Neurobiology of Language

an open access journal

SPECIAL ISSUE:
Cognitive Computational Neuroscience of Language

Computational Language Modeling and the Promise of In Silico Experimentation

Shailee Jain[1], Vy A. Vo[3], Leila Wehbe[4,5], and Alexander G. Huth[1,2]

[1]Department of Computer Science, University of Texas at Austin, Austin, TX, USA
[2]Department of Neuroscience, University of Texas at Austin, Austin, TX, USA
[3]Brain-Inspired Computing Lab, Intel Labs, Hillsboro, OR, USA
[4]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[5]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Keywords: computational neuroscience, deep learning, encoding models, experimental design, natural language processing, naturalistic stimuli

ABSTRACT

Language neuroscience currently relies on two major experimental paradigms: controlled experiments using carefully hand-designed stimuli, and natural stimulus experiments. These approaches have complementary advantages which allow them to address distinct aspects of the neurobiology of language, but each approach also comes with drawbacks. Here we discuss a third paradigm—in silico experimentation using deep learning-based encoding models—that has been enabled by recent advances in cognitive computational neuroscience. This paradigm promises to combine the interpretability of controlled experiments with the generalizability and broad scope of natural stimulus experiments. We show four examples of simulating language neuroscience experiments in silico and then discuss both the advantages and caveats of this approach.

Citation: Jain, S., Vo, Vy A., Wehbe, L., & Huth, A. G. (2024). Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, *5*(1), 80–106. https://doi.org/10.1162/nol_a_00101

DOI:
https://doi.org/10.1162/nol_a_00101

Received: 28 February 2022
Accepted: 18 January 2023

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Alexander G. Huth
huth@cs.utexas.edu

Handling Editor:
Alessandro Lopopolo

The MIT Press

INTRODUCTION

One major goal of language neuroscience is to characterize the function of different brain regions and networks that are engaged in language processing. A large body of work has investigated different aspects of language processing—such as semantic knowledge representation (Binder et al., 2009; Huth et al., 2016; Mitchell et al., 2008), syntactic processing (Friederici et al., 2000), and phonological mapping (Chang et al., 2010)—and characterized the properties of the language network like the processing timescale (Lerner et al., 2011), convergence with different sensory systems (Popham et al., 2021), role in bilingual representations (Chan et al., 2008), and more. To study these questions, language neuroscientists have developed a suite of experimental designs, ranging from highly specific controlled experiments to natural stimulus experiments and, more recently, deep learning-based approaches for computational modeling.

Each experimental design can be thought of as an investigative tool for understanding the brain's response $R_v = f_v(S)$, where f_v is the function that some brain element v (e.g., a single neuron, voxel, brain area, or magnetoencephalography [MEG] sensor) computes over a given language stimulus S to produce responses R_v. Some experimental designs—like contrast-based studies—aim to directly compare certain aspect of f_v, such as the response to different word categories. Others—like experiments with complex stimuli that are paired with encoding

models—approximate f_v using computational tools, and this allows for the prediction of activity related to new stimuli. In this paper we describe an alternative to existing paradigms: in silico controlled experimentation using computational models of naturalistic language processing. This hybrid approach combines the strengths of controlled and naturalistic paradigms to achieve high ecological generalizability, high experimental efficiency and reusability, high interpretability, and sensitivity to individual participant effects.

In silico experimentation: Simulating experiments by predicting brain responses with a computational model to test generalizability and do efficient hypothesis testing.

We first compare and contrast experimental designs based on their effectiveness and efficiency for revealing f_v. Then we introduce the in silico experimentation paradigm with deep learning models. We discuss four different neuroimaging studies that use this paradigm to investigate different linguistic phenomena in the brain. And finally, we discuss the potential of this approach to alleviate the problems of reproducibility in language neuroscience, as well as caveats and pitfalls of in silico experimentation.

EXPERIMENTAL DESIGNS IN LANGUAGE NEUROSCIENCE

Controlled Experimental Design: Contrast-Based Studies

Language is a rich and complex modality that humans are uniquely specialized to process. Given this complexity, neuroscientists have traditionally broken language down into specific processes and properties and then designed controlled experiments to test each separately (Binder et al., 2009; Friederici et al., 2000). Consider the example of investigating which areas of the brain are responsible for encoding specific types of semantic categories like "actions" (Kable et al., 2002; Noppeney et al., 2005; Wallentin et al., 2005). A simple and effective approach is to collect and compare brain responses to action words and pair them with minimally different words, perhaps similar length and frequency "object" words. If some brain element v responds more to stimuli containing the property being tested than the control stimuli—that is, f_v("action" words) > f_v("object" words)—the experimenter concludes that v is involved in processing action words. Similarly, the N400 effect (Kutas & Hillyard, 1984) is assessed by testing whether an elements' f_v reflects surprise with respect to some context. If f_v(expected word|context) < f_v(unexpected word|context), it would suggest that the brain element is capturing word surprisal.

In order for a contrast-based study to be interpretable, it is vital to remove any confounds that could corrupt observed responses and lead to false positives. Binder et al. (2009) characterize three types of confounds: the main and control conditions could differ in low-level processing demands (phonological/orthographic); the main and control conditions could differ in working memory demands, attention demands, and so forth; and, in passive tasks, the participants might engage in different mental imagery or task-related thoughts in the two conditions. If such confounds are controlled effectively, one can assume that the observed brain response will be identical in all respects unless v specifically captures the property being studied. For example, if the action and object words are matched on all other properties, f_v("action" words) and f_v("object" words) will only differ if v selectively encodes action or object concepts. Consequently, the contrast-based paradigm has high interpretability, as any variations in observed response can be attributed to the hypothesis. This clear and direct relationship between hypothesis and result ensures that the experiment has scientific value even when a hypothesis or theory is incorrect. The controlled experimental design has thus been fundamental in revealing many important aspects of brain function, such as the specialization of parts of temporal cortex for speech processing (reviewed in S. K. Scott, 2019) and distinct neural systems for concrete versus abstract concepts (Binder et al., 2005; Binder et al., 2009).

While this paradigm has been hugely influential and effective in language neuroscience, it is not without flaws. Perhaps the biggest drawback of most contrast-based designs is the lack of ecological generalizability (Hamilton & Huth, 2018; Matusz et al., 2019). To avoid confounds, controlled experiments often employ the simplest linguistic constructions required to demonstrate an effect, such as single words in the action versus object contrast. While we are fully capable of identifying action words in isolation, it is not necessary that the brain employs the same networks to understand such words in real-world settings (Matusz et al., 2019), for example, as used in a conversation or a story. In contrast to such studies, those using naturalistic stimuli have found more engagement and activation in higher order cortical regions, likely due to the incorporation of long-range structure (Deniz et al., 2021; Lerner et al., 2011). Furthermore, due to practical limitations, controlled studies typically use small stimulus sets that span a limited domain. For example, neuroimaging studies of the action contrast often use fewer than 100 words in each condition. This raises the probability that there is something peculiar or nonrepresentative about the experimental stimuli, making it more difficult to reproduce the effect or establish generalizability to a broader stimulus domain (Yarkoni, 2022). Small stimulus sets can also artificially inflate the observed statistical significance (Westfall et al., 2017).

Naturalistic stimuli:
Stimuli that subjects could be exposed to in real life; not artificially constructed for an experiment.

While controlled studies offer a very clear and direct relationship between the hypothesis and experimental result, their value depends entirely on the quality of the hypothesis. In many cases, narrowing the experimental hypothesis to focus on contrasts of a particular stimulus property may be misleading, and may fail to account for interactions between several other stimulus properties. For example, standard statistical models for assessing the "action" contrast assume that brain response is identically distributed for any subcategorization of this semantic concept. However, studies such as Hauk et al. (2004) have found that different regions across cortex selectively encode hand-related, foot-related, or mouth-related actions. This type of subcategory specificity decreases the statistical power of the overall action contrast, thereby increasing the probability of false negatives. Worse, if the overall action contrast has unevenly sampled these subcategories, the statistical power to detect action selectivity will vary in an unexpected and unknown fashion between brain areas. This issue can occur in any contrast-based experiment and is difficult or even impossible to detect by the experimenter. One potential solution would be to combine data across different contrast-based experiments, which could reveal interactions between effects. However, separate controlled experiments often do not share analysis methods, stimulus sets, or subjects, making it difficult to combine data or compare effect sizes across experiments. Lastly, for each language property that one wishes to investigate using a controlled experiment, one needs to design specific controls and repeatedly measure R_v. This results in limited reusability of experimental data, slowing down the process of scientific discovery.

