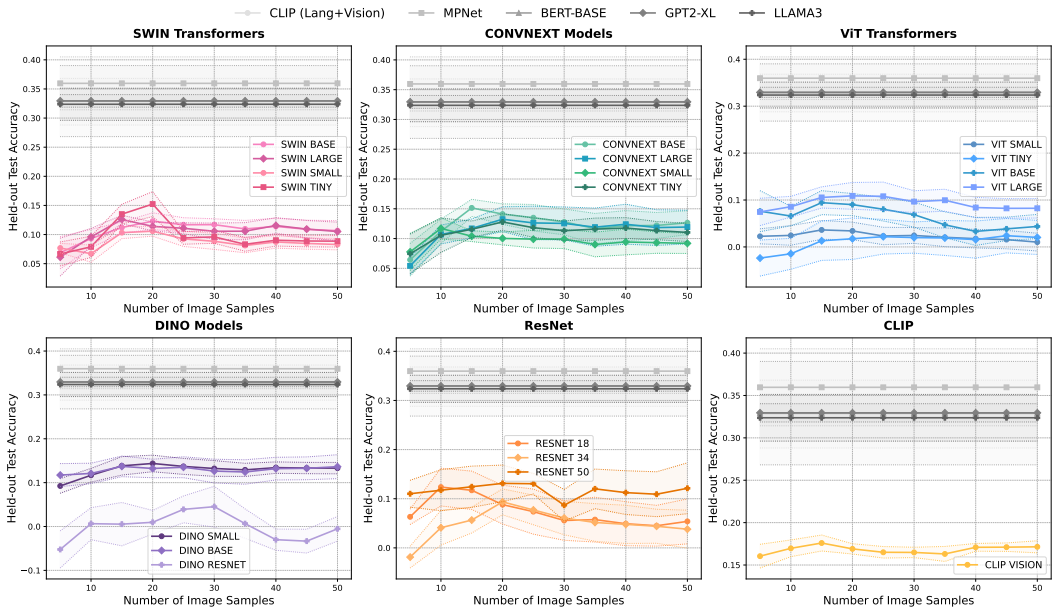

Figure 17: Evaluating LLM embeddings of Tuckute (2024) sentences versus vision-model embeddings of their visual counterparts, using only sentences with generated images of sufficient quality (mean CLIP score ≥ 0.25)

A.4 Vision Models used for Evaluation

1. **ResNet family** He et al. (2016) (ResNet-50, ResNet-34, ResNet-18, ResNet-101): Deep residual networks that use skip connections to mitigate vanishing gradients, enabling very deep convolutional architectures.
2. **ConvNeXt family** (ConvNeXt Small, Tiny, Base, Large) Liu et al. (2022): A modernized CNN design that incorporates architectural ideas from Transformers, such as large kernels and inverted bottlenecks.
3. **Swin Transformers** Liu et al. (2021) (Swin Small, Tiny, Base, Large): Hierarchical vision Transformers that process images using shifted windows, providing linear computational complexity with respect to image size.
4. **Vision Transformers (ViTs)** (ViT Small, Tiny, Base, Large) Dosovitskiy et al. (2020): Pure Transformer architectures that treat images as sequences of non-overlapping patches, capturing global context through self-attention.
5. **DINO models** (DINO Small, Base, ResNet-50) Caron et al. (2021): Self-supervised representations learned via knowledge distillation, producing strong, transferable visual features without labeled data and supporting both ViT and ResNet backbones.
6. **CLIP Vision Encoder** Radford et al. (2021): The visual branch of CLIP, trained with a contrastive objective to align images and text in a shared embedding space, enabling zero-shot recognition and robust cross-modal retrieval.

A.5 Evaluating Vision- and Language-Model Predictions of Language Cortex Activity - Tuckute (2024)

Applying the same vision-versus-language modeling framework to the Tuckute (2024) dataset as in Section 3.1.1, we do not observe the clear performance gains with increasing image counts seen for Pereira (2018). Many vision models fail to learn meaningful mappings, as indicated by their negative test-time correlation scores (Figure 16). This likely reflects the more abstract nature of the Tuckute (2024) sentences, which makes generating semantically aligned visuals more challenging (Figure 10). Nevertheless, the performance of vision-based models remains reasonably competitive,

indicating that even abstract sentences can evoke visual representations that capture meaningful information, albeit with more noise. Importantly, this observation still aligns with our core hypothesis: even when vision-based representations are noisier, they retain enough semantic signal to contribute meaningfully to neural prediction.

