Language, Computation and Cognition Project

Project 2 - Word Embeddings and the Brain

Yarden Nahum - 207362096   Maor Zelkin - 325549681

Abstract

Understanding how linguistic features influence neural responses during language comprehension is essential for bridging language models and brain activity. In this study, we extend analyses presented by Pereira et al. (2018) by leveraging the fMRI dataset from Tuckute et al. (2024) to explore two additional linguistic features. First, we followed the structured part of the project to further expand on Pereira et al. (2018), then we investigated whether sentences containing named entities (e.g., PERSON, ORG) obtained stronger neural responses compared to sentences without named entities. Using statistical analyses, we confirmed that sentences with named entities indeed produced significantly stronger activation. Additionally, we conducted a fine-grained analysis, revealing that specific entity types, such as PERSON and ORG, particularly drive this increased activation. Moreover, we explored sentiment as an additional linguistic dimension by evaluating whether sentiment classifications from a language model align with participant-based valence ratings. Our results demonstrated significant correspondence between human ratings and model-based sentiment predictions, indicating the effectiveness of current language models in capturing sentiment as a linguistic feature. Link to our code repo can be found at the bottom of our page.[1]

1 Introduction

Recent advances in computational linguistics and cognitive neuroscience have allowed researchers to explore the neural correlates of language processing using language models and fMRI data. Pereira et al. (2018) investigated how semantic information from word embeddings aligns with neural activation patterns. Specifically, they conducted three analyses: (1) decoding individual words based on voxel-level neural data (analysis 1), (2) decoding sentence-level meanings by averaging word embeddings and relating them to neural activations (analysis 2), and (3) extending analysis 2 with additional sentences covering different topics (analysis 3). Their decoder reached a mean pair-wise classification accuracy of about 0.77 and a rank accuracy of about 0.74 across 180 concepts, showing that static embeddings predict neural responses well above chance but still leave room for unexplained variance.

Building on this foundation, Tuckute et al. (2024) demonstrated that sentence embeddings derived from the GPT-2 XL language model effectively predict neural responses in human language networks, suggesting that linguistic representations captured by such models meaningfully correspond to human brain activity.

Our study builds upon these findings by examining specific linguistic features that might influence neural responses. Specifically, we explored two questions: First, do sentences containing named entities (e.g., persons, locations, organizations) trigger stronger neural responses compared to sentences without entities, and if so, which types of entities are most influential? Additionally, we performed a region-wise brain analysis to identify which type of entity affects each region the most. Second, does a RoBERTa sentiment analysis model classify sentences' sentiment similarly to the way humans do, and can it reproduce the conclusions made by Tuckute et al. (2024)?

[1] https://github.com/yarden077/fmri-brain-response-sentence-decoding/tree/main

Through our analysis, we found that sentences containing named entities elicited significantly stronger neural activations, as confirmed by t-tests. This suggests that named entities serve as prominent linguistic features capable of robustly activating language regions. Furthermore, we observed an alignment between sentiment classification and human valence ratings, consistent with results reported by Tuckute et al. (2024), indicating that modern language model effectively captures emotional nuances of sentences. These findings highlight the practical usefulness of a BERT-based sentiment model for predicting how strongly sentences activate the brain based on emotional content, which has valuable applications in medical settings, such as assessing brain activity following surgery. Using our approach, doctors could quickly evaluate the level of brain response to any given sentence, making assessments faster and more consistent without relying on additional human analysis. Knowing which kinds of sentences consistently amplify or dampen language-region responses could help clinicians design quick, standardized probes for assessing language-network integrity in patients (e.g., before or after neurosurgery) and guide future work on brain-aligned text generation in computational linguistics. Overall, these results provide practical insights into generating linguistic stimuli that reliably influence neural activity, emphasizing named entities and emotional valence as critical dimensions for future research in cognitive neuroscience and computational linguistics.

2 Data

2.1 Structured part dataset

In the structured part, we used the Pereira et al. (2018) dataset, consisting of fMRI recordings from 16 participants. The dataset consists of 3 parts for each experiment done in the article. One is a dataset of 180 words which were shown to the participants, along with brain activation imagery of the participant when reading the word. Second is a set of 96 text passages, each consisting of 4 sentences, together with the brain activity of each participant. Lastly, a second set of 72 passages, each consisting of 3 or 4 sentences (unrelated topics to the first set of sentences) and the brain imagery. In all parts, brain activity was represented using a vector of voxels, allowing for detailed spatial analysis across individual brain regions.

2.2 Open-ended part dataset

The dataset from Tuckute et al. (2024) includes fMRI recordings from 14 participants exposed to linguistically diverse stimuli. Unlike Pereira et al.'s voxel-level approach, Tuckute et al. aggregated neural responses into functional Regions of Interest (ROIs), defined specifically to target language-processing brain areas. Participants viewed 2,000 sentences chosen to maximize linguistic variety, including syntactic, semantic, and stylistic diversity. Of these sentences, 1,000 were baseline sentences selected from naturalistic corpora, while the remaining 1,000 sentences were specifically chosen to either strongly activate ("drive") or minimize ("suppress") responses in language-related brain regions. Sentences were presented individually, and neural responses were recorded. The dataset further includes behavioral ratings collected from 3,600 additional participants, who evaluated each sentence on multiple linguistic dimensions (e.g., grammaticality, plausibility, emotional valence).

3 Experiments and Results

3.1 Structured tasks

In the structured tasks, we utilized the Pereira et al. (2018) dataset, conducting a series of analyses aimed at exploring the relationships between neural signals and semantic representations of words and sentences. In their paper, they utilized a text embedding model to decode the imaging data to determine which words and pictures the subjects were looking at during the experiment. Initially, we replicated analysis 1 of Pereira et al. (2018), extending it by using fastText embedding model from Mikolov et al. (2017) and comparing the results. Our comparison indicated that fastText embeddings provided slightly superior decoding "average rank" accuracy (53.51) compared to GloVe (61.91). (see Fig.1., appendix.1. and appendix.2.)