Naturalistic Stimuli

To combat the lack of stimulus generalization and limited reusability, there has been a rising trend toward naturalistic experimental paradigms (Brennan et al., 2012; Hamilton & Huth, 2018; Hasson et al., 2008; Lerner et al., 2011; Regev et al., 2013; Shain et al., 2020). With the development of better neuroimaging/recording technology, we now have access to high quality brain recordings of humans while they perceive engaging, ecologically valid stimuli like podcasts (Huth et al., 2016; Lerner et al., 2011; J. Li et al., 2022; Nastase et al., 2021; S. Wang et al., 2022), fictional books (Bhattasali et al., 2020; Wehbe, Murphy, et al., 2014), and movies (J. Chen et al., 2017)—all examples of stimuli humans encounter or seek out in their everyday lives. Recent work has further developed this naturalistic paradigm to incorporate communication and social processing, beyond passive perception (Bevilacqua et al.,

2019; Redcay & Moraczewski, 2020). Naturalistic stimulus data sets are easier to construct and often larger than controlled stimuli. For example, J. Chen et al. (2017) publicly released a data set collected on a 50 min movie, Wehbe, Murphy, et al. (2014) released data collected on an entire chapter from the Harry Potter books, comprising more than 5,000 words, and LeBel et al. (2022) released data collected on over 5 hr of English podcasts per participant. These stimuli also provide a diverse test bed of linguistic phenomena—from a broad array of semantic concepts to rich temporal structure capturing discourse-level information. Furthermore, they do not directly constrain the hypotheses the experimenter can test and thus facilitate high reusability of the data. However, this also means that natural stimulus data have low statistical power with respect to any specific hypothesis, and it is necessary to carefully design analyses to control for confounding effects. This makes interpretation of the observed effects much more challenging than contrast-based experiments.

Naturalistic Experimental Design: Controlled Manipulations of Naturalistic Stimuli

To reap the benefits of both interpretable controlled experiments and generalizable naturalistic stimuli, some studies have deployed a hybrid experimental design (Chien & Honey, 2020; Deniz et al., 2019; Lerner et al., 2011; Overath et al., 2015; Yeshurun et al., 2017). Here, natural stimuli are manipulated to change or remove some specific language cue or property (e.g., scrambling the words in a story) and the sensitivity of different brain regions to this manipulation is measured, for example, f_v(intact story) vs. f_v(scrambled story). This can reveal properties across the brain like the timescale of information represented (Lerner et al., 2011, 2014) or specificity to the type of naturalistic stimulus, such as human speech (Overath et al., 2015). This experimental design accounts for ecological validity by restricting analyses to brain regions that robustly respond to the naturalistic stimuli. Furthermore, it has the same advantage of controlled experiments when it comes to interpretation: Assuming effective control of confounds, any observed change in brain activity is likely to be an effect of the stimulus manipulation. However, this approach also has disadvantages: The manipulated stimuli are often unnatural (like reversed or scrambled speech) and restrict the types of interactions the experimenter can observe. For example, the scrambled story experiment assumes that all regions processing short timescale information will behave identically. The manipulated stimuli also limit the reusability of the experiment, meaning that a new experiment needs to be designed for each effect of interest.

Ecological validity: Determiniation whether an experiment is likely to faithfully reflect and generalize to situations encountered in real life.

Naturalistic Experimental Design: Predictive Computational Modeling

Encoding models are an alternative computational approach for leveraging naturalistic experimental data (Bhattasali et al., 2019; Caucheteux & King, 2022; Goldstein et al., 2021; Huth et al., 2016; Jain et al., 2020; Jain & Huth, 2018, p. 20; Schrimpf et al., 2021; Wehbe, Vaswani, et al., 2014). These predictive models learn to simulate elicited brain responses $R_v = f_v(S)$ to natural language stimuli S by building a computational approximation to the function f_v for each brain element v, typically in every participant individually. Here, R can be captured by any neuroimaging or neural recording technique. Given limitations on data set sizes, the search for f_v is typically constrained to *linearized encoding models*, $g_v(Ls(S))$ (M. C.-K. Wu et al., 2006), where g_v is a linear combination of features extracted from the stimulus by a function Ls. While g_v is termed a *linear model*, of particular interest is the *linearizing transform* Ls. Contrast-based experimental designs test a hypothesis by comparing responses elicited by different conditions. Each condition is composed of stimuli that share some features (e.g., all words that describe actions). Encoding models can test the same hypothesis by incorporating these features into Ls. For example, for every word in the natural stimulus, one could create an

Linearized encoding model: Model that learns to predict elicited response in a brain element as a linear function of features of interest extracted from stimuli.

indicator feature I_{action} that is 1 if the word describes an action and 0 otherwise. Feature spaces consisting of 1s and 0s are equivalent to a contrast-based experimental design, assuming other confounds have been eliminated.

Encoding models can also adopt much more complex and high-dimensional functions for *Ls*. This makes it possible to account for multiple, interacting stimulus properties that may affect the response R_v. For example, *Ls* could indicate multiple levels of semantic categories. In the example of action and object words, the feature space could indicate that hand-related, foot-related, and mouth-related words were all types of actions, and distinguish all action words from multiple subcategories of objects. One recent example of such a high-dimensional feature space that captures semantic similarity (Mikolov et al., 2013; Pennington et al., 2014) is *word embeddings*, which have been used to characterize semantic language representations across the human brain (de Heer et al., 2017; Huth et al., 2016; Wehbe, Murphy, et al., 2014; Wehbe, Vaswani, et al., 2014). With a suitably rich linearizing transform *Ls*, this approach vastly expands the set of hypotheses that can be reasonably explored with a limited data set. The expandable feature space also allows encoding models great flexibility to test additional hypotheses without collecting new data, leading to high reusability. Estimating the brain response as a function of the nonlinear feature space is made possible by collecting large data sets that are partitioned into a portion for training (estimating) the model and a portion for testing the model on unseen data. Typically, regularized linear regression is used to estimate the linear relationship g_v based on the feature space *Ls*. This is used to predict new responses

$$\hat{R}_v = g_v(Ls(S))$$

to unseen stimuli. Finally, the model is evaluated by measuring how well it predicts brain responses, $\rho(f_v(S_{new}), g_v(Ls(S_{new})))$. Thus, unlike other approaches, encoding models explicitly measure generalizability by testing on new, naturalistic stimuli. In contrast-based designs, a generalization test is usually achieved through replication with an independent data set, often from a different lab where protocols and analysis details may differ. With encoding models, the same experimenter usually runs their own generalization test and directly estimates how much of the neural response R_v is explained by the model, holding all other variables constant. Encoding models can also be used to investigate if the same brain region under different tasks have the same tuning. For example, Deniz et al. (2019) show that semantic tuning is preserved between reading and listening, while Çukur et al. (2013) show that the tuning of different regions in visual cortex when attending to a given category is biased toward the attended category. Encoding models can also be used to compare tuning of two different regions (Toneva et al., 2022).

Artificial Neural Networks as a Rich Source of Linguistic Features

Neural language models: Types of artificial neural networks that learn to predict the next word in a sequence from past context.

The most important choice that an experimenter makes when using encoding models is that of the linearizing transform. To find useful linearizing transforms, neuroscience has mostly followed advances in computational linguistics or natural language processing (NLP) where, in recent years, deep learning (DL) models trained using self-supervision have seen great success. One such cornerstone model is the *neural language model*—a self-supervised artificial neural network (ANN) that learns to predict the next word in a sequence, w_{t+1}, from the context provided by previous words (w_1, w_2 ... w_t). Several recent studies have shown that representations derived from LMs capture many linguistic properties of the preceding sequence (w_1, w_2 ... w_t) like dependency parse structure, semantic roles, and sentiment (see Mahowald et al., 2020, for a review; Clark et al., 2019; Conneau et al., 2018; Gulordava et al., 2018; Haber & Poesio, 2021; Hewitt & Liang, 2019; Hewitt & Manning, 2019; Lakretz et al., 2019; B. Z. Li et al., 2021; Linzen & Leonard, 2018; Marvin & Linzen, 2018; Prasad et al., 2019; Tenney et al., 2018;

Tenney et al., 2019). While this by no means is a complete representation of phrase meaning (Bender & Koller, 2020), using a language model as a linearizing transform has been shown to effectively predict natural language responses in both the cortex and cerebellum, with different neuroimaging techniques and stimulus presentation modalities (Abnar et al., 2019; Anderson et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2021; Jain et al., 2020; Jain & Huth, 2018; Kumar et al., 2022; LeBel et al., 2021; Reddy & Wehbe, Murphy, 2020; Schrimpf et al., 2021; Toneva et al., 2020; Toneva & Wehbe, 2019; S. Wang et al., 2020; Wehbe, Murphy, et al., 2014; Wehbe, Vaswani, et al., 2014). Moreover, these models easily outperform earlier word embedding encoding models that use one static feature vector for each word in the stimulus and thus ignore the effects of context (Antonello et al., 2021; Caucheteux & King, 2022; Jain & Huth, 2018). Deep LMs have also been used to investigate the mapping between ANN layer depth and hierarchical language processing (Jain & Huth, 2018). Along similar lines and at a lower level, supervised and self-supervised models of speech acoustics have been used to develop the best current models of auditory processing in human cortex to date (Kell et al., 2018; Y. Li et al., 2022; Millet et al., 2022; Millet & King, 2021; Vaidya et al., 2022).