To account for this variability, we filtered the Tuckute (2024) dataset to retain only those images with a CLIP-based quality score above 0.25. Restricting our analysis to these higher-quality samples allowed us to recover the expected performance trend: as the number of relevant images increases, model accuracy improves. However, this trend is less not observed in ViT-based models (Figure 17). One possible explanation is that standard ViT architectures rely on fixed-size, non-overlapping image patches and lack mechanisms for hierarchical feature extraction or localized inductive biases. As a result, they may be less sensitive to fine-grained variations in image quality or spatial detail, especially in scenarios where the semantic content is subtle or distributed across small regions of the image.

fMRI datasets tailored for multimodal modeling, especially those that align closely with the hypotheses we aim to test are exceptionally rare. In particular, not all datasets contain sentences that are as visually grounded and easily translatable into images as those in the Pereira (2018) dataset. For example, many sentences in the Tuckute (2024) dataset are more abstract in nature, making it substantially more difficult to generate meaningful visual representations. This challenge is evident in the lower quality and occasionally off-topic images produced by the diffusion model, as shown in Appendix Figure 10.

A.6 Evaluating Vision- and Language-Model Predictions of Language Cortex Activity (Caption Scene Dataset 2025)

Lastly, we compare language-model embeddings of the full original sentence with vision-model embeddings of the corresponding visual scenes using the Caption Scene Dataset (2025), as described in Section 3.1.1. Because participants first read captions describing MS-COCO images, we begin by contrasting the language-model embeddings of these captions with the vision-model embeddings of the original COCO images.

Although the language models better capture brain activity in the language cortex, the vision models achieve performance that is only slightly lower. This reinforces our earlier finding that language cortex prediction remains highly sensitive to the exact linguistic form, giving language models an advantage, while the underlying semantic content can be represented nearly as well through visual modality.

We then extend this analysis by generating multiple synthetic images for each caption with Stable Diffusion and averaging their vision embeddings, following the procedure in Section 3.1.1. The same pattern emerges as in the Pereira (2018) and Tuckute (2024) datasets (Figures 2 and 17): vision-model performance improves as the number of generated samples increases, in some cases approaching that of the language models.

A.7 THE USAGE OF LARGE LANGUAGE MODELS (LLMS)

LLMs are primarily used to identify typos and make the language more aligned with conventions of academic writing.

Prompt to generate paraphrases

```
Prompt:
f"""You are an expert image captioner. I'll show you some existing captions for an
     image, and your task is to generate 70 NEW captions that:
   1. Are similar in style and detail level to the existing captions
   2. Capture the same meaning but with different wording
   3. Are direct, concise descriptions (around 10-15 words each)
   4. Are worded differently from each existing caption and from each other

Here are the existing captions:
{insert all captions text for the image here}

Generate 10 new captions formatted exactly as:
1. [First new caption]
2. [Second new caption]
3. [Third new caption]
4. [Fourth new caption]
5. [Fifth new caption]
6. [Sixth new caption]
7. [Seventh new caption]
8. [Eighth new caption]
9. [Ninth new caption]
10. [Tenth new caption] ... """
```

Table 1: Prompt used for generating paraphrase using Gemini-2.5-Flash.

Prompt to generate paraphrases with extra context

```
Prompt:
f"""You are an expert image captioner. I'll show you existing captions for an image,
and your task is to generate 70 NEW captions that:

1. Are similar in style but include more detail than the existing captions
2. Capture the same meaning but with different wording
3. Are worded differently from each existing caption and from each other
4. Contain additional commonsense context to the original caption

Example:
For the sentence 'The boy is eating pancakes for breakfast', some paraphrases
with additional context would be:
  1. The boy is eating pancakes with maple syrup in the morning for breakfast
  2. The boy is sitting at the dining table and having pancakes for breakfast

Here are the existing captions:
{insert all captions text for the image here}

Generate 10 new captions formatted exactly as:
1. [First new caption]
2. [Second new caption]
3. [Third new caption]
4. [Fourth new caption]
5. [Fifth new caption]
6. [Sixth new caption]
7. [Seventh new caption]
8. [Eighth new caption]
9. [Ninth new caption]
10. [Tenth new caption] ... """
```

Table 2: Prompt used for generating paraphrases with additional commonsense context using Gemini-2.5-Flash.

