Non-literal Meaning Representation in the Brain during Naturalistic Listening

Zhengwu Ma, Yuhan Huang, Chengcheng Wang, Jixing Li*
Department of Linguistics and Translation, City University of Hong Kong,
{zhengwu.ma, yuhan.huang, chengcheng.wang}@my.cityu.edu.hk,
jixingli@cityu.edu.hk

Abstract

Naturalistic language comprehension often involves interpretations that go beyond literal meaning. In continuous narratives, literal and non-literal meanings are tightly intertwined, making them difficult to distinguish computationally. Here, we combined literal sentence representations and human-annotated non-literal interpretations for model-brain alignment. Using fMRI data recorded during passive listening to the Chinese version of *The Little Prince*, we annotated sentences containing non-literal meaning with human-written interpretations of their implied meaning. We then derived the literal and non-literal representations from LLaMA3.1-8B and evaluated their correspondence with neural activity using whole-brain encoding models. Literal representations aligned strongly with left-lateralized frontotemporal regions, whereas non-literal interpretations showed broader right-hemisphere involvement. Combining the two further improved encoding performance in the bilateral temporal and dorsal frontal cortices, suggesting that naturalistic comprehension engages complementary levels of meaning.

1 Introduction

When we listen to a story, we do not simply process the words that are explicitly stated. We constantly infer "implied" messages based on context, such as metaphors, symbolic meanings, or the speaker's intentions. This suggests at least two co-existing meaning dimensions in everyday understanding: literal meaning, grounded in the surface sentence, and non-literal (inferred) meaning, reflecting what the sentence is taken to convey in context (Grice, 1975; Zwaan and Radvansky, 1998). A central question is how literal and non-literal meaning relate to patterns of brain activity during naturalistic comprehension.

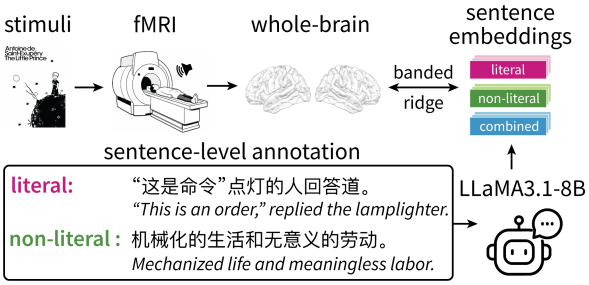

Figure 1: Aligning literal and non-literal meaning embeddings with whole-brain fMRI data during naturalistic listening in the Chinese version of *The Little Prince*.

Much of what we know about non-literal interpretation in the brain comes from neuroimaging studies that use carefully controlled contrasts, often focusing on specific phenomena such as metaphor or irony (Rapp et al., 2004; Eviatar and Just, 2006; Zempleni et al., 2007). These studies have been invaluable for isolating particular processes, but their simplified stimuli and task settings do not fully capture how meaning unfolds in the naturalistic context.

In real narratives, multiple cues co-occur, and information accumulates over time. Neural responses therefore reflect moment-to-moment processing and integration in an extended context (Hasson et al., 2008; Lerner et al., 2011). This makes it unclear how findings from controlled paradigms carry over to naturalistic story comprehension. What we still lack is a way to compare literal text and inferred interpretations for the same narrative content in a matched manner.

To make such a matched comparison possible, we need predictors that are quantifiable, time-aligned, and directly comparable across meaning dimensions. Recent large language models (LLMs) are well suited to this role: they cover a wide range of language tasks, and their representations have been shown to systematically relate to neural responses during language comprehension (Schrimpf

*Corresponding author.

et al., 2021; Goldstein et al., 2022; Caucheteux and King, 2022; Caucheteux et al., 2023; Gao et al., 2024, 2025; Kumar et al., 2024).

Here, we introduce a measurement framework for disentangling multiple meaning dimensions during naturalistic language comprehension. Using whole-brain functional magnetic resonance imaging (fMRI) recorded during continuous listening to the Chinese audiobook of *The Little Prince*, we aligned the narrative with the neural time series at the sentence level and constructed paired predictors for the same content (see Figure 1). For narratives, we used two matched texts: the original sentence and a human-annotated non-literal interpretation that summarized what the sentence is understood to convey in the story. Using LLMs, we extracted sentence-level features for both literal and non-literal sentences, and then evaluated these predictors in whole-brain encoding models. Our results reveal:

- Literal and non-literal meaning were related to both shared and dissociable cortical response patterns during continuous story comprehension, with a tendency for non-literal signals to extend more into right-hemisphere regions.
- We introduce a feasible measurement for separating language comprehension in naturalistic narratives, enabling comparisons in brain encoding analysis.
- We provide sentence-level non-literal annotations aligned to the original narrative text, enabling matched comparisons of surface content and inferred meaning in future work.

2 Related Work

2.1 Neural basis of literal and non-literal language comprehension

Language comprehension is supported by a left-lateralized fronto-temporal network, with additional bilateral involvement in speech and semantic processing (Hagoort, 2013; Hickok and Poeppel, 2007). Within this network, the left temporal and parietal areas, including the middle temporal gyrus (MTG), the superior temporal gyrus (STG) and the angular gyrus (AG), handle accessing and representing word meanings (Binder et al., 2009), whereas the inferior frontal gyrus (IFG) contributes to integration and control during sentence processing (Hagoort, 2013; Matchin and Hickok, 2020). This network is often framed as interactions between a transmodal semantic hub in the anterior temporal lobes (ATL) and fronto-temporal mechanisms that build hierarchical meaning from simpler constituents (Patterson et al., 2007; Ralph et al., 2017; Pallier et al., 2011; Lau et al., 2008). This framework serves as a general model of meaning construction, but it is less specific about how to dissociate complex context-driven meaning in naturalistic narratives.