Subsequently, we compared Analyses 1, 2, and 3 from Pereira et al. (2018). Analysis 1 focused on decoding individual words, whereas Analyses 2 and 3 examined sentence-level decoding. Then We applied the GloVe-based decoder model, that was trained on the data of analysis 1, to the sentence-

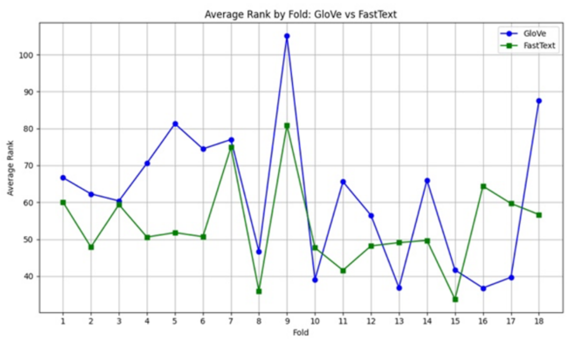


Fig.1. Comparison of average rank between GloVe (blue) and FastText (green) across 18 cross-validation folds.
Average rank measures how close each decoded vector is to its true semantic vector by ranking all possible vectors by similarity. Lower ranks indicate better decoding (in our case, 1 is best, 180 is worst, 90 is chance level). The graph shows that FastText outperforms GloVe in 14 out of 18 folds, indicating a slight overall advantage.

level datasets from analyses 2 (384 sentences) and 3 (243 sentences), and compared the decoded vectors to sentence-level embeddings. The sentence-level embeddings were generated by averaging individual word embeddings within each sentence. The results are relatively poor (Mean average rank across 18 cross-validation folds: analysis 2 - 156.92; analysis 3 - 100.74) which may be expected given that the decoder was originally trained on individual word representations.

A detailed topic-based analysis revealed variability in decoding accuracy, suggesting that sentences about dreams, stress, and castles were decoded with relatively low average ranks, suggesting that the neural patterns associated with these topics were captured more effectively by the decoder. On the other hand, topics like beekeeping, owl, and lawn-mowers resulted in much higher ranks, indicating that the decoder struggled more with these concepts (see fig.2.)

In Task 2, we investigated sentence representation models further by training decoders on the sentence-level data from analysis 3 using two types of sentence embeddings: static embeddings from the original Pereira et al. (2018) paper (average GloVe embeddings) and contextual BERT embeddings from Wang et al. (2020). Our results demonstrated that contextualized embeddings outperformed static embeddings (Mean average rank across 18 cross-validation folds: BERT - 89.28; GloVe - 100.02), underscoring the enhanced representational

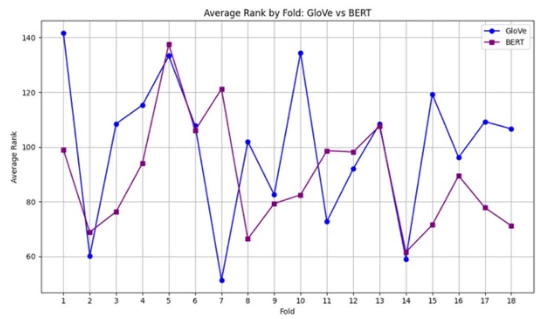


Fig.3. Average rank across 18 cross-validation folds for static GloVe embeddings (blue) versus contextualized BERT embeddings (purple). Bert outperforms GloVe in 10 out of 18 folds - overall means 89.28 for BERT vs. 100.02 for GloVe - showing that contextualized sentence representations capture neural variance slightly more accurately than static word vectors.

capacity of contextualized models like BERT for sentence-level neural decoding (see fig.3., appendix.3. and appendix.4.)

Finally, in Task 3, we switched from neural decoding to neural encoding, aiming to predict voxel-level brain responses from sentence embeddings using linear regression. We computed the R^2 scores to evaluate model performance. Contextualized embeddings again showed superior performance, capturing neural signal variability more effectively(see table 1 and fig.4.). However, both models produced poor results, as they were able to predict fewer than 0.05% (23 / 185,866) of the voxels.

3.2 Open – ended task

In the open-ended task, our primary goal was to extend the analyses performed by Pereira et al. (2018) by exploring additional linguistic features and their relation to neural activations, utilizing the dataset from Tuckute et al. (2024).

We first investigated whether sentences containing named entities led to stronger neural responses compared to sentences without named entities. To accomplish this, we explored multiple named entity recognition (NER) models including BERT-based NER (Devlin et al., 2019), spaCy's transformer-based model (Honnibal et al., 2020), and the Stanza NER model (Qi et al., 2020). After evaluating model performance qualitatively on a subset of sentences, we selected spaCy's transformer model for its comprehensive entity classification.