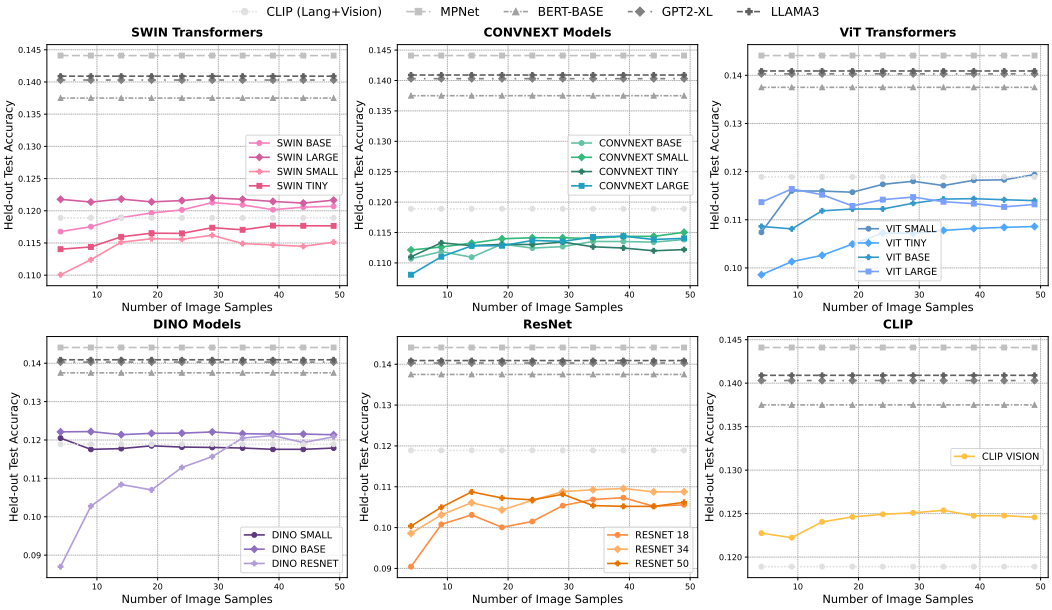

Figure 18: Performance comparison between LLM embeddings of the original linguistic stimuli from the CSD (2025) dataset and vision model embeddings of their corresponding visual counterparts.

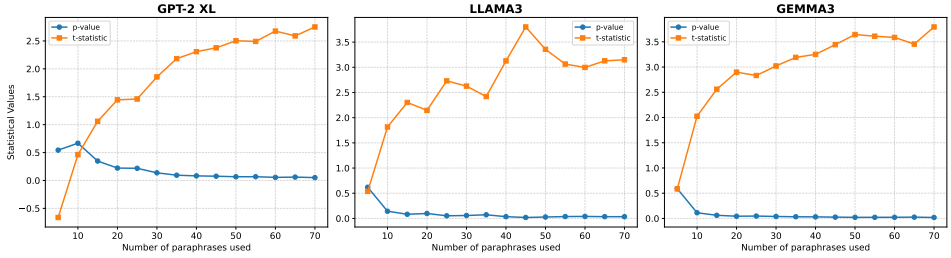

Figure 19: Expanded analysis of the experiments described in Section 3.2.2 across five data splits. For each condition, we compute test-time accuracies for averaged paraphrases and compare them with the accuracies obtained using the original single sentence. The t-statistics increase with the number of paraphrases, further supporting the claims in Section 3.2.2.

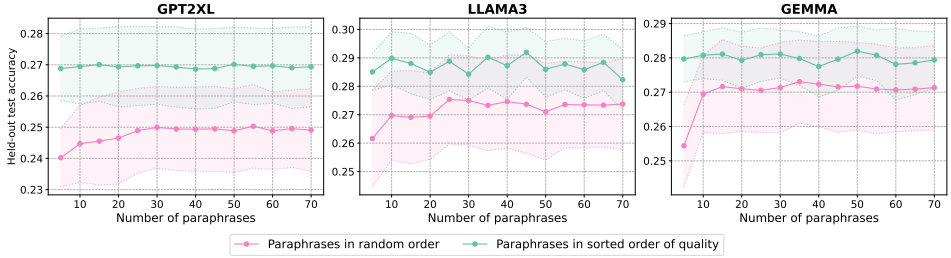

Figure 20: Pereira (2018) dataset - Comparison of averaged LLM embeddings of paraphrases in random order with those arranged in sorted order of semantic similarity to the original sentence. These paraphrases do not have added commonsense context.