Work on context-dependent interpretation suggests that going beyond the surface sentence can recruit broader integrative and control processes. The right hemisphere, for example, has been argued to maintain broader contextual information or diffuse semantic activation that supports remote associations (Beeman et al., 1994; Jung-Beeman, 2005). Studies of metaphors and symbolic expressions often report additional recruitment of regions in the right hemisphere, such as rIFG and rMTG, and when interpretive demands increase, while dorsal frontal areas (e.g., middle/superior frontal areas) are activated for monitoring and cognitive control (Bohrn et al., 2012; Rapp et al., 2012). Interpretation related to social intent or affect can also involve salience-related regions such as the anterior insula and the anterior cingulate cortex (Citron and Goldberg, 2014; Seeley et al., 2007). However, much of this evidence comes from controlled contrasts that isolate specific phenomena (e.g., Rapp et al., 2004; Eviatar and Just, 2006; Zempleni et al., 2007), leaving open how literal content and non-literal interpretations co-exist, and whether they can be disentangled during continuous naturalistic narrative comprehension.

2.2 Model-derived language features for naturalistic encoding

Naturalistic encoding work has shown that semantic features derived from a text-based representational space can be linked to distributed cortical responses during continuous story listening (Huth et al., 2016). This framework is widely used to map semantic organization onto the brain cortex under naturalistic stimulation, and for our question it is crucial to use sentence-level features that capture meaning in context. Modern transformer-based language models provide a practical shared representational space for this purpose because they generate contextualized representations that can be extracted at the sentence-level or even the discourse-level in a uniform way (Vaswani et al., 2017; Devlin et al.,

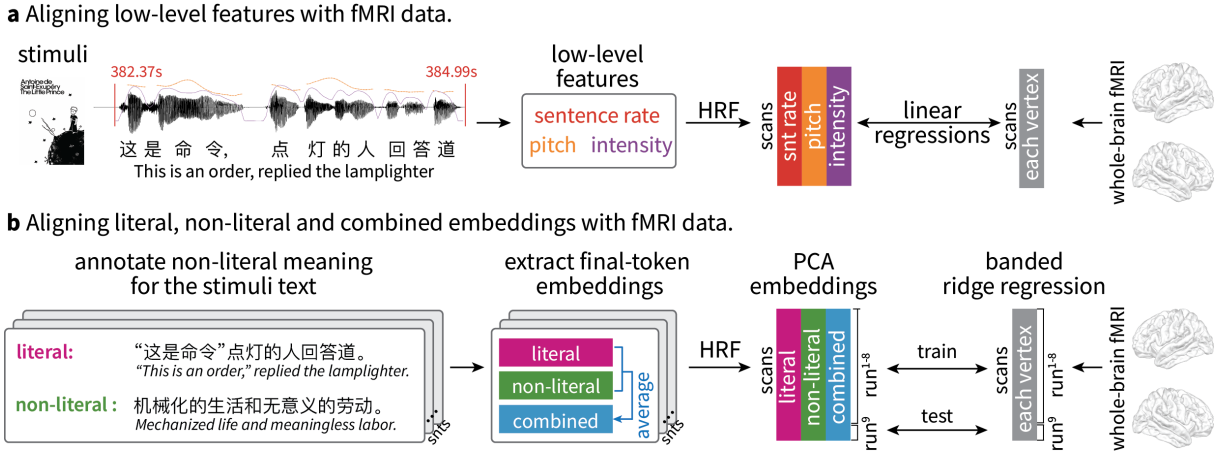

Figure 2: Methods overview. **a** Aligning the low-level regressors, sentence rate, pitch, and intensity with whole-brain fMRI data using GLMs. **b** Aligning literal, non-literal and combined sentence representations with whole-brain fMRI data using banded ridge regression.

2019; Radford et al., 2018). Recent model-derived features of LLMs have shown robust correspondence with neural responses during naturalistic language processing (Caucheteux et al., 2023; Kumar et al., 2024; Goldstein et al., 2024).

LLM-derived representations are sensitive to the amount of input context. Recent work has explicitly manipulated the amount of preceding context and shown that alignment can change substantially with the window size of the context (Yu et al., 2024). Long-context inputs have also been used to better capture complex information integration (Goldstein et al., 2025a,b). Here, in this work, we intend to apply a matched sentence-level comparison across meaning dimensions. We extracted sentence-locked representations by feeding each sentence on its own, without adding extra discourse context, to keep the comparison clean and interpretable. We then contrasted features derived from the narrative's surface sentence with features derived from explicit interpretation annotations. This setup motivated our paired-predictor design, enabling comparable tests of literal versus non-literal meaning during naturalistic comprehension.