The unprecedented success of DL-based approaches over earlier encoding models can likely be attributed to several important factors. First, features extracted from the DL-based models have the ability to represent many different types of linguistic information, as discussed above. Second, DL-based models serially process words from a language stimulus to generate incremental features. This mimics causal processing in humans and thus offers an important advantage over static representations like word embeddings, which cannot encode contextual properties. Third, recent work has shown that these models often recapitulate human errors and judgments, such as effectively predicting behavioral data of human reading times (Aurnhammer & Frank, 2018; Futrell et al., 2019; Goodkind & Bicknell, 2018; Merkx & Frank, 2021; Wilcox et al., 2021). This again suggests some isomorphism between human language processing and DL-based models. The next word prediction objective also enables language models to perform well on psycholinguistic diagnostics like the cloze task, although there is substantial room for improvement (Ettinger, 2020; Pandia & Ettinger, 2021). Finally, self-supervised ANNs, that is, networks that predict the next word or speech frame, transfer well to downstream language tasks like question answering and coreference resolution, and to speech tasks like speaker verification and translation across languages (Z. Chen et al., 2022; A. Wu et al., 2020). This suggests that the self-supervised networks are learning representations of language that are useful for many tasks that humans may encounter.

These factors have contributed to the increasing popularity of DL-based encoding models as an investigative tool of brain function. This approach has revealed aspects of how the brain represents compositional meaning (Toneva et al., 2020), provided fine-grained estimates of processing timescales across cortex (Jain et al., 2020), and uncovered new evidence for the cerebellum's role in language understanding (LeBel et al., 2021).

Yet despite these successes, DL-based encoding models are hard to interpret. The representations produced by language models are entirely learned by black-box neural networks, and thus cannot be understood with the same ease as the indicator features described in the Naturalistic Experimental Design: Predictive Computational Modeling section above. While the representations themselves are opaque, one potential avenue is to interpret the success of a DL-based model at predicting some brain area as suggesting a commonality between that brain area and the objective that model was trained for (e.g., word identification [Kell et al., 2018] or 3D vision tasks [A. Wang et al., 2019]). However, the fact that similar representations can be derived from DL-based models that are trained for different objectives puts this type of interpretation on shaky ground (Antonello & Huth, 2024; Guest & Martin, 2023). These

difficulties have left the field at something of an impasse: We know that DL-based models are extremely effective at predicting brain responses, but we are unsure why and unsure what these models can tell us about how the brain processes language.

Pièce De Résistance: In Silico Experimentation With DL-Based Encoding Models

Controlled experiments and encoding models using naturalistic stimuli both have distinct advantages and disadvantages. However, it may be possible to combine these paradigms in a way that avoids the disadvantages and retains the advantages. To this end, we present an experimental design that combines these two paradigms: in silico controlled experimentation using encoding models. This paradigm first trains encoding models on an ecologically valid, highly generalizable naturalistic experiment. Then, it uses the encoding models to simulate brain activity to controlled stimulus variations or contrasts. Notably, this does not require additional data to be collected for every condition.

The first use of in silico experimentation is to test if effects discovered in controlled, non-ecologically valid setups generalize to naturalistic stimuli. This experimental design also facilitates quick and efficient hypothesis testing. Experimenters can prototype new controlled experiments and narrow down the desired contrasts or stimuli without having to repeatedly measure in vivo. While this is a complement to and not a substitute for in vivo experiments that should follow the prototyping phase, in silico experimentation can greatly reduce the cost of generalizability and hypothesis testing, and accelerate scientific discovery.

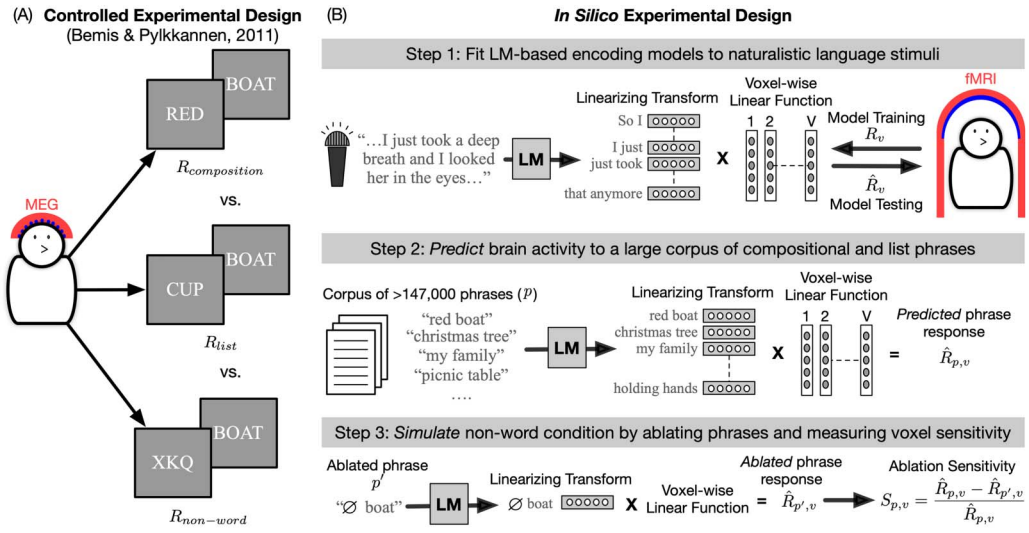

Figure 1. Example of an in silico adaptation of a controlled experiment. (A) The original MEG study investigated composition over two-word phrases (Bemis & Pylkkänen, 2011). This was done by presenting three different types of phrases to participants to solve a picture matching task. By contrasting the elicited brain responses in the composition condition with the responses in the list and non-word conditions, the authors could infer which brain regions are engaged in compositional processing of two-word phrases. (B) This experimental paradigm can be conceptually simulated with LM-based fMRI encoding models of naturalistic stimuli. The composition and list conditions can be tested by using the learned encoding model to predict each voxel's response to a large, diverse corpus of phrases. The non-word condition can be simulated by replacing the first word in a phrase with a non-word, extracting new *ablated* features of the phrase from the LM and using the encoding model to predict the brain's response to the ablated phrase. If a voxel's response is highly sensitive to the removal of the first word, it would suggest that the voxel combines information over both words to arrive at meaning. This provides a data-efficient way to test for compositional processing across diverse types of phrase constructions. fMRI = functional magnetic resonance imaging; LM = language model; MEG = magnetoencephalography.

In Figure 1, we present a controlled experimental design with its in silico counterpart. Figure 1A shows an experimental paradigm that was designed to understand linguistic composition of two-word phrases (Bemis & Pylkkänen, 2011). Participants were presented with phrases in which meaning can be composed across constituent words and contrasting conditions where it cannot (word list and non-word). This experiment can be conceptually simulated in silico, as shown in Figure 1B (Jain & Huth, 2023). Instead of collecting separate neuroimaging data for each type of phrase construction, the in silico experiment was done with DL-based encoding models trained on two-word sequences. The learned models were first used to predict brain responses to a large, diverse corpus of phrases that contained both noun–noun and adjective–noun constructs among others. Next, the non-word condition was simulated by replacing the first word in the phrase with a non-word, extracting a new *ablated* feature, and finally predicting each functional magnetic resonance imaging (fMRI) voxel's response to the ablated phrase. Assuming that the DL-based encoding model captures compositional effects, this in silico experiment can ameliorate the disadvantages of both controlled and encoding model-based experimental designs. First, since simulating responses is trivial in both time and cost, the simulated experiment can use thousands or even millions of two-word phrases instead of the hundreds that can be tested in vivo. This ameliorates problems that arise with limited stimulus sets that may fail to account for key properties or generalize to naturalistic contexts. Second, by simulating and then comparing responses under conditions that are derived from linguistic theory (composition vs. single word, or word list), this in silico experiment provides results that are easily and explicitly interpretable, unlike encoding models with natural stimuli. However, one major concern raised by this approach is whether the encoding model can capture how the brain responds to the language properties of interest. To address this it is important to verify both that the encoding model is highly effective at predicting brain activity, and that it is sufficiently complex to capture the desired property.

Similar in silico experimentation has recently become popular in vision neuroscience. There, DL-based encoding models of the visual system are first trained on ethologically valid tasks like object recognition. Then they are probed to understand brain function (Yamins & DiCarlo, 2016). For example, Bashivan et al. (2019) used DL-models to synthesize images that maximally drive neural responses. This provided a noninvasive in silico technique to control and manipulate internal brain states. Similarly, Ratan Murty et al. (2021) synthesized images from end-to-end DL models trained on brain data to provide stronger evidence for the categorical selectivity of different brain regions. In silico experimentation with explicit computational models has also been used in studies of the medial temporal lobe. In Nayebi et al. (2021), computational models of the medial entorhinal cortex were used to investigate the functional specialization of heterogeneous neurons that do not have stereotypical response profiles. By doing ablation in silico, they found that these types of cells are equally important for downstream processing as are grid- and border-cells. Each of these studies first relied on the encoding model's ability to generalize to new stimuli. This was an indication that the features learned by the DL-based models encoded similar information to the brain regions that they predicted well. Second, these studies leveraged the predictive ability of encoding models to simulate brain activity in new, controlled conditions as a lens into brain function. This enabled the researchers to explore aspects of brain function that would otherwise be highly data intensive or impossible to do.