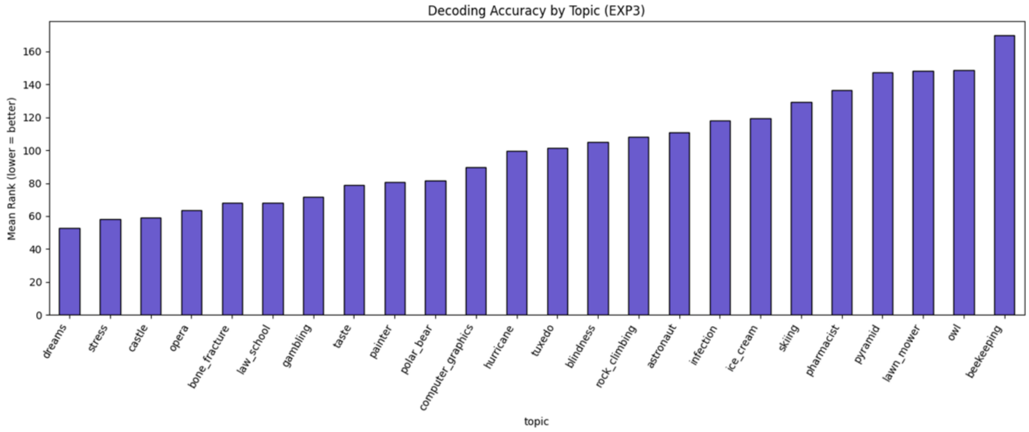


Fig. 2. Topic-wise decoding performance for Pereira et al.'s analysis 3 (EXP3). Bars show the mean rank assigned by GloVe-based decoder to each topic; lower values indicate better decoding accuracy. Topics such as *dreams (52.8)*, *stress (58.2)*, and *castle (59.1)* yielded the lowest mean ranks, suggesting that their neural patterns were captured most reliably. In contrast, topics like *beekeeping (169.8)*, *owl (148.6)*, and *lawn mower (148.1)* produced the highest ranks, indicating weaker decoder performance. Overall, the figure highlights substantial variability in how well semantic content is decoded across topics, pointing to topic-specific differences in the mapping between fMRI responses and embedding space.

Metric	**GloVe**	**BERT**
Mean R^2	-1.994	-1.538
Voxels with $R^2 > 0$	2 / 185,866	23 / 185,866
Voxels with $R^2 > 0.1$	0 / 185,866	8 / 185,866
Mean R^2 for voxels with $R^2 > 0.1$	N/A	0.178

Table 1: Linear regression performance (R^2) in voxel-wise neural encoding using sentence embeddings. BERT-based contextualized embeddings outperform static GloVe embeddings, yielding higher R^2 values and more voxels with predictive signal ($R^2 > 0.1$), indicating better alignment with neural responses.

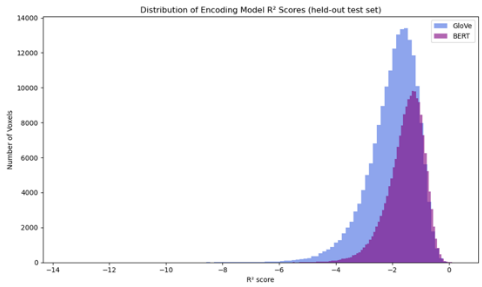


Fig 4. Distribution of voxel-wise encoding performance for GloVe vs. BERT embeddings (held-out test set). Histogram shows the number of voxels as a function of the linear-regression R^2 obtained when predicting fMRI activity from sentence embeddings. Both models explain very little variance (scores cluster well below 0), but the BERT curve (purple) is shifted modestly rightward relative to GloVe (blue), confirming the small yet reliable advantage of contextualized embeddings noted in the text (<0.05 % of voxels reach $R^2 > 0$).

We tagged each sentence from Tuckute et al.'s (2024) dataset (2,000 unique sentences) with entity labels such as PERSON, ORG, GPE, LOC, DATE, TIME, MONEY, QUANTITY, and several others.

Subsequently, we created binary indicators marking the presence or absence of each entity type per sentence. Out of 2,000 unique sentences, 568 contain at least one named entity, while 1,432 contain none. We plotted the distribution of entity type among sentences (see fig.5.) which shows that ORG and PERSON entities are three to four times more frequent than other entities.

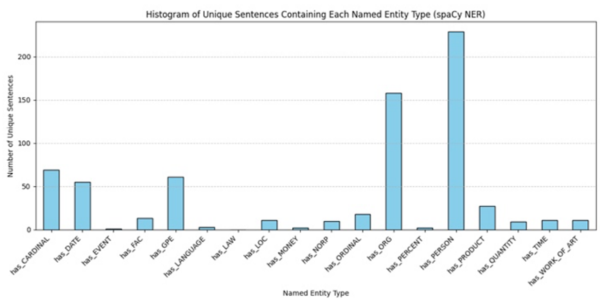


Fig. 5. Distribution of named-entity types across the 2,000 unique sentences in Tuckute et al.'s (2024) corpus. Bars show the count of sentences that contain at least one entity of each spaCy NER class. PERSON and ORG entities are the most frequent, whereas categories such as EVENT or LAW are rare.

A t-test was conducted to statistically examine differences in neural responses between sentences containing entities and those without. A visual inspection of the histograms (see Fig.6. and appendix.5.) showed that the BOLD-response distributions for sentences with and without named entities were roughly symmetric and bell-shaped, so the normality assumption for a t-test was reasonable.

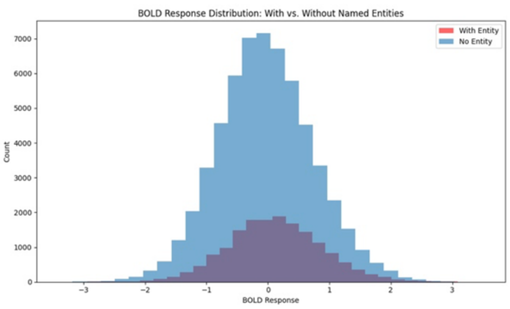


Fig. 6. Distribution of BOLD responses for sentences with (red) versus without (blue) named entities. Both curves are bell-shaped and roughly symmetric, supporting the assumption of normality required for the subsequent independent-samples *t*-test. Visual inspection confirms similar variance, and the noticeable rightward shift of the purple histogram foreshadows the significant increase in mean activation for entity-bearing sentences reported in Section 3.2.

Our analysis showed that sentences containing named entities had significantly higher neural responses (t-test: $T = 22.57$, $p < 0.001$), confirming that named entities play an important role in modulating language-region responses (see table 2).

	Mean response	STD	T-statistic	P-value
with entities	0.1193	0.7630	22.5706	<0.001
without entities	-0.0339	0.7733		

Table 2: T - test results. Mean BOLD responses are shown for sentences that contain at least one named entity and for sentences that do not. A t-test between the groups ($T = 22.57$, $p < 0.001$) reveals a highly significant difference: sentences with entities evoke stronger brain activity than sentences without entities.