3 Methods

3.1 Non-literal meaning annotations

We adopted the classic storybook *The Little Prince* as our experimental text because its language is concise, while its intended meanings are often rich and implicit. The Chinese version of *The Little Prince* audiobook contains 15,603 words distributed across 1,289 sentences (mean sentence length = 12.10 words, SD = 9.37). We operationalized *non-literal meaning* as sentences whose intended interpretation involves metaphorical, symbolic, inferential, or otherwise implied meanings that extend beyond their surface lexical-semantic composition.

Due to the lack of an established dataset with non-literal meaning annotations, we recruited two native Mandarin speakers with formal training in linguistics to annotate all sentences for the presence or absence of non-literal meaning. For sentences identified as non-literal, annotators additionally provided a written paraphrase of the intended non-literal interpretation. Inter-annotator disagreements were resolved through discussion until consensus was reached, ensuring annotation reliability. Detailed annotation instructions are provided in Section 6.

A total of 429 sentences (33.3% of the total sentences; mean sentence length = 13.67 words, SD = 10.25) were labeled as containing non-literal meaning. For the following analyses, we only focused on these 429 sentence pairs which both contain literal and non-literal meanings. Representative examples of annotated literal and non-literal contrasts are shown in Table 1.

3.2 fMRI data

We analyzed the Chinese subset of a publicly available naturalistic listening fMRI dataset (Li et al., 2022), including 35 native Mandarin speakers (15 females; age = 19.3 $\pm$ 1.6 years). During scanning, participants passively listened to the complete Chinese audiobook of *The Little Prince* (approximately 99 minutes) in a single session divided into

ID	Literal Meaning	Non-literal Meaning
1	“这是命令,” 点灯的人回答道。 “This is an order,” replied the lamplighter.	机械化的生活和无意义的劳动。 Mechanized life and meaningless labor.
2	他的那朵花曾对他说，她是整个宇宙中独一无二的一种花。 His flower had once told him that she was a flower unique in all the universe.	小王子怀念爱情，感到心疼后悔。 The little prince misses his love, feeling heartache and regret.
3	当我们默默地走了好几个小时以后，天黑了下来，星星开始发出光亮。 When we had trudged along for several hours, in silence, the darkness fell, and the stars began to come out.	星星的亮起提醒人们即使在最艰难的情况下也总有希望存在。 The lighting up of the stars reminds people that even in the most difficult situations, there is always hope.

Table 1: Examples of literal and non-literal meaning annotations from the Chinese version of *The Little Prince*.

nine runs of approximately 10 minutes each. The preprocessed volumetric data were reconstructed onto a “fsaverage4” surface-based template. The fMRI signals are z-scored across the time dimension for each participant, surface vertex and session independently. Details for fMRI data acquisition and preprocessing procedure are provided in Appendix A.

3.3 Model

To quantify sentence-level representational features of literal and non-literal meanings, we extracted final-token embeddings from the open-source large language model LLaMA3.1 (Grattafiori et al., 2024) for both the original sentences and their annotated interpretations. LLaMA3.1 has been widely used in model-brain alignment research (e.g., Binz et al., 2025; Gao et al., 2025; An et al., 2025) and offers strong multilingual capabilities. To manage computational resources (see Appendix B), we used the 8B-size model.

3.4 Representation differences between literal and non-literal meanings

Linear probing from literal embeddings to non-literal embeddings. We first tested whether non-literal embeddings can be recovered from literal embeddings via a simple linear mapping. For each model layer, we trained a ridge regression model to predict the non-literal embedding vector from the corresponding literal embedding vector. Prediction performance was evaluated with 5-fold cross-validated coefficient of determination (R^2). As a permutation control, we randomly shuffled the pairing between literal and non-literal embeddings and repeated the same layer-wise probing procedure.

Dimensionality reduction of literal and non-literal representations. We applied principal component analysis (PCA) to the sentence-level embeddings for the literal and non-literal conditions. For each sentence, we first averaged the model hidden states across layers to obtain a single 4096-dimensional vector, and then projected these vectors onto the first three principal components. We visualized the resulting 3D projections to examine the global geometry and separability of literal vs. non-literal representations.

3.5 Aligning low-level features with whole-brain fMRI data

We included two acoustic features and one sentence-level regressor: **pitch**, **intensity**, and **sentence rate**, which are known to correlate with activity in core language-related regions and served as baseline predictors against which the contribution of sentence-level semantic representation was evaluated (Momenian et al., 2024; Wang et al., 2025; Li et al., 2022). Pitch and root-mean-square (RMS) intensity were computed from the Chinese audiobook at 10 ms resolution using the Voicebox toolbox[1]. Sentence rate was modeled as a binary regressor marking the offset of each sentence in the audiobook.

For each run, these low-level regressors were convolved with the canonical hemodynamic response function (HRF). The run-wise convolved regressors were then concatenated across runs, z-scored, and fit to each subject’s concatenated nine-run fMRI time series at each cortical vertex using ordinary least squares (OLS), as illustrated in Figure 2a.

[1] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

3.6 Aligning literal and non-literal embeddings with whole-brain fMRI data

Identify the best-performing layer. We performed layer-wise encoding analyses at each cortical vertex using the extracted literal and non-literal embeddings. For each layer and condition (literal, non-literal), we constructed a time series of sentence-level embeddings, time-locked to each annotated sentence's offset. This embedding time series was then convolved with the HRF to match the temporal resolution of the fMRI data and used as predictors in vertex-wise ridge regression models. Models were trained on embeddings and fMRI responses concatenated across the first eight runs and evaluated on the held-out ninth run. Encoding performance was quantified as the Pearson correlation (r) between the predicted and observed fMRI time series in the test run. To identify the best-performing layer, we averaged r across vertices to obtain a single whole-brain encoding score for each layer.