In language, in silico experimentation is a promising area that is under development, bolstered by the successes in vision neuroscience and growing efforts to understand artificial language systems. One of its earliest uses is the BOLDpredictions simulation engine (Wehbe et al., 2018; Wehbe et al., 2021), an online tool that allows the user to simulate language experiments that contrast two conditions, each defined by a list of isolated words. BOLDpredictions relies on an

encoding model from a natural listening experiment that predicts brain activity as a function of individual word embeddings (Huth et al., 2016). In the following sections, we review in silico adaptations of four different language experiments based on four separate data sets. Each of these in silico experiments uses a single naturalistic experiment to train the encoding models, illustrating how a single data set and experimental approach can provide a flexible test bed for many different hypotheses about natural language. The first experiment uses the BOLDpredictions engine to simulate a semantic contrast comparing concrete and abstract words (Binder et al., 2005), testing its generalizability to naturalistic settings. The next experiment focuses on a contrast-based study of composition in two-word phrases (Bemis & Pylkkänen, 2011), testing generalizability over a broader, more diverse stimulus set. The third experiment adopts contrasts from a group-level study investigating the temporal hierarchy for language processing across cortex by manipulating naturalistic stimuli (Lerner et al., 2011). This simulation checks if effects persist at the individual-level and demonstrates how a successful replication can be used to validate computational model constructs themselves. Finally, the last experiment conceptually replicates a study on forgetting behavior in the cortex that also uses controlled manipulations of naturalistic stimuli (Chien & Honey, 2020). This simulation demonstrates the possibility of misinterpretation with the in silico approach, arising from fundamental computational differences between neural language models and the human brain.

Generalizability testing: Testing to see if effects observed on a particular data set extend to a new data set not used for model estimation.

In the experimental simulations described below, voxelwise encoding models were fit to fMRI data collected from a naturalistic speech perception experiment. Participants listened to natural, narrative stories from *The Moth Radio Hour* (Allison, 2009–) while their whole-brain BOLD responses were recorded ($N = 8$ for study 1; $N = 7$ for studies 2 and 3). In each study, encoding models were fit for each voxel in each subject individually using ridge regression. The learned models were then tested on one held-out story that was not used for model estimation, and encoding performance was measured as the linear correlation between predicted and true BOLD responses. Statistical significance of the encoding performance was measured using temporal blockwise permutation tests ($p < 0.001$, false discovery rate (FDR) corrected; Benjamini & Hochberg, 1995). Finally, in silico analyses were conducted on voxels that were significantly predicted by the encoding model, broadly covering the temporal, parietal, and frontal lobes.

Semantic contrasts:* *Wehbe et al. (2018)

Binder et al. (2005) investigated the brain regions responsive to abstract and concrete concepts. Subjects read individual stimulus strings and pressed one of two buttons to indicate whether each one was a word or a non-word. The study reported that concrete words activated bilateral language regions such as the angular gyrus more than abstract words, and abstract words activated left inferior frontal regions more than concrete words. In total, the authors found 15 cluster peaks.

Wehbe et al. (2018) evaluated the reproducibility of these results using an encoding model trained on naturalistic stimuli. They simulated a contrast between the lists of concrete words and abstract words that were kindly shared by Binder et al. (2005). Figure 2 shows the significance map for subject 1 and the group-averaged significance map showing the number of subjects for which the null hypothesis is rejected. The reported regions of interest (ROIs) are shown as an overlay on the flattened cortical maps. Each ROI is originally reported as a single coordinate in brain space and is estimated to have a radius of 10 mm. For every one of the eight subjects, many voxels were significantly more activated for concrete words over abstract words (with $p < 0.05$, FDR corrected permutation test over the words in each condition), specifically in areas bordering the visual cortex and parts of the inferior frontal gyrus. Some reported ROIs had a high overlap with the significance map (specifically in the angular gyri,

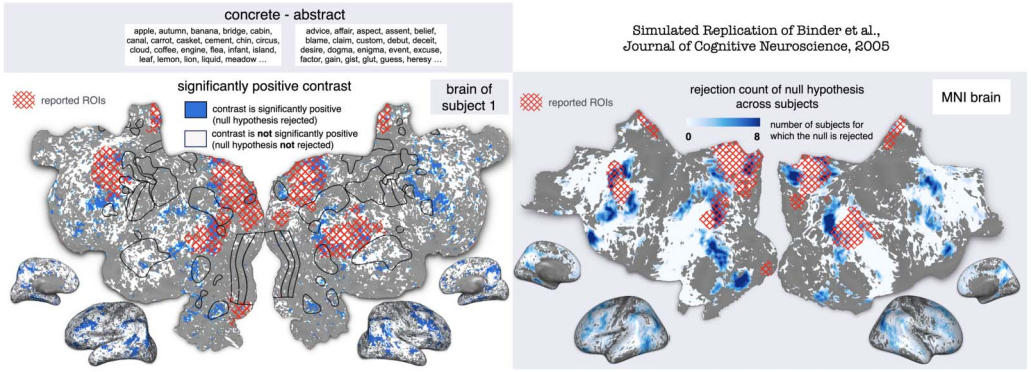

Figure 2. Generalizability test using BOLDpredictions of the concrete vs. abstract contrast of Binder et al. (2005). The authors compared fMRI activity when subjects processed concrete and abstract words. Wehbe et al. (2018) used the published stimulus to simulate the contrast for each subject and run a permutation test. After MNI space transformation, the number of subjects for which the null hypothesis was rejected is computed at each voxel. The simulated statistical maps are shown on flattened maps and inflated 3D hemispheres. Results for subject 1 are shown in subject 1's native cortical space. Results for the average of eight subjects are shown in the MNI space. Published ROIs are estimated as 10 mm radius spheres, shown in red hatch on the flatmaps (distortion due to the flattening process). A comparison of the overlap of the reported ROIs and the statistical maps reveals that Wehbe et al. (2018) achieve a relatively high overlap for specific ROIs (in the angular gyri, in the posterior cingulate gyri, the right precuneus, and the middle temporal gyri) and not for others. Therefore, BOLDpredictions predicts that the contrast from Binder et al. (2005) generalizes to naturalistic conditions, to a certain extent. MNI = Montreal Neurological Institute; ROIs = regions of interest.

in the posterior cingulate gyri, the right precuneus, and the middle temporal gyri). The significant effect in those ROIs can be considered to be replicated by BOLDpredictions. However, the reported ROIs and the significance map did not always agree, with the effect in some regions being reported only by Binder et al. (2005) or only by Wehbe et al. (2018).

There are many possible reasons for non-generalizability of individual reported ROIs, including the stochasticity of brain activity, variations in experimental paradigms and analysis techniques, and lack of reproducibility. The authors of BOLDpredictions (Wehbe et al., 2018; Wehbe et al., 2021) note that any scientific finding needs to be reproduced in a series of experiments that would create a body of evidence toward this finding, and the in silico experimentation using BOLDpredictions is one additional piece of evidence. The authors also note that expanding the engine to different data sets, models, and so forth will establish the robustness of the in silico effects and help determine if the original contrast-based experiment lacks reproducibility (Wehbe et al., 2018).

Semantic composition contrasts: Jain and Huth (2023)

In the second in silico experiment, Jain and Huth (2023) simulated and expanded on studies of combinatorial processing in 2-word phrases across cortex, first described in Bemis and Pylkkänen (2011). The original controlled experiment consisted of participants reading two word adjective–noun phrases ("red boat") and doing a picture matching task while brain responses were recorded using MEG. To contrast this compositional condition, a list control was introduced wherein participants were presented with two-word noun–noun phrases ("cup boat") along with a non-word control consisting of a non-word and a word ("xkq boat"). Note that participants were instructed to avoid composing meaning in the word list, but no explicit control was introduced. To isolate regions involved in two-word composition, the study contrasted the adjective–noun condition with the controls. The experimenters tested 25 base nouns, six color adjectives, and six non-words. Overall, they found that areas in ventral medial

prefrontal cortex (vmPFC) and left anterior temporal lobe both selectively responded to the composition condition.

Jain and Huth (2023) conceptually replicated the original study by building encoding models that approximate every voxel's response to a naturalistic two-word phrase as a non-linear function of the words in the phrase. For each (overlapping) two-word phrase in the natural language stimuli, features were first extracted from a powerful language model, the generative pretrained transformer (GPT; Radford et al., 2018). Then, voxelwise encoding models were trained to learn a linear function from the phrase features to the elicited response after the second word. Using the encoding models, each voxel's response to a large corpus of over 147,000 two-word phrases was predicted and ranked. This stimulus set comprised both adjective–noun phrases like "red boat" and noun–noun phrases like "picnic table." Next, for a given phrase selected by a voxel, the first word was replaced with a non-word (i.e., the word was ablated) and the ablated phrase feature was extracted from GPT. Using the learned encoding model, the voxel's response to the ablated phrase was predicted. Finally, the sensitivity of the voxel to the presence of the ablated word was measured. If the ablated word is important for the voxel to process the phrase, removing it should notably change its response and give high sensitivity. This was done to simulate the compositional versus non-word condition in the original study.