To understand this effect more deeply, we performed additional analyses to explore which specific entity types influenced neural responses most strongly. We moved from a two-group t-test to a sentence-level linear regression to examine the influence of each entity. To isolate the impact of each individual entity type, we first removed any sentence that contained two or more different entity labels. By restricting the analysis to sentences with a single entity, we could be confident that any change in BOLD response was attributable to that specific entity category rather than a mixture of several co-occurring entities. Out of the 568 sentences containing entities, 456 contain only one type of entity. For every sentence we created a binary feature for each entity label (1 if the entity appears in the sentence). We then fit an ordinary least-squares model that predicted the BOLD response from this set of dummy variables. This approach lets us estimate how much the presence of each entity type raises or lowers neural activity while holding all other types constant. Coefficients were tested with two-tailed t-statistics; terms with $p<0.05$ were deemed significant (see Table 3). The model revealed positive weights for EVENT ($\beta = 0.43$), LANGUAGE ($\beta = 0.41$), MONEY ($\beta = 0.24$), WORK_OF_ART ($\beta = 0.16$), ORG ($\beta = 0.14$), PERSON ($\beta = 0.12$), and PRODUCT ($\beta = 0.10$) indicating these entities relate to stronger language-network responses. These results refine our initial finding: not all entities are equal, entities related to social interaction (e.g., PERSON, ORG), cultural content (e.g., WORK_OF_ART, LANGUAGE), and abstract or symbolic concepts (e.g., MONEY, EVENT, PRODUCT) are associated with stronger neural activation in language networks, while spatial or quantitative references show no significant effects.

We also conducted a region-wise analysis across several regions of interest (ROIs) in the brain to pinpoint which specific areas were most sensitive to named entities. For every ROI we split the single-trial BOLD values into two groups - sentences with at least one named entity and sentences without - and ran an independent-samples t-test. The script loops over all ROI labels, records the group means, Δ-mean (mean_with_entities - mean_without_entities), t-statistic, and p-value, and then ranks ROIs by absolute effect size. The left inferior frontal gyrus (IFG) and middle frontal gyrus (MFG) showed the strongest and most reliable boosts in activation for entity-bearing sentences (IFG: Δmean = 0.21, $t = 12.22$, $p < 0.001$; MFG: Δmean = 0.1722, $t = 9.42$, $p < 0.001$), whereas temporal ROIs displayed smaller or non-significant differences (see table 4). This pattern indicates that frontal language areas are particularly sensitive to the presence of named entities, highlighting functional heterogeneity within the language network.

	Coef.	std err	T - statistic	P-value	95% CI	Metrics	values
const	0.0519	0.015	3.521	**<0.001**	[0.023, 0.081]	R^2	0.010
PERSON	0.1298	0.019	6.966	**<0.001**	[0.093, 0.166]	Adj- R^2	0.009
ORG	0.1472	0.021	6.898	**<0.001**	[0.105, 0.189]	F-statistic	10.83
GPE	-0.1099	0.061	-1.808	0.071	[-0.229, 0.009]	P-value	<0.001
LOC	0.043	0.028	1.573	0.116	[-0.011, 0.098]	AIC	$3.073 * 10^4$
DATE	-0.0418	0.026	-1.607	0.108	[-0.093, 0.009]	BIC	$3.084 * 10^4$
TIME	0.0314	0.045	0.704	0.481	[-0.056, 0.119]		
MONEY	0.2486	0.088	2.832	**0.005**	[0.076, 0.421]		
QUANTITY	-0.0553	0.065	-0.846	0.397	[-0.183, 0.073]		
PERCENT	0.0696	0.100	0.698	0.485	[-0.126, 0.265]		
EVENT	0.4366	0.221	1.974	**0.048**	[0.003, 0.870]		
PRODUCT	0.1067	0.041	2.594	**0.009**	[0.026, 0.187]		
WORK_OF_ART	0.2680	0.062	4.335	**<0.001**	[0.147, 0.389]		
LANGUAGE	0.4130	0.140	2.943	**0.003**	[0.138, 0.688]		

Table 3: OLS Results. Columns report the coefficient (effect size), standard error, t-statistic, and two-sided p-value for each binary indicator that a sentence contains a given entity type. All statistically significant terms have positive coefficients (e.g., PERSON, ORG, MONEY, EVENT, PRODUCT, WORK_OF_ART, LANGUAGE), which indicate that these entities are associated with higher language-network activation. Model-level statistics (rightmost columns) show that the overall regression is significant ($F = 10.83$, $p < 0.001$), though the explained variance is small (Adj. $R^2 = 0.009$). Together, the table supports the conclusion that not all entity types modulate neural responses equally, with culturally and socially loaded entities driving the strongest effects.

	Δmean	T-statistic	P-value
lang_LH_IFG	0.2134	12.22	<0.001
lang_LH_MFG	0.1722	9.42	<0.001
lang_LH_netw	0.1443	10.08	<0.001
lang_LH_PostTemp	0.1359	8.72	<0.001
lang_LH_IFGorb	0.1306	7.03	<0.001
lang_LH_AntTemp	0.1226	8.18	<0.001

Table 4: Region-wise effect of named entities on BOLD response. For each left-hemisphere language ROI we report the mean activation difference between sentences with vs. without named entities (Δ mean), together with an independent-samples t statistic and the corresponding p-value. IFG and MFG show the largest boosts (Δ mean = 0.21 and 0.17, respectively; both $p < 0.001$), indicating that frontal language areas are especially sensitive to the presence of named entities; all language ROIs exhibit significant positive effects.

To investigate what kind of entity drives these ROI effects, we again excluded sentences that mentioned more than one entity type, leaving only "single-entity type" sentences, so that any boost could be attributed unambiguously to that tag. Within each ROI we then fit an ordinary-least-squares model in which the BOLD response was predicted from one-hot flags for every tag (PERSON, ORG, DATE, etc.).