Banded ridge regression. We next used banded ridge (multi-kernel) regression (Dupré la Tour et al., 2022) to jointly model fMRI responses using literal and non-literal representations. We selected embeddings from the best-performing LLaMA3.1 layers identified in the layer-wise encoding analysis and fit multi-kernel ridge models with Himalaya's MULTIPLEKERNELRIDGECV. Based on these selections, we constructed three predictor sets: (i) literal sentence embeddings, (ii) non-literal sentence embeddings, and (iii) a *combined* embedding defined as the average of the best-layer literal and non-literal embeddings. Each predictor set was reduced from 4096 to 100 dimensions using PCA, and run-wise predictors were concatenated across the first eight runs to form three separate linear kernels.

The regression models were trained in the first eight concatenated runs and evaluated in the ninth run. For each vertex and participant, kernel-specific regularization weights were selected via grid search with nested cross-validation over 10 values log-spaced from 10^0 to 10^{20} within the training data (Gao et al., 2025; Goldstein et al., 2025b; Huth et al., 2016). Kernel-specific performance was obtained by splitting each kernel's prediction and computing the Pearson correlation between that kernel's predicted time series and the observed vertex-wise fMRI time series in the test run. (see Figure 2b and Appendix C for regression details).

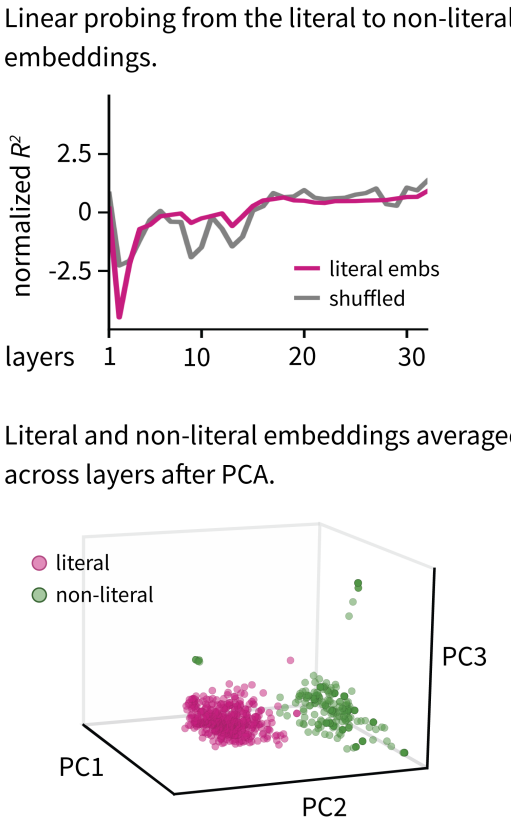

Figure 3: a Layer-wise linear probing from literal to non-literal embeddings using ridge regression. **b** PCA visualization of sentence-level embeddings after averaging across layers. Each point is a sentence projected onto the first three principal components.

3.7 Statistical significance testing

Linear probe statistical testing. We first averaged the 5-fold R^2 scores for the literal→non-literal mapping and for the shuffled-pairing control. We then compared probing performance between the two conditions using a two-tailed paired-sample t-test.

GLM and banded ridge regressions. At the group level, vertex-wise GLM regression coefficients (β) and banded ridge encoding performance (Pearson's r) were first z-scored and evaluated with one-sample, one-tailed cluster-based permutation tests (Maris and Oostenveld, 2007) with 10,000 permutations. Clusters were formed from statistics corresponding to a p-value less than 0.05, and only clusters spanning a minimum of 20 vertices were included in the analysis.

Contrasts between literal and non-literal encoding maps. To compare the encoding performance between conditions, we computed vertex-wise contrast maps by subtracting the non-literal embedding $r$map from the corresponding literal

embedding *r*map (literal > non-literal). The reverse contrast (non-literal > literal) was computed conversely. The resulting contrast maps were z-scored and evaluated using the same cluster-based permutation procedure described above. All statistical analyses were performed using custom Python codes, making heavy use of the Scipy (v1.12.0; Virtanen et al., 2020), MNE (v1.10.2; Gramfort et al., 2014) and Eelbrain (v0.41; Brodbeck et al., 2023) packages.

4 Results

4.1 Representational differences between literal and non-literal embeddings

Linear probing from literal to non-literal embeddings. We found no significant difference in probing performance between the true literal→non-literal mapping and the shuffled control. Across layers, ridge-based linear probes from literal to non-literal embeddings did not outperform the shuffled baseline (Figure 3a), indicating that non-literal embeddings are not reliably predictable from literal embeddings using a simple linear mapping in this embedding space.

Dimensionality reduction of literal and non-literal embeddings. PCA visualization of two embeddings showed that literal and non-literal representations form separable clusters in the low-dimensional space (Figure 3b), suggesting systematic differences in their global geometry. Together with the null linear-probing result, this pattern indicates that the two representations are separable in embedding space, while their relationship is not well captured by a single linear mapping.