The resultant ablation sensitivity of voxels across cortex is visualized in Figure 3. Overall, the in silico experimentation produced similar results to the original study in vmPFC and left anterior temporal lobe—both of these regions exhibit sensitivity to the presence of a

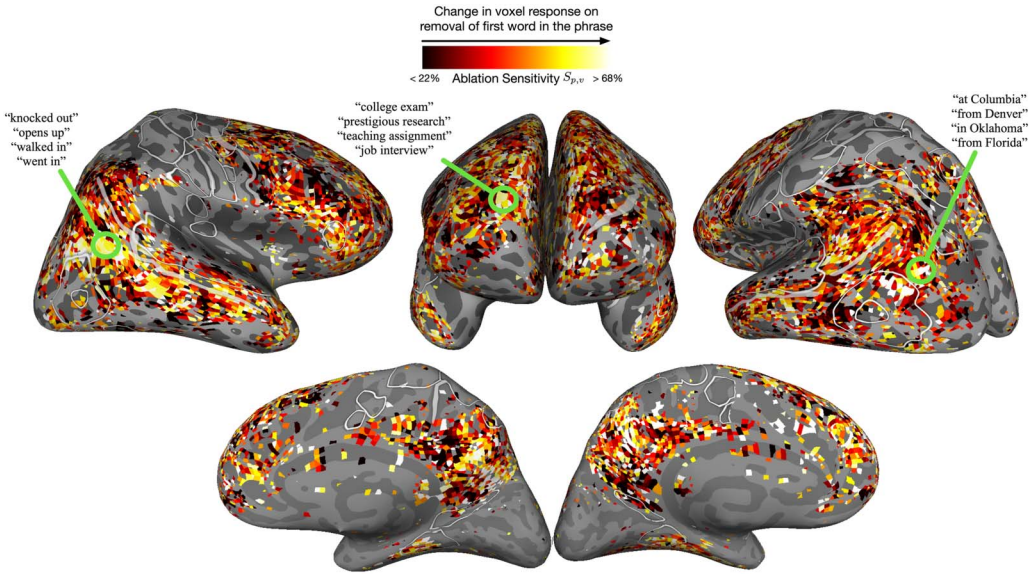

Figure 3. In silico adaptation of a study examining compositional processing in two-word phrases. The original study compared the MEG responses of participants with three different types of two-word phrases: adjective–noun, noun–noun and non-word–noun (Bemis & Pylkkänen, 2011). The in silico simulation of the first two conditions was done by constructing a larger diverse corpus of phrases and using LM-based encoding models to predict fMRI voxel responses to each phrase. The non-word–noun condition was simulated by replacing the first word in a phrase with a non-word (i.e., word ablation), extracting new phrase features from the LM, and then predicting the voxel's response to the ablated phrase. A large change in a voxel's response upon word ablation indicated its sensitivity to the first word in the phrase and suggested that the voxel relied on the first word to process the phrase. Similar to the original study, the in silico experiment revealed high sensitivity in ventral medial prefrontal cortex and left anterior temporal lobe. However, the experiment also found that several other areas across cortex combined meaning over the two words in a phrase and moreover, captured diverse semantic concepts arising from the composition. fMRI = functional magnetic resonance imaging; LM = language model; MEG = magnetoencephalography.

compositional word. Beyond areas reported originally, other regions like right inferior parietal and dorsal prefrontal also showed high sensitivity. This finding corroborates other studies of phrase composition (e.g., Boylan et al., 2015; Graves et al., 2010). The in silico study was able to analyze two-word composition in broader regions of cortex by simulating activity for each voxel independently and over a much larger stimulus set that comprises diverse concepts and constructions. While the simulation does not guarantee causal involvement of any region in two-word composition, it demonstrates the utility of broadly sampling stimuli and raises the possibility that many more regions are involved in this process. Moreover, in the in silico study Jain and Huth (2023), this paradigm was extended to much longer phrases (10 words) to understand the relationship between semantic representations and word-level integration across cortex. This would be difficult to implement in real-world settings as doing single-word ablations on increasingly longer phrases is combinatorially explosive.

Construction timescale contrast: Vo et al. (2023)

In the third in silico experiment, Vo et al. (2023) tested whether voxelwise encoding models based on features from a neural LM can capture the timescale hierarchy observed during human natural language comprehension. In Lerner et al. (2011), subjects listened to a first-person narrative story that was either intact, reversed, or temporally scrambled at the word level, sentence level, or paragraph level. The scrambling manipulations altered the temporal coherence of the natural language stimulus, and allowed the researchers to measure the reliability of fMRI responses to each condition using intersubject correlation. This revealed an apparently hierarchical organization of temporal receptive windows, with information over short timescales processed in auditory cortex and long timescales processed in parietal and frontal cortex. For the in silico adaptation, the authors trained a multi-timescale long short-term memory (MT-LSTM) network as a language model (Mahto et al., 2020). Then they used the features from the MT-LSTM to predict fMRI responses for each voxel using the data set described above. To mimic the manipulations of the original study, they generated 100 scrambled versions of a held-out test story. This enabled the authors to examine the predicted fMRI responses within each voxel in each subject. Rather than measuring intersubject reliability, they chose to measure an analogous intrasubject reliability value, testing whether the scrambling condition caused a significant drop in this value across conditions. The authors show through simulations that their metric (based on the variance in the simulated fMRI response) is directly analogous to intersubject correlation measures, which is supported by other work (Blank & Fedorenko, 2017; Hasson et al., 2009).

The results of this experiment compared to a schematized version of the original results are shown in Figure 4. This in silico experiment reproduced the pattern of the temporal hierarchy along the temporoparietal axis. It did find that some regions in frontal cortex appear to integrate over shorter timescales than the original work, similar to a later replication of the work (Blank & Fedorenko, 2020) and to a different in-silico replication of the experiment that used GPT-2, rather than a MT-LSTM language model (Caucheteux et al., 2021). Furthermore, the fine-grained resolution of the single-voxel analyses revealed substantial variability across subjects. Taken together, the in silico results suggest that timescales are not as uniform across broad regions as previously reported. This is in agreement with single-neuron studies that show a heterogeneity of intrinsic timescales within a brain region (Cavanagh et al., 2020).

Forgetting timescale contrast: Vo et al. (2023)

In the last experiment, the authors used the same MT-LSTM encoding models as experiment 3 to simulate how different brain regions forget information in naturalistic narrative stimuli

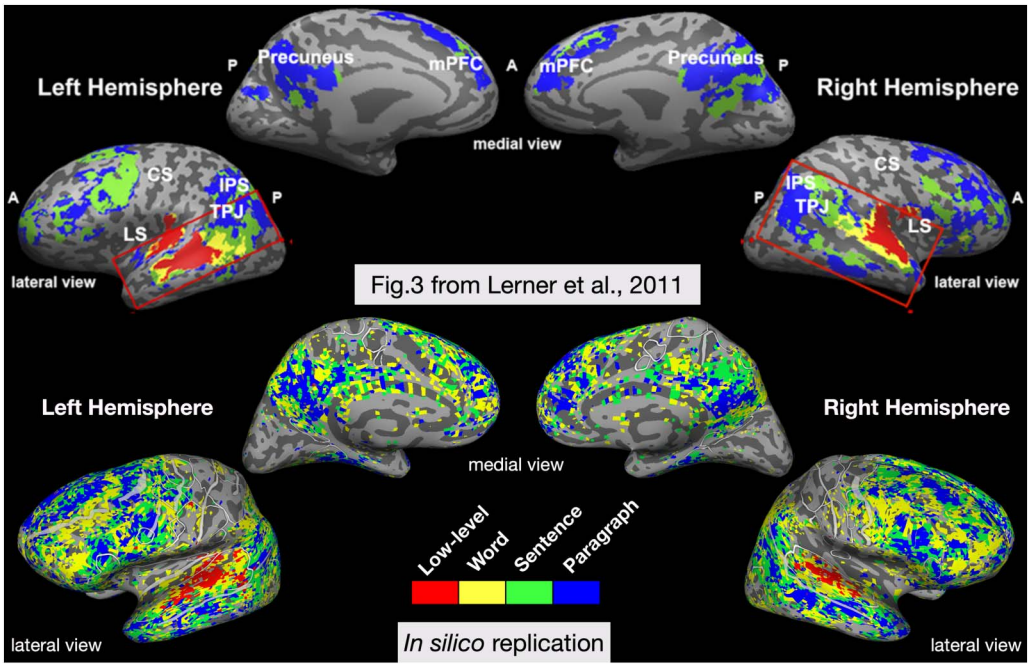

Figure 4. In silico adaptation of a study mapping the hierarchy of temporal receptive windows. (Top) Original results adapted from Lerner et al. (2011). The authors played an audio clip of a narrative story, either intact, reversed, or scrambled at different temporal scales. The figure shows an overlay of several intersubject correlation maps, which measured the cross-subject reliability of the fMRI response in each condition. (Bottom) Results from the in silico experiment of temporal receptive windows, shown for every significantly predicted voxel on a single subject. The in silico experiment suggests that temporal processing windows for different brain regions are not as uniform as previously reported. CS = central sulcus; IPS = intraparietal sulcus; LS = lateral sulcus; mPFC = medial prefrontal cortex; TPJ = temporoparietal junction.