After removing the intercept, we kept only predictors whose coefficients were significant at $p < 0.05$, and selected the tag with the largest influence. As Table 5 shows, the most impactful tag varied by region. In both the anterior temporal (AntTemp) and posterior temporal (PostTemp) regions, the entity WORK_OF_ART had the largest influence on activation levels ($\beta = 0.28$, $p = 0.040$; $\beta = 0.42$, $p = 0.003$, respectively). In the inferior frontal gyrus (IFG), ORG entities had the strongest effect ($\beta = 0.16$, $p = 0.002$), while PERSON entities were most predictive in the orbital part of IFG (IFGorb; $\beta = 0.12$, $p = 0.017$). Finally, EVENT entities had the most pronounced effect in the middle frontal gyrus (MFG), showing the highest coefficient overall ($\beta = 1.17$, $p = 0.049$). These results might suggest that different frontal and temporal language areas are tuned to different semantic cues, refining our earlier observation that named entities as a class heighten neural responses.

Our second major analysis focused on evaluating emotional valence and how sentiment analysis language models align with human valence judgments. We employed the RoBERTa-based sentiment analysis model (Camacho-Collados et al., 2022) to classify the sentiment (positive, neutral, negative) of each sentence. Model predictions were then compared to human ratings provided in the Tuckute et al. (2024) dataset. Before turning to the brain data, we first quantified the distribution of sentiment labels assigned by the RoBERTa classifier, confirming that the dataset is strongly skewed toward *neutral* sentences, with roughly balanced but much smaller *positive* and *negative* classes (see Fig.7. for the bar plot). This step ensured that later statistical tests would be interpreted in light of the class imbalance. We observed a strong correlation between model predictions and human judgments, validating the

ROI	Top entity	coefficent	P-value
lang_LH_AntTemp	WORK_OF_ART	0.281322	0.040063
lang_LH_IFG	ORG	0.164640	0.002684
lang_LH_IFGorb	PERSON	0.122313	0.017106
lang_LH_MFG	EVENT	1.173589	0.048585
lang_LH_PostTemp	WORK_OF_ART	0.417988	0.003387

Table 5. Top contributing named entity type for each ROI based on regression analysis using sentences with only one entity type (see appendix.6. for result metrics of each model). For each region, we identified the named entity that had the strongest positive association with brain response. WORK_OF_ART showed the highest impact in both the anterior and posterior temporal regions. ORG was most predictive in the IFG, PERSON in the IFGorb, and EVENT had the strongest effect in the MFG. All effects reported are statistically significant ($p < 0.05$), indicating distinct regional sensitivity to different semantic categories.

model's capacity to accurately capture emotional valence (see Fig.8. for boxplot analysis).

Additionally, we analyzed how neural responses corresponded to sentiment classifications. Our results indicated that sentences labeled as negative obtained significantly higher brain responses compared to neutral and positive sentences (see table 6). However, no significant difference was observed between neutral and positive sentences. These findings support the conclusion that negative emotional content robustly activates language networks in the brain as mentioned in Tuckute et al. (2024).

Overall, our experiments demonstrate that linguistic features such as named entities and emotional valence significantly influence neural responses. These insights offer practical implications for designing linguistic stimuli in cognitive neuroscience research and potential applications in clinical settings, where standardized probes could be used for assessing language-network integrity, for example in pre- or post-neurosurgical contexts.

	mean	T - statistic	P - value
negative vs neutral	Neg = 0.052 Neut = -0.011	8.715	< 0.001
neutral vs positive	Neut = -0.011 Pos = -0.023	1.575	0.1152
negative vs positive	Neg = 0.052 Pos = -0.023	8.162	< 0.001

Table 6: Pairwise t-tests linking sentiment labels to mean BOLD response.
Negative sentences ($M = 0.052$) produce reliably stronger activation than both neutral ($M = -0.011$) and positive ($M = -0.023$) sentences, with sizable t-statistics ($t = 8.72$ and 8.16, respectively; $p < 0.001$). The neutral–positive contrast is non-significant ($t = 1.58$, $p = 0.12$), indicating that elevated neural activity is specific to negative valence.

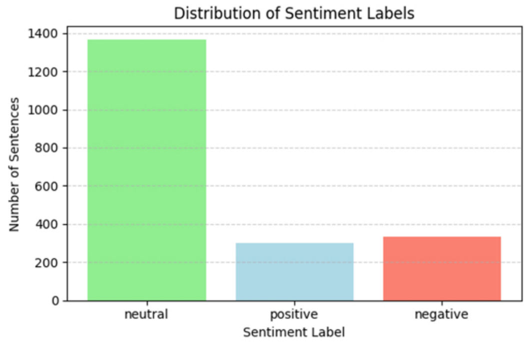


Fig.7. Distribution of RoBERTa sentiment labels in the 2,000-sentence dataset. The bar chart shows that the sentiment model assigns most sentences to the *neutral* class (1,366 items), whereas *negative* (333) and *positive* (301) sentences are far less frequent and roughly balanced with each other.

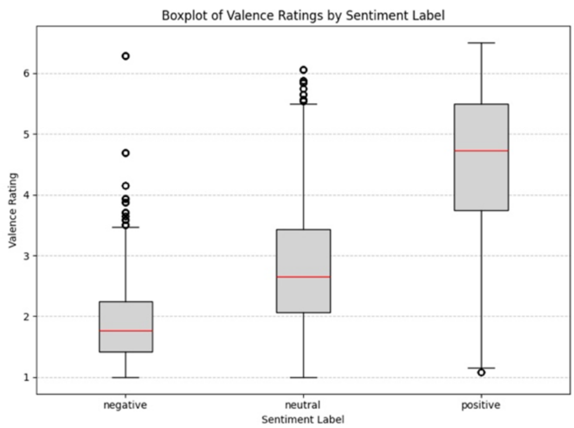


Fig.8. Human valence ratings versus model-assigned sentiment labels. Box-and-whisker plots show the distribution of crowd-sourced valence scores (1 = very negative, 7 = very positive) for sentences the RoBERTa sentiment classifier labeled *negative* (left), *neutral* (center), and *positive* (right). Medians (red lines) rise monotonically from negative to positive, and inter-quartile ranges are well separated, indicating that the model's categorical predictions track participants' affective judgments. This alignment might support the use of modern sentiment models as proxies for human emotional appraisal in brain-language studies.