4.2 Encoding performance across LLaMA3.1 model layers

The best-performing encoding layers for literal and non-literal embeddings were layer 11 and layer 21, respectively (see Figure 4a). This pattern is consistent with prior model-brain alignment work showing that mid-to-late layers often yield the strongest encoding performance.

Correlations among the literal and non-literal sentence embeddings from their best-performing layers, as well as the combined embedding defined as their average, are shown in Figure 4b. Literal and non-literal embeddings exhibited low similarity (r=0.081), consistent with the representational differences above. The combined embedding was strongly correlated with the non-literal embedding

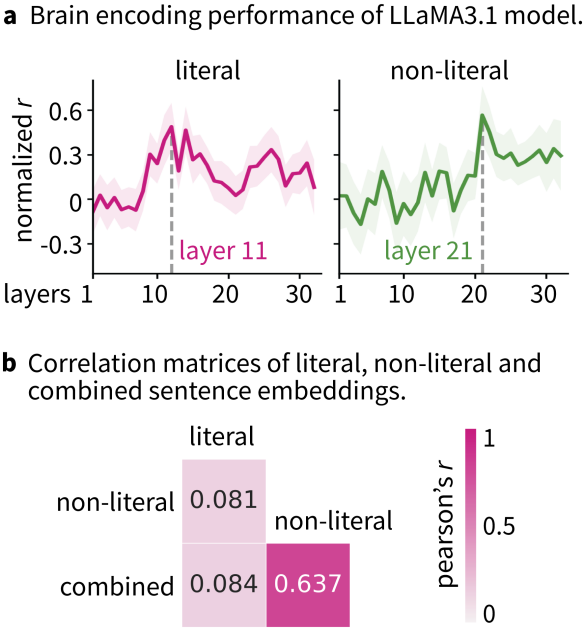

Figure 4: **a** Layer-wise encoding performance of LLaMA3.1 against fMRI data using literal and non-literal embeddings. **b** Correlation matrices of literal, non-literal, and combined embeddings from their best-performing LLaMA3.1 layers.

(r=0.637) but only weakly correlated with the literal embedding (r=0.084).

4.3 Brain regions associated with low-level features

Correlations among the HRF-convolved sentence rate, pitch, and intensity regressors are shown in Figure 5a. The strongest association was between pitch and intensity (r=0.433), whereas correlations involving two acoustic features and sentence rate were close to zero (both r<0.02).

We observed significant bilateral effects of the low-level features in temporal and frontal regions, consistent with prior work on speech-related processing (Li et al., 2022; Momenian et al., 2024; Wang et al., 2025). The acoustic regressors (pitch and intensity) showed robust effects in the temporal cortex: pitch exhibited a significant association in the bilateral ATL and the MTG, with a stronger effect in the right hemisphere, while intensity was primarily associated with the bilateral STG. Sentence rate was associated with responses in left ATL, left IFG, bilateral STG, MTG, and ventromedial prefrontal cortex (vmPFC). Detailed clusters and corresponding statistics for the low-level features are reported in Figure 5b and Table 2.

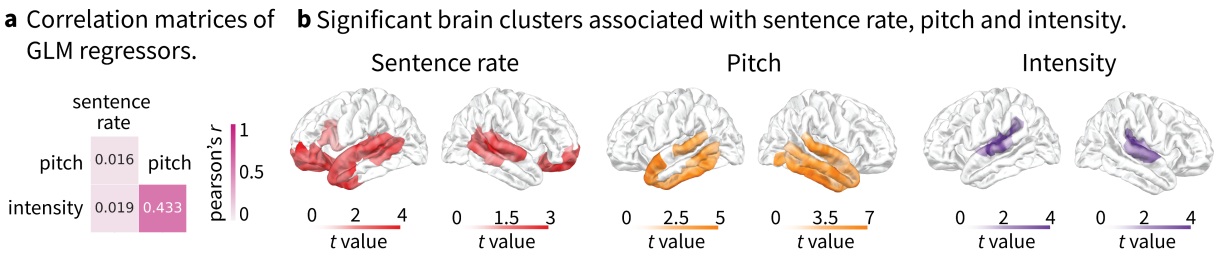

Figure 5: **a** Correlation matrices among three low-level features. **b** Significant brain clusters associated with sentence rate, pitch and intensity.

Embedding group	Left hemisphere			Right hemisphere		
	N vertices	p	Cohen's d	N vertices	p	Cohen's d
Sentence rate	758	0.0061	3.7642	722	0.0199	3.1042
Pitch	257	0	2.4034	452	0	2.3383
Intensity	249	0.0189	2.4325	222	0.0374	3.1270

Table 2: Cluster statistics for sentence rate, pitch, and intensity effects from the GLM analysis.

4.4 Brain regions associated with literal, non-literal and combined embeddings

We observed widespread frontal and temporal clusters across the literal, non-literal, and combined conditions, with lateralization patterns that differed by condition (Figure 6a).

In the literal condition, left-lateralized effects spanned canonical language regions from frontal to temporoparietal areas, including the IFG, insula, STG/MTG, and AG. Smaller right-hemisphere clusters were also detected in the frontal/insular cortices and temporal regions, including IFG, insula, STG/MTG.

The non-literal condition showed a more right-lateralized pattern. Only a small cluster was observed in the left insula, whereas the right hemisphere exhibited widespread effects across frontal and temporal regions, including the precentral gyrus (PrG), caudal middle frontal gyrus (cMFG), IFG, and STG.