(Chien & Honey, 2020). While Chien and Honey found that all brain regions forget information at the same rate (Figure 5A), the in silico results suggested that low-level regions such as auditory cortex forget information at a faster rate than high-level regions like the precuneus (Figure 5B). To better understand this discrepancy, the authors investigated forgetting behavior in the MT-LSTM itself. The results first indicated that every unit in MT-LSTM forgot information at a specific rate tied to its processing timescale (Figure 5C). The authors further hypothesized that the discrepancy could stem from the MT-LSTM's inability to forget information, even if the preceding context is noisy/uninformative (Figure 5D). To test this, they measured the language model's cross entropy (lower is better) for a paragraph in three conditions: preceded by the correct paragraph (*actual context*), preceded by no paragraph (*no context*) and preceded by random paragraphs in the story (box plot of 100 different *incorrect contexts*). The story was scrambled by dividing it into non-overlapping chunks of 9, 55, 70, 80, or 200 words or at the actual sentence and paragraph boundaries (*hand-split*). Overall, smaller differences were observed between the conditions as the scrambled context became longer (increased chunk size) and closer to the intact story. With fixed-size chunks, the model performed better when it had no context than when it had access to incorrect information. In contrast, with actual sentences/paragraphs, the model had better performance with incorrect context than no context at all. In both cases, the type of context influences the model performance suggesting that the model retains information from the past. Second, it retains this context even if it is not useful, as in the fixed-chunk conditions. The model could have simply ignored the wrong context to perform better but it did not (or was unable to). This highlights the language model's inability to forget information that is then reflected in the encoding model results. The authors

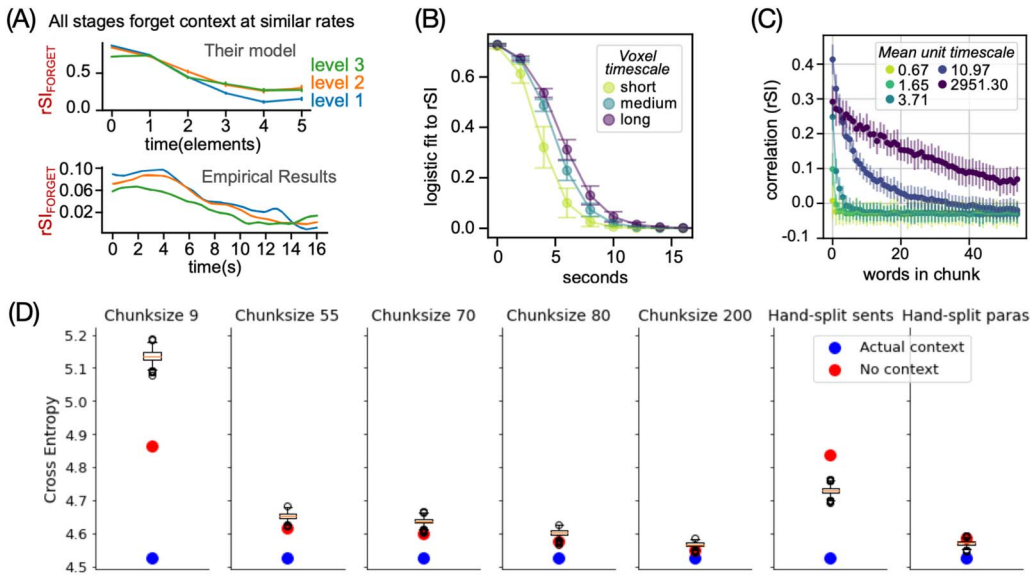

Figure 5. In silico adaptation of a study on forgetting behavior during natural language comprehension. In the original study, Chien and Honey (2020) scrambled paragraphs in a story and analyzed how quickly different brain regions forgot the incorrect context preceding each paragraph. The in silico adaptation used the MT-LSTM based encoding model to predict brain activity at different points in a paragraph when it was preceded by incorrect context. (A) The original study reported that each brain region (denoted by different colored lines) forgot information at a similar rate, despite differences in construction timescales. (B) In contrast, the in silico replication estimated that regions with longer construction timescales also forgot information slowly. (C) Within the MT-LSTM itself, the forgetting rate of different units was related to its attributed timescale. (D) Next, the MT-LSTM's language modeling abilities were tested on shuffled sentences or paragraphs. The DL model achieved better performance at next-word prediction by using the incoherent, shuffled context as opposed to no context at all. This shows that the DL model retains the incoherent information, possibly because it helps with the original language modeling task it was trained on or because the model has no explicit mechanism to flush-out information when context changes (at sentence/paragraph boundary). The computational model's forgetting behavior thus differs from the brain, revealing specific flaws in the in silico model that could be improved in future versions, such as a modified MT-LSTM. DL = deep learning; MT-LSTM = multi-timescale long short-term memory; rSI = correlation between scrambled and intact chunks.

hypothesized that with hand-split sentences and paragraphs, the incorrect context still provides relevant information to predict the next word, leading to better performance than no context at all.

DISCUSSION

Advantages of the in Silico Experimental Design

In the following sections, we discuss the advantages of using in silico experimental design with DL-based encoding models and its potential impact on language neuroscience.

Hypothesis development and testing

Each of the studies above conceptually replicated controlled analyses of different linguistic properties using voxelwise encoding models fit on a single naturalistic fMRI data set. Overall, the studies reproduced several effects reported in the original experiments. Interestingly, however, the in silico experiments also found new effects that had not been explored originally. For example, the first experiment suggested that regions in inferior frontal gyrus were more active for concrete words than abstract words. In the second experiment, the investigation of composition in phrases was expanded to a much larger stimulus set and longer phrases. This corroborates earlier results, but the in silico paradigm also enables experimenters to explore

interactions between the effect of interest (here, linguistic composition) and other important linguistic properties, such as semantic category. The third experiment found more diversity in timescales in regions like prefrontal cortex than previously reported, closely matching more recent studies of timescale distribution across cortex (Blank & Fedorenko, 2020; Caucheteux et al., 2021) and single-neuron studies (Cavanagh et al., 2020). This demonstrates how the in silico experimental design can be used not only to reproduce and test the generalizability of controlled studies, but also to conduct large-scale exploratory and data-driven analyses that can reveal new aspects of brain function.

Beyond these examples, it is possible to test new hypotheses using in silico experiments before collecting data for a controlled experiment. Wehbe et al. (2018) showcase how BOLD-predictions and in silico experimentation can be used to design new experiments. While the in silico results may not precisely match the eventual data collected on a human population, they would reveal areas where the underlying DL model has failed to match human-level processing and present possible areas of improvement. This has the potential to advance our understanding of both neural network language models and biological language processing. In particular, the in silico paradigm can both draw upon large-scale multidisciplinary efforts to build tools and methods for interpreting neural network language models (Ettinger et al., 2018; Hewitt & Manning, 2019; Ravfogel et al., 2020; Sundararajan et al., 2017), as well as contribute to them by providing a human neural benchmark. Furthermore, more interpretable models allow for novel causal intervention experiments that perturb and control ANNs in ways that biological neural networks cannot be perturbed (Zhang et al., 2022).

Testing for generalizability of effects and experimental construct validity

One way to ensure the observed effects of the in silico experiments are not due to the specific task design is to test the generalizability of effects across model architectures, training tasks, neuroimaging data sets and modalities. Unlike reproducibility tests in traditional neuroimaging experiments, these tests do not rely on laborious and time-consuming data collection. Moreover, there are increasingly more tools and techniques to interpret DL models (Clark et al., 2019; Ettinger, 2020), and we can target investigations to these. For example, in the forgetting experiment, the authors checked how the model represented the cognitive process itself. We note that some drawbacks of DL models persist across architectures and tasks. For instance, current language models still perform poorly on common sense reasoning and struggle with capturing long-term dependencies in language. However, with technological advancements, the types of inferences we can make with the in silico paradigm will greatly improve. A case in point is the modeling of contextual processing in the brain. Until recently, language encoding models were largely restricted to static word embeddings that made it difficult to analyze how the brain processed word sequences. However, with the advent of neural language models, this has changed dramatically.

In vision neuroscience, the functional profile of the fusiform face area was established through contrast experiments that evolved over a long period of time (Anzellotti et al., 2014; Gauthier et al., 2000; Kanwisher et al., 1997; Liu et al., 2010; Sergent et al., 1992; Suzanne Scherf et al., 2008; Xu, 2005). Each new experiment was designed to address a confound that was not accounted for previously. Today, however, in silico experiments with vision models have enabled neuroscientists to efficiently contrast large, diverse sets of stimuli and establish the functional specificity of different regions (Ratan Murty et al., 2021). Similarly, in language neuroscience, encoding models have been used to evaluate the semantic selectivity of many regions going beyond semantic contrasts that are tested for a handful of conditions at a time (Huth et al., 2016; Mitchell et al., 2008). This demonstrates how the in silico

Construct validity: Determination whether a theoretical, experimental, or computational construct faithfully reflects the true phenomena.

paradigm allows scientists to quickly design and test multiple experiments that get at the same underlying question. This means that in silico experiments, despite using similar manipulations to controlled experiments, can provide an additional way to test the *construct validity* of the experiment (Yarkoni, 2022). When coupled with generalizability testing, we run a lower risk of over-claiming or over-generalizing.