4 Discussion and Conclusions

In this study, we extended prior analyses of neural language processing by examining how specific linguistic features such as word/sentence embedding, named entities and emotional sentiment influence neural activation patterns in language-related brain areas. Our analyses build on foundational work by Pereira et al. (2018) and recent insights from Tuckute et al. (2024), contributing novel empirical evidence and practical implications.

We first demonstrated that sentences containing named entities reliably elicit stronger neural activations than sentences without entities, supporting our hypothesis that named entities serve as prominent linguistic cues in language comprehension. A more detailed regression analysis further revealed that certain types of entities, especially those referencing languages, individuals, organizations, monetary concepts, events, products, works of art and languages, strongly enhance neural responses. Conversely, spatial and temporal entities (e.g., locations, dates, quantities) seem to have no effect on neural activation. This nuanced finding emphasizes that not all named entities equally influence brain activity, highlighting the importance of semantic content, specifically cultural and social relevance, in driving language-region responses.

Region-wise analysis underscored the frontal cortex's critical role in processing named entities, particularly in regions like the left inferior frontal gyrus (IFG) and middle frontal gyrus (MFG). These results support the idea of functional specialization within the language network, suggesting that frontal areas may prioritize socially and culturally salient information. Furthermore, region-wise analysis of entity types revealed distinct sensitivities within each region of interest. ORG entities elicited the strongest responses in the IFG, while EVENT-related entities had the highest impact in the MFG. These findings align with the idea of functional specialization, suggesting that different frontal regions may be tuned to specific semantic categories.

Our second major finding relates to emotional sentiment. We showed alignment between human sentiment ratings and predictions made by a RoBERTa-based language model, confirming the model's effectiveness in capturing emotional nuances. Furthermore, we established that negative emotional content consistently evokes stronger neural responses compared to neutral or positive content. This aligns with and extends the findings of Tuckute et al. (2024), adding empirical evidence for the heightened sensitivity of the language network to negative emotional stimuli.

Practically, these insights have significant implications. Clinically, understanding which linguistic features reliably activate language regions could aid in quickly creating and testing linguistic probes for assessing brain function, especially useful before or after neurosurgical interventions. Methodologically, our work emphasizes the value of combining computational linguistics with neural data to refine models that predict brain responses to language.

Nonetheless, this study has limitations. Our entity analysis relied on automated tagging, which, despite careful model selection, may still contain errors influencing our conclusions. We are also limited by the entity tags provided by the spaCy model. There may be more plausible entity categories or more precise definitions, for example, what qualifies as a WORK_OF_ART entity, which could affect our results.

Moreover, the sentiment data are highly imbalanced: about two-thirds of the 2,000 sentences were neutral, leaving far fewer positive and negative items. While our statistics are robust, a more balanced corpus would be able to create a better comparison between the sentiment model and humans. Thirdly, the public release of the Tuckute et al. (2024) dataset contains only 14 participants and focuses on left-hemisphere language ROIs; stronger or different entity effects might emerge in right-hemisphere or extra-language regions that were not examined. In addition, having more participants will add variance to the data which could alter the results.

Unlike the dataset from Pereira et al. (2018), which contains voxel-wise brain response data, the Tuckute et al. (2024) dataset provides only average brain responses. This loss of information limited our ability to conduct more precise analyses in the open part of the project.

Future research could focus on dealing with the limitations of our study. However, we hope that our analysis could be applied in the field of generative AI. A possible experiment could focus on generating high and low brain response sentences based on the features we have found out. Moreover, exploring these effects in diverse languages and populations could enhance the generalizability of these findings, potentially broadening their applicability in clinical and linguistic research contexts.

References

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. Nature communications, 9(1), 1–13.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. Nature Human Behaviour, 8(3), 544–561.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776-5788.

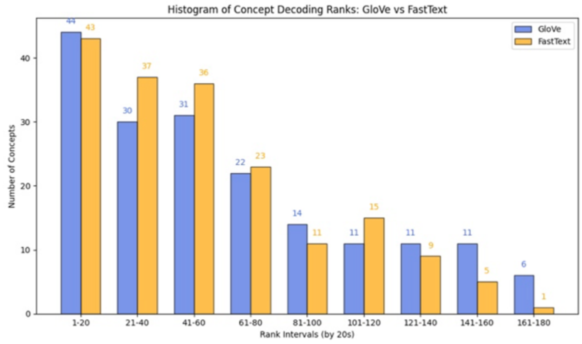
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python.

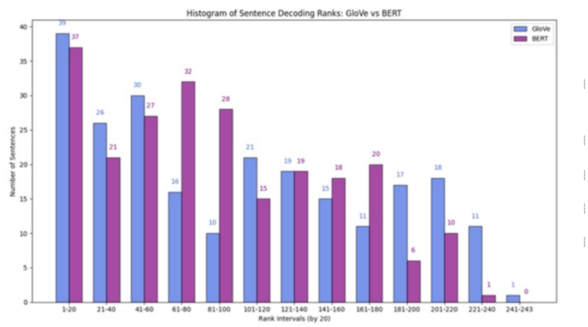
Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., ... & Barbieri, F. (2022). TweetNLP: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*.