In the combined embedding, the effects were more prominent in the right hemisphere overall. Left-hemisphere effects were observed mainly in superior frontal gyrus (SFG), STG/MTG. In the right hemisphere, the effects extended across frontal, insular, and temporal cortices, including SFG, insula, STG/MTG, and the temporal pole. Table 3 reports detailed literal, non-literal and combined cluster statistics.

4.5 Brain regions associated with contrasts between literal and non-literal embeddings

Contrast analyses also revealed lateralized differences between conditions: the literal > non-literal contrast was strongest in the left frontal and temporal regions, whereas the non-literal > literal contrast was strongest in the right frontal regions (see Figure 6b). Specifically, the literal > non-literal contrast yielded significant clusters in the left IFG and STG, while the non-literal > literal contrast yielded marginally significant clusters confined to the right frontal cortex, with peaks in the precentral gyrus, SFG, and cMFG. Detailed statistics are reported in Table 3.

5 Discussion

A large body of model–brain alignment work shows that contextual embeddings derived from large language models predict neural activity during naturalistic language comprehension (e.g., Gao et al., 2024, 2025; Goldstein et al., 2022; Huth et al., 2016; Toneva et al., 2022; Schrimpf et al., 2021). Most prior studies, however, use embeddings derived directly from surface text and therefore primarily index literal meaning. Here, we offered a measurement framework by aligning fMRI responses to embeddings derived from manually literal and annotated non-literal interpretations of the same narrative content. We found literal and non-literal representations derived from LLM were clearly distinct in embedding space, and non-literal representations could not be reliably recovered from literal embeddings with a simple linear probe.

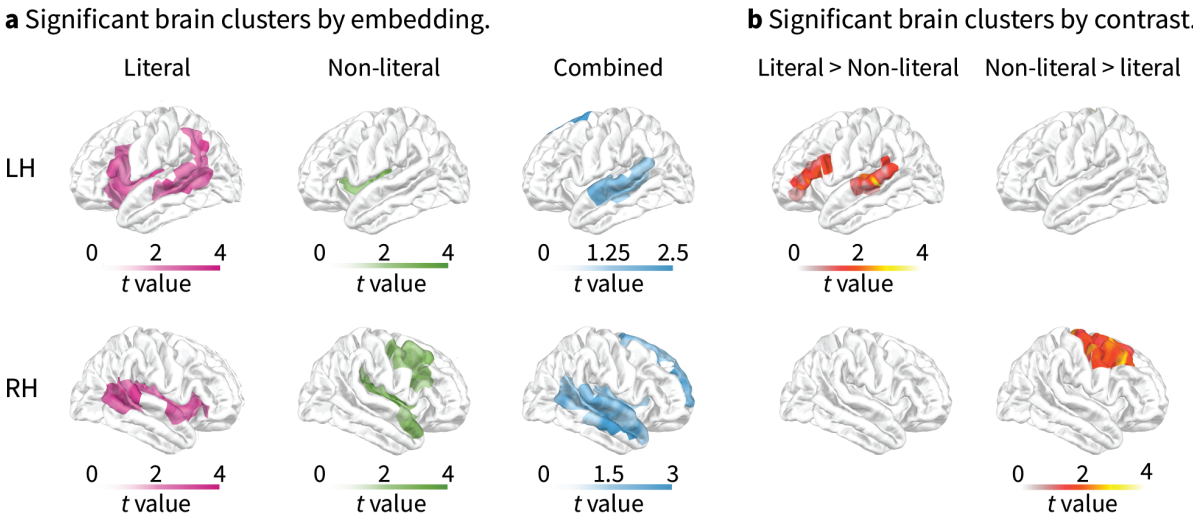

Figure 6: **a** Significant brain clusters associated with literal, non-literal and combined embeddings. **b** Significant brain clusters for literal > non-literal and non-literal > literal contrasts.

Embedding group	Left hemisphere			Right hemisphere		
	N vertices	p	Cohen's d	N vertices	p	Cohen's d
Literal	123	0.001	6.2609	112	0.0007	6.5042
Non-literal	30	0.0091	5.2102	129	0.0056	3.8235
Combined	55	0.0145	3.8592	212	0.0052	4.3092
Contrast						
Literal > Non-literal	43	0.002	2.5873	/	/	/
Non-literal > Literal	/	/	/	76	0.0631	3.8239

Table 3: Cluster statistics for literal, non-literal, and combined sentence embeddings from banded ridge regression.

In the brain, both literal and non-literal embeddings significantly predicted widespread responses, but with condition-dependent spatial patterns. Literal embeddings aligned most strongly with left-lateralized temporal and frontal regions typically implicated in core semantic processing, consistent with prior model-brain alignment findings. Whereas non-literal embeddings showed broader right-hemisphere involvement in frontal and temporal cortices, consistent with accounts linking the right hemisphere to integrative and interpretive aspects of meaning (Jung-Beeman, 2005; Binder et al., 2009; Rapp et al., 2004; Fedorenko et al., 2024). Importantly, combining the best-layer literal and non-literal embeddings improved encoding performance and produced bilateral effects, suggesting that neural responses during narrative comprehension reflect multiple levels of meaning in parallel rather than a single surface-based semantic representation.