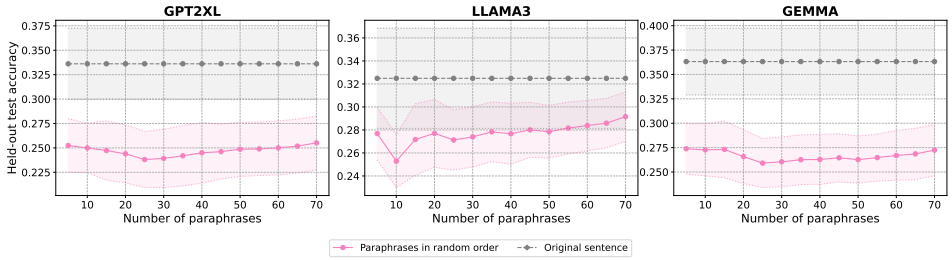

Figure 21: Tuckute (2024) dataset - Comparison of LLM embeddings of the original linguistic stimuli presented to subjects with averaged embeddings of generated paraphrases without additional context.

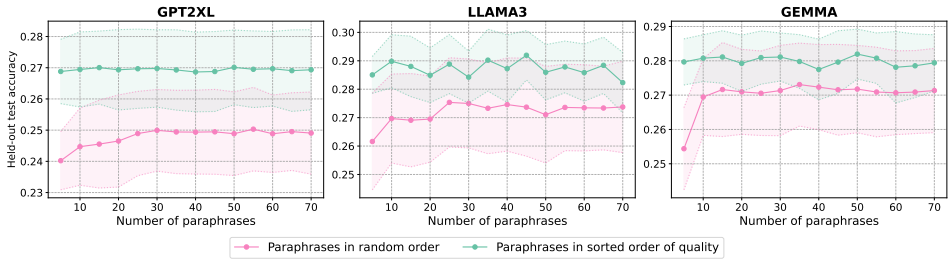

Figure 22: Pereira (2018) dataset - Comparison of averaged LLM embeddings of paraphrases in random order with those arranged in sorted order of semantic similarity to the original sentence. These paraphrases do not have added commonsense context.

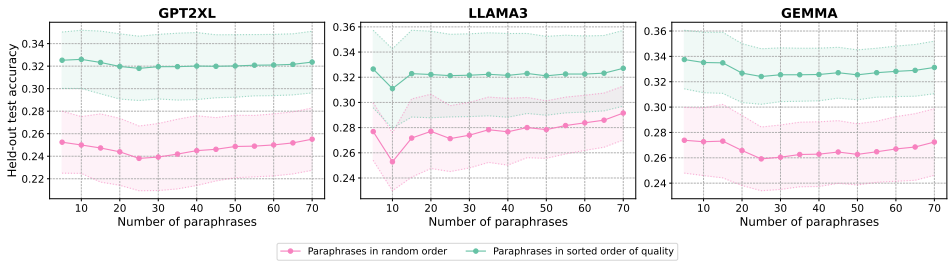

Figure 23: Tuckute (2024) dataset - Comparison of averaged LLM embeddings of paraphrases in random order with those arranged in sorted order of semantic similarity to the original sentence. These paraphrases do not have added commonsense context.

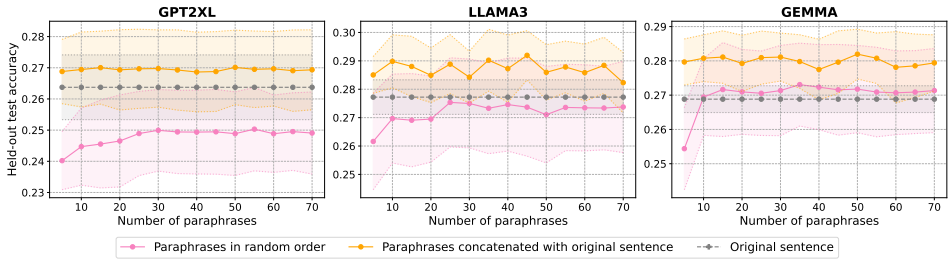

Figure 24: Pereira (2018) dataset - Comparison of averaged LLM embeddings of paraphrases alone with those that are concatenated with the original sentence. These paraphrases do not have added commonsense context.