A Appendices

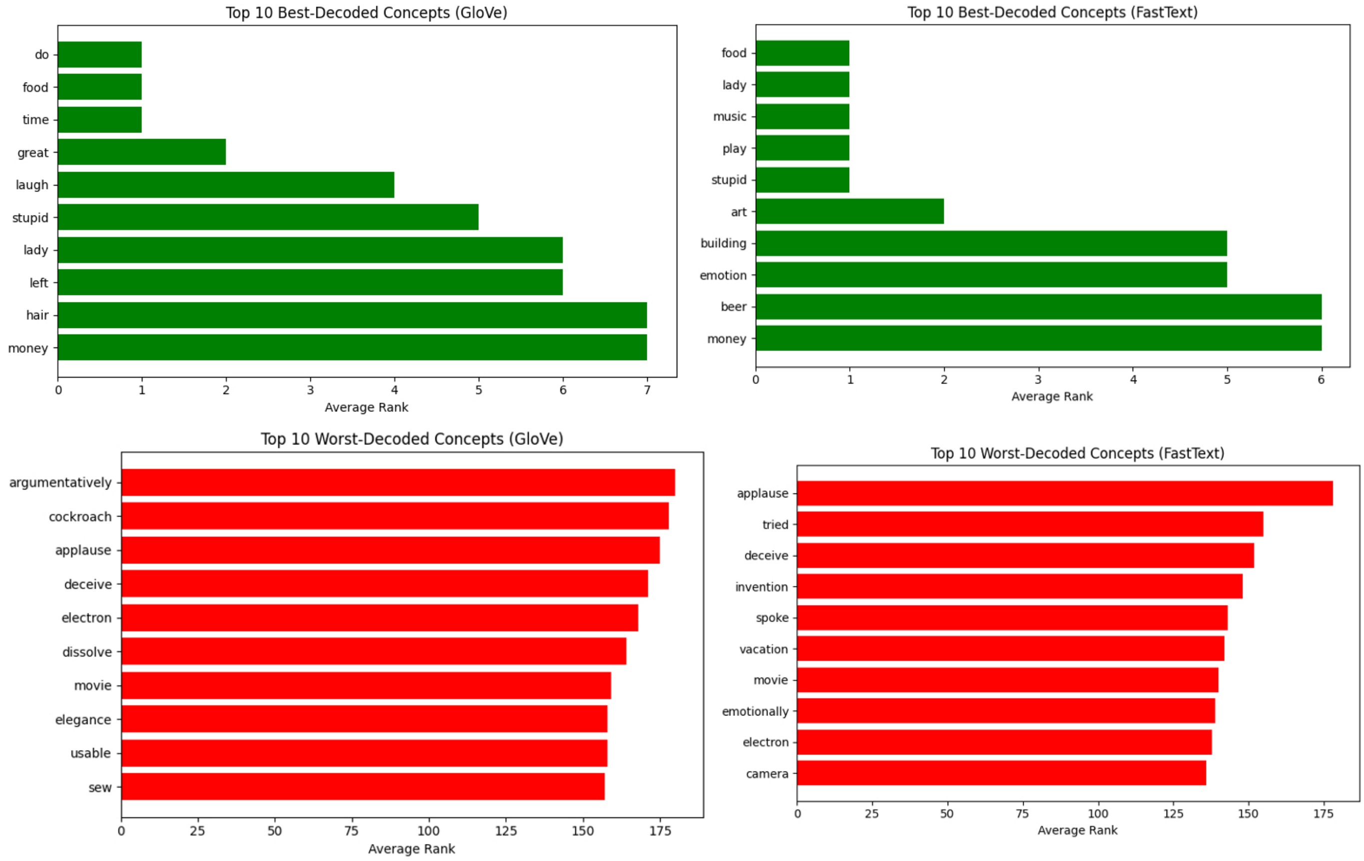


Appendix.2. Histogram comparing how many concepts fall within successive 20-rank bins (1–20, 21–40, …, 161–180) when decoded with GloVe (blue) versus FastText (orange). Both models place most concepts in the top-60 ranks, eventhough FastText is behind (by 1) in the very best bin (1–20) it shows fewer failures in the hardest bins (121–180). This pattern reinforces FastText's overall advantage in decoding accuracy while confirming that either embedding can recover a large share of concepts well above chance.



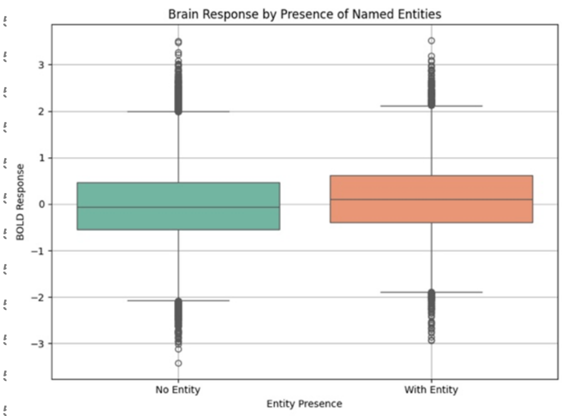
Appendix.4. Histogram of sentence-decoding ranks for GloVe (blue) and BERT (purple) on the EXP-3 dataset.

This histogram shows the number of sentences falling into each decoding rank interval for GloVe (blue) and BERT (purple) embeddings. Both models show a concentration of sentences in the lowest rank bins (1–20, 21–40), but GloVe has a slightly higher proportion in the very best bin (1–20), while BERT has more sentences in the middle rank intervals (61–100, 81–100, etc.). Notably, GloVe's decoding performance is more spread out, with a larger tail extending into the higher (worse) rank intervals (181–243), while BERT has fewer sentences in these upper bins. This pattern suggests that BERT embeddings yield more consistent decoding performance overall, with fewer extremely poorly decoded sentences compared to GloVe.



Appendix.1. Top 10 Best and Worst Decoded Concepts Using GloVe and FastText Embeddings.