It is also worth noting that standard LLM training objectives are not designed to explicitly separate surface-form semantics from interpretive or non-literal dimensions of meaning. Solely relying on literal-derived embeddings may miss aspects of meaning that are critical for naturalistic comprehension.

6 Conclusion

We investigated whether model-derived representations of non-literal meaning align with neural activity during naturalistic narrative comprehension. Embeddings derived from manually annotated non-literal interpretations significantly predicted brain responses, and combining literal and non-literal representations improved encoding performance in bilateral temporal and frontal regions, suggesting that comprehension is sensitive to both compositional and interpretive information.

However, LLM-derived embeddings from the surface form alone did not fully reflect non-literal meaning: literal embeddings did not reliably predict non-literal embeddings with a simple linear mapping, and the two representations were separable in embedding space. These findings motivate incorporating explicit interpretive semantic dimensions in model-brain alignment to better capture naturalistic language understanding.

Limitations

Several limitations should be acknowledged. First, our non-literal embeddings were derived from manually annotated interpretations. While this ensures interpretive specificity, non-literal meaning is inherently context-dependent and may vary across individuals. Future work could develop computational approaches to model interpretive meaning at scale and assess inter-annotator and inter-reader variability more directly.

Second, our definition of *non-literal meaning* pooled multiple types of implied meaning (e.g., metaphorical, symbolic, and inferential readings) that may recruit partially distinct neural mechanisms. Current analyses do not dissociate these subtypes, and future studies should test whether categories such as metaphor, irony, and implicature show distinct alignment patterns.

Finally, we analyzed only the Chinese subset of a multilingual naturalistic fMRI dataset (Li et al., 2022). This controlled the annotation pipeline, but limited claims about cross-linguistic generalizability. Given that non-literal meaning is shaped by language structure and cultural conventions, future work should leverage the English and French data to examine shared versus language-specific neural signatures of interpretive meaning.

Ethics Statement

The authors declare no competing interests. The non-literal meanings were manually annotated by two trained annotators following the instructions to "identify and specify the non-literal interpretation of each sentence whenever applicable." Annotators were recruited through advertisements and were compensated for their work in accordance with local rates; they reported no conflicts of interest. The annotation task involved no risks.

The fMRI dataset used in the analysis is publicly available and does not contain sensitive content, such as personal information. The adaptation and use of the fMRI dataset were conducted in accordance with its license.

The model states of LLaMA3.1-8B are utilized solely for research purposes, aligning with its intended use.

References

Jingmin An, Yilong Song, Ruolin Yang, Nai Ding, Lingxi Lu, Yuxuan Wang, Wei Wang, Chu Zhuang, Qian Wang, and Fang Fang. 2025. Hierarchical frequency tagging probe (HFTP): a unified approach to investigate syntactic structure representations in large language models and the human brain. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mark Beeman, Rhonda B. Friedman, Jordan Grafman, Enrique Perez, Sherri Diamond, and Miriam Beadle Lindsay. 1994. Summation Priming and Coarse Semantic Coding in the Right Hemisphere. *Journal of Cognitive Neuroscience*, 6(1):26–45.

Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. 2025. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009.

Isabel C. Bohrn, Ulrike Altmann, and Arthur M. Jacobs. 2012. Looking at the brains behind figurative language—A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. *Neuropsychologia*, 50(11):2669–2683.

Christian Brodbeck, Proloy Das, Marlies Gillis, Joshua P Kulasingham, Shohini Bhattasali, Phoebe Gaston, Philip Resnik, and Jonathan Z Simon. 2023. Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions. *eLife*, 12:e85012.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Francesca M. M. Citron and Adele E. Goldberg. 2014. Metaphorical Sentences Are More Emotionally Engaging than Their Literal Counterparts. *Journal of Cognitive Neuroscience*, 26(11):2585–2595.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. 2022. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728.

Oscar Esteban, Rastko Ciric, Karolina Finc, Ross W. Blair, Christopher J. Markiewicz, Craig A. Moodie, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Daniel E. P. Gomez, Zhifang Ye, Taylor Salo, Romain Valabregue, Inge K. Amlien, Franziskus Liem, Nir Jacoby, Hrvoje Stojić, Matthew Cieslak, Sebastian Urchs, Yaroslav O. Halchenko, Satrajit S. Ghosh, Alejandro De La Vega, Tal Yarkoni, Jessey Wright, William H. Thompson, Russell A. Poldrack, and Krzysztof J. Gorgolewski. 2020. Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nature Protocols*, 15(7):2186–2202.

Zohar Eviatar and Marcel Adam Just. 2006. Brain correlates of discourse processing: An fMRI investigation of irony and conventional metaphor comprehension. *Neuropsychologia*, 44(12):2348–2359.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. 2024. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312.

Changjiang Gao, Jixing Li, Jiajun Chen, and Shujian Huang. 2024. Measuring meaning composition in the human brain with composition scores from large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11295–11308. Association for Computational Linguistics.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. 2025. Increasing alignment of large language models with language processing in the human brain. *Nature Computational Science*, 5(11):1080–1090.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A. Nastase, Zaid Zada, Eric Ham, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, Werner Doyle, Sasha Devore, Patricia Dugan, Roi Reichart, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Orrin Devinsky, Adeen Flinker, and Uri Hasson. 2024. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15(1):2768.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, Omer Levy, Roi Reichart, and Uri Hasson. 2025a. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16(1):10529.