Establishing the validity of model constructs

The in silico approach uniquely facilitates experimenters to evaluate and improve the design of computational models based on observed in silico behavior, going beyond good prediction performance. For example, in the forgetting experiment, the authors identified that the MT-LSTM language model does not instantaneously switch context, and this could influence the observed effects. One possible solution to the nonreplication would be to then train the language model on a new task that encourages forgetting. Alternatively, it could prompt the need for designing alternate architectures that have a built-in forgetting mechanism closer to observed human behavior. Artificial neural networks can be investigated through causal intervention experiments and perturbations, whereas it is very difficult to impossible to do this for human language systems. By analyzing the behavior of DL models in many in silico experiments, we can create a check-and-correct cycle to build better computational models of the brain and establish the validity of model constructs.

An analogous paradigm has also risen in popularity in NLP. Moving beyond better performance with larger language models, there has been a growing effort toward curating diverse language tasks like syntactic reasoning and multistep inference to understand the limitations of current models and establish a benchmark for future innovation. In the same vein, we believe that many different in silico experiments can be used together to establish the validity of different model constructs and provide a benchmark to test future innovations in computational language modeling. We hope that this pushes the field past solely testing encoding model performance on different architectures. Ultimately, this paradigm is a bridge between computational models and experimental designs in neuroscience, such that we can make joint inferences on both and improve them in tandem.

Preserving individual participant differences

One potential advantage of in silico encoding model experiments is that the models are typically estimated with single-subject data, allowing experiments to test for effects in individuals rather than over a group average. While group averaging is a common method to improve the signal-to noise ratio (SNR) of neuroimaging studies, it can lead to an underestimation of effect size (Fedorenko, 2021; Handwerker et al., 2004) and hide effects that can be seen with more fine-grained functional anatomical data (Popham et al., 2021). Finally, individual participant analysis does not preclude the estimation of how prevalent an effect is at the population level (Ince et al., 2021); however, it does enable experimenters to account for individual differences, which can be critical to establish links between brain and behavior (Hedge et al., 2018). Consequently, there has been a rising trend toward language studies that analyze participants individually and report consistency of effects across the group (Blank & Fedorenko, 2020; Huth et al., 2016; Wehbe, Vaswani, et al., 2014). While this requires the experimenter to collect more samples per subject to improve the SNR, this approach does not make assumptions about anatomical alignment and preserves spatial resolution important for inferring brain function. The improved sensitivity provides better control for Type 1 errors (by allowing the experimenter to see which effects replicate across participants) and Type 2 errors (by allowing a flexible mapping that can identify important regions in each participant, even if they do not match perfectly in anatomy).

However, the individual-participant analytic approach raises important questions about how to isolate functionally consistent regions across participants and infer consistency of effects. One solution is to use a common set of functional localizer stimuli across participants to isolate functionally homologous networks. For example, the auditory and visual language localizers developed by Fedorenko et al. (2010) and T. L. Scott et al. (2017) have been shown to robustly identify regions across cortex that are important for language processing. This approach enables future studies to consistently isolate language processing regions and characterize their function. Modeling approaches such as hyperalignment (Haxby et al., 2020) and probabilistic mapping of the cortical surface (Huth et al., 2015) offer solutions to compute group-level maps from functional data of individual participants. Nevertheless, these approaches do not provide a computational framework to model individual-participant effects. Encoding models, on the other hand, learn a different function for each brain element in each subject. This enables them to effectively model individual participants and retain high spatial resolution.

Improving reproducibility in language neuroscience

There has been an increasing concern in the sciences about the lack of reproducibility for many types of experiments (Pashler & Harris, 2012; Simmons et al., 2011), a problem to which neuroscience is not immune. Several papers have discussed the prevalence of analysis variability, software errors, nontransparent reporting of methods, and lack of data/code sharing as primary causes for low reproducibility and generalizability in neuroscience (see Barch & Yarkoni, 2013, and Poldrack et al., 2020, for introductions to special issues on reproducibility in neuroimaging; Button et al., 2013; Evans, 2017). These studies have also identified issues in statistical analyses, like low statistical power (and, consequently, inflated effect sizes), HARKing, and p-hacking. We believe that the in silico experimentation paradigm can help alleviate some of these issues by providing access to and encouraging open tools for scientific research. When combined with open access to naturalistic data, preprocessing methods, and analysis code, the in silico paradigm can enable scientists to use a standard setup as they test a variety of different hypotheses and thus reduce the "researcher degrees of freedom." Platforms such as BOLDpredictions can help with this. Indeed, BOLDpredictions is intended as a community tool to allow easy in silico experimentation and generalization testing. It is intended to allow other researchers to contribute their encoding models for other experiments (even outside of language) so that in silico experiments can be available to all. Furthermore, competitions such as Brain-Score (https://www.brain-score.org/competition/) and the SANS'22 Naturalistic fMRI data analysis challenge (https://compsan.org/sans_data_competition/content/intro.html) can align scientific work toward a common goal and facilitate verifiability. Since naturalistic experiments broadly sample the stimulus space, the in silico paradigm can also act as a test bed for generalizability.

Caveats and the Possibility of Overinterpretation

The in silico paradigm leverages the advantages of both controlled and naturalistic experimental designs with DL-based encoding models. However, it is important to recognize the caveats of this approach so as to minimize the risk of overinterpretation. Here we discuss a number of potential issues.

Limitations in the natural language stimulus

One critical advantage of naturalistic stimuli over controlled designs is the access to many diverse examples of language use. However, this also means that the experimenter has little

control over the rate of occurrence of different types of linguistic features. Word frequency is an example of uncontrolled variation in natural stimuli (e.g., high frequency of words describe everyday objects like "table" and "book" as opposed to low-frequency words like "artillery" and "democracy"). This presents an important challenge in naturalistic paradigms as the rare variables will have low power and could lead to incorrect or incomplete inferences of brain function. For example, if a voxel encodes semantic concepts related to politics and governance, but this category is not well represented in the naturalistic stimuli, the experimenter runs the risk of incorrectly inferring the voxel function. This can be addressed by building larger, freely available data sets collected from diverse sources and encouraging replication of effects on them.

Another issue with naturalistic paradigms is that they currently rely on the passive perception of language. Many studies have shown that turn-taking in conversation is an important, universal aspect of communication and has implications on how we learn, understand and generate language (Levinson, 2016). Despite a rising trend toward stimuli that take into account social and contextual information, we are still far from studying truly natural use of language with neuroimaging. Some work has investigated aspects of conversational communication (Bögels et al., 2015; Gisladottir et al., 2015; Magyari et al., 2014; Sievers et al., 2020), but the field is still behind in modeling these effects with encoding models or ANNs. Richer data sets will be key to developing these approaches, such as the real-world communication data collected in Bevilacqua et al. (2019) or multimodal movie stimuli discussed in Redcay and Moraczewski (2020). This is an important future direction for the naturalistic paradigm to understand the brain mechanisms of language processing in ethological settings.

Limitations in the DL-based feature space

Perhaps the most important factor guiding the in silico experimental design is the success of DL models at predicting brain activity. This paradigm allows neuroscientists to inspect brain function by conducting simulations on the computational model instead, which is easier to perturb, interpret, and control. However, this also means that the types of effects we can observe are limited by the capabilities of the DL model. For example, the forgetting experiment by Vo et al. (2023) demonstrates how the computational model has different behavior than the human brain, affecting the observed in silico behavior. Domain shift presents another common issue for neural networks, although recent studies has proposed that fine-tuning on the target domain/task (Radford et al., 2018) and dynamic data selection during training (Aharoni & Goldberg, 2020; van der Wees et al., 2017) can greatly alleviate this problem for language models. Several encoding model studies explicitly fine-tuned the language model to operate in set (Jain et al., 2020) or trained the language model on a corpus specifically curated to resemble the experimental task (Jain & Huth, 2018; Wehbe, Vaswani, et al., 2014). Furthermore, while ANNs like language models have been successfully employed for a wide range of tasks, their syntactic, common sense, and logical reasoning abilities are still far from those of humans (Ettinger, 2020; Linzen, 2020; Pandia & Ettinger, 2021; Wilcox et al., 2021). Overall, it is important to note that building good encoding models of brain activity and understanding brain function with the in silico paradigm are both contingent on better artificial models of language processing.

Fine-tuning:
A secondary learning procedure that modifies an already trained artificial neural network to adapt to a new task or data set.