The figure shows the top 10 concepts decoded with highest accuracy (lowest average ranks, green bars) and lowest accuracy (highest average ranks, red bars) when using GloVe (left column) and FastText embeddings (right column). Both embeddings decoded concepts like "money," "stupid," and "lady" very effectively. However, FastText performed slightly better overall, achieving lower ranks for its top-decoded concepts. Conversely, concepts such as "applause," "electron," and "movie" were consistently difficult to decode accurately, resulting in high ranks for both embedding methods.



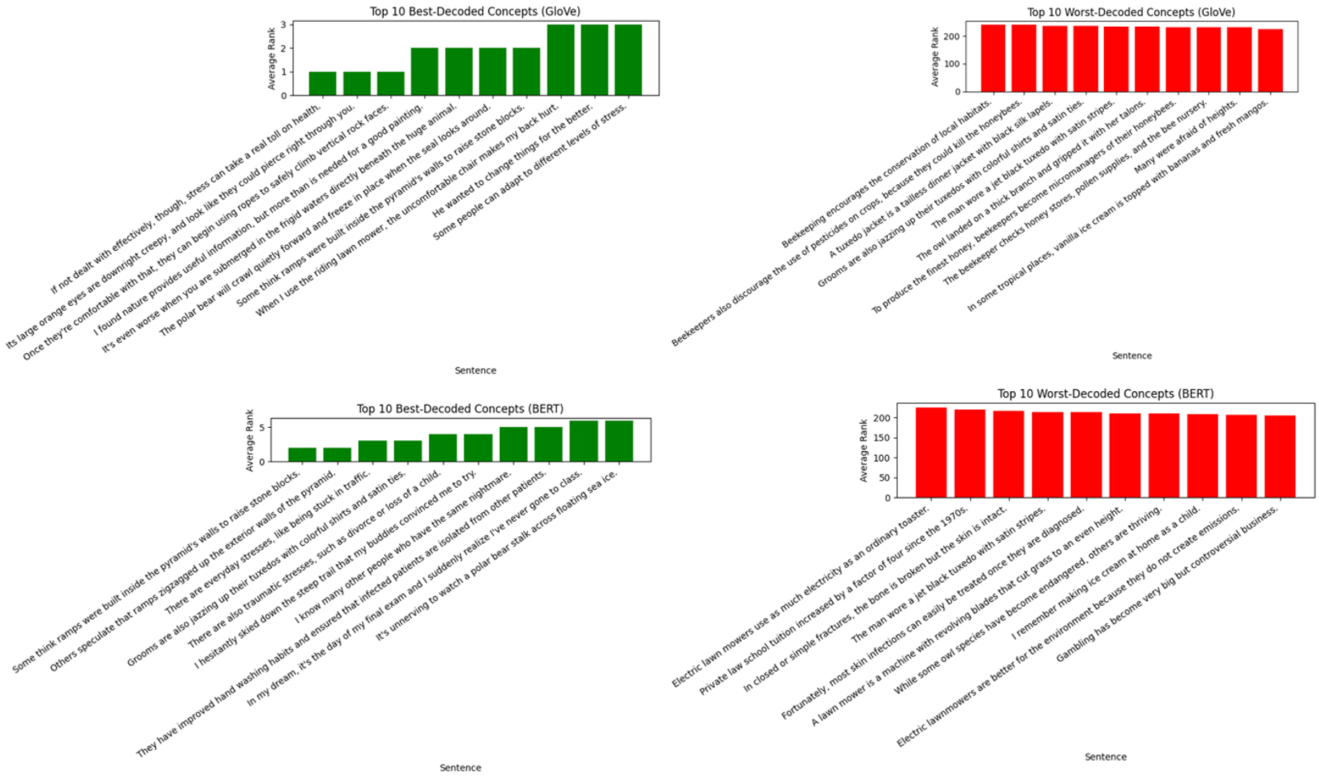
Appendix.5. Brain Response by Presence of Named Entities.

Boxplots show the distribution of language-region BOLD responses for sentences with and without named entities. The boxplot clearly shows that sentences containing named entities obtain higher brain responses on average compared to sentences without named entities. The median BOLD response is greater for the "With Entity" group, and the overall distribution is shifted upward. This difference was highly significant (t-test: T = 22.57, $p < 0.001$), supporting the conclusion that named entities robustly boost brain activity in language regions.

ROI	R^2	Adj-R^2	F - statistic	P - value	AIC	BIC
LH_AntTemp	0.007	0.002	1.357	0.173	5668	5751
LH_IFG	0.012	0.007	2.540	**0.001**	6473	6556
LH_IFGorb	0.008	0.003	1.688	0.057	6818	6901
LH_MFG	0.012	0.007	2.521	**0.001**	6732	6815
LH_PostTemp	0.012	0.008	2.611	**0.001**	5859	5942
LH_netw	0.011	0.006	2.340	**0.004**	5401	5484

Appendix.6. Regression Results for Named Entity Presence Across Left Hemisphere Language ROIs.

This table reports linear regression statistics testing whether the presence of named entities predicts BOLD responses in various left hemisphere (LH) language-related regions of interest (ROIs). Significant effects were observed in LH_IFG, LH_MFG, LH_PostTemp, and the overall LH_netw ($p < 0.01$), with modest R^2 values, indicating that entity presence explains a small but statistically reliable portion of neural variance. The LH_AntTemp and LH_IFGorb did not reach significance. Lower AIC and BIC values in LH_netw and LH_PostTemp suggest relatively better model fit for these ROIs.



Appendix.3. Top-10 best- and worst-decoded sentences under GloVe (Top) and BERT (Bottom) sentence embeddings.

Bar height shows the mean decoding rank obtained across 18-fold cross-validation. GloVe and BERT each recover a handful of sentences with ranks below 3, but BERT's easiest items are, on average, decoded with slightly smaller error. In addition, the models struggle with different informational sentences with the exception of the sentence *"The man wore a jet black tuxedo with satin stripes."* Which both models fail on.

581