Ariel Goldstein, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel A. Nastase, Harshvardhan Gazula, Aditi Singh, Aditi Rao, Gina Choe, Catherine Kim, Werner Doyle, Daniel Friedman, Sasha Devore, Patricia Dugan, Avinatan Hassidim, Michael Brenner, Yossi Matias, Orrin Devinsky, Adeen Flinker, and Uri Hasson. 2025b. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, 9(5):1041–1055.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S. Hämäläinen. 2014. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan

Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin,

Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 herd of models. *Preprint*, arxiv:2407.21783 [cs].

H. P. Grice. 1975. Logic and Conversation. In *Speech Acts*, pages 41–58. Brill.

Peter Hagoort. 2013. MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology*, 4.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. 2008. A Hierarchy of Temporal Receptive Windows in Human Cortex. *Journal of Neuroscience*, 28(10):2539–2550.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

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

Mark Jung-Beeman. 2005. Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, 9(11):512–518.

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

Ellen F. Lau, Colin Phillips, and David Poeppel. 2008. A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, 9(12):920–933.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. 2011. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, 31(8):2906–2915.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022. Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, 9(1):530.

Eric Maris and Robert Oostenveld. 2007. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–190.

William Matchin and Gregory Hickok. 2020. The Cortical Organization of Syntax. *Cerebral Cortex*, 30(3):1481–1498.

Mohammad Momenian, Zhengwu Ma, Shuyi Wu, Chengcheng Wang, Jonathan Brennan, John Hale, Lars Meyer, and Jixing Li. 2024. Le petit prince hong kong (lpphk): Naturalistic fmri and eeg data from older cantonese speakers. *Scientific data*, 11(1):992.

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

Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12):976–987.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Matthew A. Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers. 2017. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55.

Alexander M Rapp, Dirk T Leube, Michael Erb, Wolfgang Grodd, and Tilo T. J Kircher. 2004. Neural correlates of metaphor processing. *Cognitive Brain Research*, 20(3):395–402.

Alexander M. Rapp, Dorothee E. Mutschler, and Michael Erb. 2012. Where in the brain is nonliteral language? A coordinate-based meta-analysis of functional magnetic resonance imaging studies. *NeuroImage*, 63(1):600–610.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

William W. Seeley, Vinod Menon, Alan F. Schatzberg, Jennifer Keller, Gary H. Glover, Heather Kenna, Allan L. Reiss, and Michael D. Greicius. 2007. Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9).

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272.

Qixuan Wang, Qian Zhou, Zhengwu Ma, Nan Wang, Tianyu Zhang, Yaoyao Fu, and Jixing Li. 2025. Le petit prince (lpp) multi-talker: Naturalistic 7 t fmri and eeg dataset. *Scientific data*, 12(1):829.

Shaoyun Yu, Chanyuan Gu, Kexin Huang, and Ping Li. 2024. Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension. *Science Advances*, 10(21):eadn7744.

Monika-Zita Zempleni, Remco Renken, John C. J. Hoeks, Johannes M. Hoogduin, and Laurie A. Stowe. 2007. Semantic ambiguity processing in sentence context: Evidence from event-related fMRI. *NeuroImage*, 34(3):1270–1279.

Rolf A. Zwaan and Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.

Appendix

A fMRI data acquisition and preprocessing

MRI data were acquired with a 3 T GE Discovery MR750 scanner using a 32-channel head coil. Structural scans were acquired using a T1-weighted magnetization-prepared rapid gradient-echo sequence. Functional scans were obtained using a multi-echo echo-planar imaging (EPI) sequence (TR = 2000 ms; TEs = $[12.8, 27.5, 43]$ ms; flip angle = 77°; matrix size = 72×72; FoV = 240.0×240.0 mm; in-plane acceleration factor = 2; 33 axial slices; voxel size = $3.75 \times 3.75 \times 3.8$ mm). Each scanning run began with a trigger followed by an 8 s delay before stimulus onset. Anatomical and functional MRI data were preprocessed with fMRIPrep (v25.0.0; Esteban et al., 2020) with all default parameters, with final resampling to the "fsaverage4" surface performed in a single interpolation step using the `mri_vol2surf` function.

B Computational resources

All experiments were conducted on a high-performance computing (HPC) cluster with nodes equipped with dual AMD EPYC 7742 processors (64 cores per socket; 128 physical cores per node) and 512 GB of RAM. For each participant, the GLM required approximately 0.5 CPU-hours, and the banded ridge regression required approximately 5 CPU-hours.

C Banded ridge methods

In the banded ridge regression, fMRI responses y were predicted as

$$\hat{y} = \sum_i K_i w_i \tag{1}$$

where $K_i = X_i X_i^\top$ and w_i are the corresponding kernel weights. Model fitting minimized the objective

$$\|y - \sum_i K_i w_i\|^2 + \sum_i \alpha_i \, w_i^\top K_i w_i \tag{2}$$

with independent ridge penalties α_i for each kernel. We used the `precomputed` kernel option and performed random search over $\alpha_i \in [10^0, 10^{20}]$, optimizing log-weights $\delta_i = -\log \alpha_i$ via cross-validation. Data were split into the first eight runs for training and the final run for testing. Per-kernel predictions $\hat{y}_i$ were obtained using `predict(split=True)`, and Pearson correlations between predicted and observed responses were computed as model-specific performance scores.