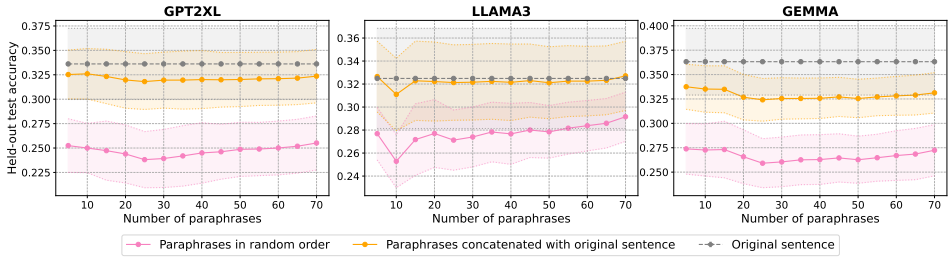

Figure 25: Tuckute (2024) dataset - Comparison of averaged LLM embeddings of paraphrases alone with those that are concatenated with the original sentence. These paraphrases do not have added commonsense context.

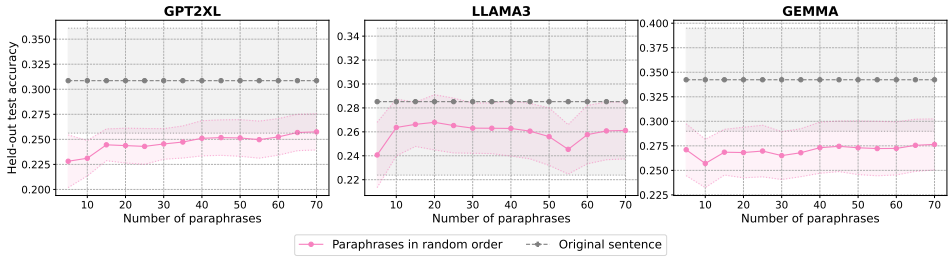

Figure 26: Tuckute (2024) dataset - Comparison of LLM embeddings of the original linguistic stimuli presented to subjects with averaged embeddings of generated paraphrases with additional context.

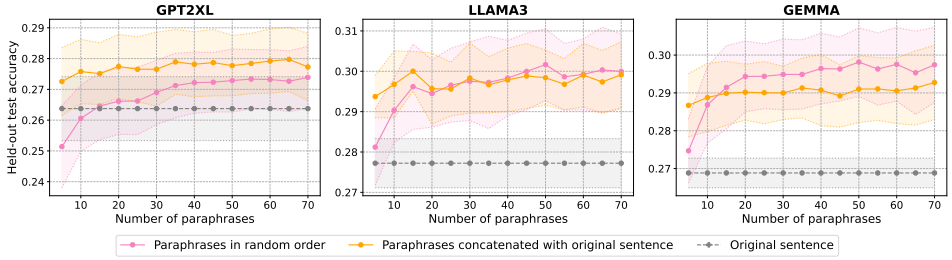

Figure 27: Pereira (2018) dataset - Comparison of averaged LLM embeddings of paraphrases alone with those that are concatenated with the original sentence. These paraphrases have added commonsense context.

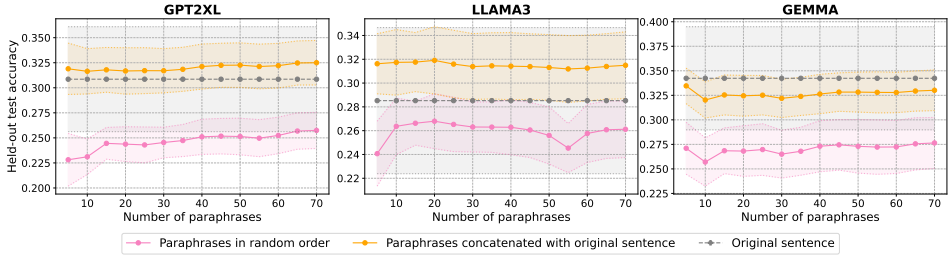

Figure 28: Tuckute (2024) dataset - Comparison of averaged LLM embeddings of paraphrases alone with those that are concatenated with the original sentence. These paraphrases have added commonsense context.