Limitations in computational modeling

Another source of confounds in encoding models and the in silico paradigm is incorrect modeling assumptions. For example, Jain et al. (2020) highlight that many fMRI encoding models rely on a downsampling technique that incorrectly transforms slowly varying features,

making them highly correlated with local word rate. Consequently, an experimenter may (incorrectly) conclude that a brain region that is well predicted by the downsampled features is selective for the slowly varying information (e.g., discourse) it captures, when, in fact, the brain region merely responds to the rate of words. In other cases, it may be important to model several different sources of noise, which has been pursued in other work simulating fMRI data (Ellis et al., 2020). Current neuroimaging modalities also have low SNR, limiting the predictive performance of computational models. Because all modeling approaches likely have caveats and edge cases for which their assumptions fail, it is important to clearly articulate and discuss these issues in future work.

Zone generalization: Determination whether two brain regions process stimuli similarly.

Inappropriate causality and zone generalization

Unlike contrast-based experiments and encoding models with simple interpretable features like indicator variables, DL-based encoding models rely on ANNs that are themselves hard to interpret. To this end, any significant correlation observed between brain activity and model predictions leaves many possibilities for interpretation. An experimenter may conclude that the task or objective function the DL model was trained on closely resembles a task the brain solves, when this may not be the case. For example, one might falsely infer that the brain does predictive coding for language because it is well predicted by features from a language model that is trained to predict the next word in a sequence. Guest and Martin (2023) elaborate on this issue by discussing the logical fallacies in inferences drawn between brain behavior or activity, and DL models of language. Specifically, they highlight that studies analyzing parallels between the brain and computational models of language often attribute inappropriate causality by assuming that predictive ability is sufficient to claim task similarity or model equivalence. On the contrary, the direction of causality should be that if an artificial model closely resembles the brain, it can mimic brain behavior and activity, or that a lack of prediction abilities clearly indicates a lack of model equivalence. This is a pertinent issue for in silico experimentation as the paradigm uses computational models of language processing in the brain to simulate its behavior. However, it is important to note that in all of the in silico examples presented here, the authors were using the generalizability of the encoding models to predict brain responses in different conditions. This only suggests that the encoding models can effectively capture the brain's behavior for language tasks but is not a sufficient account to conclude model equivalence.

Another issue with logical inference in DL-based encoding models relates to the functional equivalence of two brain regions that are both well predicted by a given feature space. In their recent study, Toneva et al. (2022) discuss this issue in detail for language encoding models and provide a computational framework to analyze the extent to which brain regions share computational mechanisms solely based on their encoding performance.

The three main sources of confounds—naturalistic stimuli, DL-based feature spaces, and modeling assumptions—can intersect in interesting ways and raise the probability of incorrect interpretation. False causality stemming from spurious correlations are a problem for in silico experiments, much like controlled experiments. To this end, it is important to emphasize transparent analysis methods, better interpretability tools for DL models, and rigorous tests of reproducibility with diverse data sources.

Although reproducibility is traditionally viewed as the replication of effects across participant pools/data sets, with in silico experimentation we can add another layer of replicability, across different models that have different intrinsic biases (architectures, training objectives, etc.), and learn different types of representation.

Important Factors of Consideration

Before doing in silico experimentation, one important consideration is determining if the encoding model is "good enough." While there is no quantitative threshold above which a model can be considered suitable, we suggest the following.

Statistical significance of encoding model performance on a diverse, held-out test set

It is imperative that experimenters test whether encoding model performance is statistically significant at the individual brain element level. Any in silico experimentation should only be done on brain elements that are significantly predicted by the computational model. A well-established approach in language encoding modes is to correlate the true and predicted responses for a held-out test set. Following this, a permutation test can be done to check if the correlation is statistically significant.

We also emphasize the importance of using diverse test sets to effectively gauge generalization. If a brain element is selective for features that are not present in the test set, then it may falsely be labeled as poorly predicted. One feasible solution is to use a leave-one-out testing procedure. This can be done by fitting an ensemble of encoding models, each of which excludes one unique set of training data in the model estimation. Statistical significance can then be measured for encoding model predictions on all held-out data. This procedure increases diversity in the test set and improves statistical power.

Feature-space selection

Given the diversity of function across the brain, it is possible that no one feature space or computational model best predicts all brain regions. Thus, experimenters should test several different features spaces and models (Nunez-Elizalde et al., 2019) and individually choose the one with best held-out set performance for each brain element. This is especially important for DL models as different neural language models or their layers predict different brain regions well. In this case, we would use the neural language model (layer) that best predicts held-out stories for each element and, further, passes construct validity tests (ie., has well-understood behavior to the controlled manipulation). For example, in the in silico semantic composition experiment, Jain and Huth (2023) found that the lower layers of the neural language model were generally indifferent to ablating words farther in the past (Khandelwal et al., 2018). Consequently, these layers cannot be used to conduct the ablation study, as they do not respond to the manipulation in the first place.

Interpreting the DL-models

To establish the validity of computational model constructs, we suggested the use of interpretability tools and techniques to understand how the DL-model itself represents a cognitive process. This would allow the experimenter to directly investigate sources of confounds.

It is also important to consider the types of questions the in silico paradigm is most suited to answer. As demonstrated here, this paradigm can be used to estimate functional properties in the brain, such as selectivity to different word categories or the processing timescale. It cannot, however, be used to test the causal involvement of a brain area or the exact computational mechanism. For example, many regions in the experiments above are shown to capture semantic properties in language. Whether these regions play a causal role in semantic tasks, can only be determined by an in vivo measurement.

Conclusion and Future Directions

In this article, we highlight the promises of in silico experimentation and detail how it brings together the advantages of controlled experiments and naturalistic experiments paired with encoding models. We showcase four different in silico experiments that all rely on naturalistic language experiments to simulate four different previous studies. We survey the advantages and potential caveats of in silico experimentation and highlight how it can take advantage of recent work in DL to simulate experiments with diverse types of language stimuli.

Current work on DL-based encoding models for language is largely restricted to self-supervised models. This is expected since self-supervised models have been trained on large amounts of data and consequently learn highly useful and transferable linguistic representations. However, it remains to be seen if task-based experimental designs in neuroscience can be simulated and adapted with more goal-directed artificial language networks. Additionally, it is also important to investigate and characterize which types of neuroscientific results can be explored with self-supervised models and what aspects of language meaning are beyond the scope of the next-word-prediction objective.

Lastly, DL-based language encoding models rely on feature extraction from language or speech ANNs (linearizing transform) and learn a linear function atop the features. We believe that the in silico paradigm can become more powerful if language encoding models directly update the parameters of the ANN itself, resulting in an end-to-end system. While this has been popularized in vision (e.g., Bashivan et al., 2019), it is yet to be explored for language. This approach can potentially introduce diversity into the computational mechanisms of the ANNs, such as recurrence, linear readout from a memory store, and so forth, to integrate processing in different brain structures (hippocampus, cortex, etc.). This could allow us to understand parallel mechanisms like linguistic function, working memory access, and attention using this same approach.

ACKNOWLEDGMENTS

Research reported in this article was also supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health as part of the Collaborative Research in Computational Neuroscience (CRCNS) program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Nicole Beckage and Javier Turek for useful discussions on this work.

FUNDING INFORMATION

Shailee Jain, Foundations of Language Fellowship, William Orr Dingwall Foundation. Alexander G. Huth, Burroughs Wellcome Fund (https://dx.doi.org/10.13039/100000861). Alexander G. Huth, Intel Corporation (https://dx.doi.org/10.13039/100002418). Alexander G. Huth, National Institute on Deafness and Other Communication Disorders (https://dx.doi.org/10.13039/100000055), Award ID: R01DC020088. Leila Wehbe, National Institute on Deafness and Other Communication Disorders (https://dx.doi.org/10.13039/100000055), Award ID: R01DC020088.

AUTHOR CONTRIBUTIONS

Shailee Jain: Conceptualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead. **Vy A. Vo**: Conceptualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting. **Leila Wehbe**: Conceptualization: Supporting; Funding

acquisition: Equal; Supervision: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting. **Alexander G. Huth**: Conceptualization: Supporting; Funding acquisition: Equal; Supervision: Lead; Writing – original draft: Supporting; Writing – review & editing: Supporting.

REFERENCES

Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019). Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 191–203). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4820

Aharoni, R., & Goldberg, Y. (2020). Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 7747–7763). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.692

Allison, J. (Producer). (2009–). *The moth radio hour* [Radio program]. The Moth; Atlantic Public Media/PRX.

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Grimm, S., & Lalor, E. C. (2021). Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, *41*(18), 4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021, PubMed: 33753548

Antonello, R., & Huth, A. (2024). Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*, *5*(1), 64–79. https://doi.org/10.1162/nol_a_00087

Antonello, R., Turek, J. S., Vo, V. A., & Huth, A. (2021). Low-dimensional structure in the space of language representations is reflected in brain responses. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems*. NeurIPS. https://openreview.net/forum?id=UYI6Sk_3Nox

Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2014). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, *24*(8), 1988–1995. https://doi.org/10.1093/cercor/bht046, PubMed: 23463339

Aurnhammer, C., & Frank, S. L. (2018). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. *PsyArXiv*. https://doi.org/10.31234/osf.io/wec74

Barch, D. M., & Yarkoni, T. (2013). Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 687–689. https://doi.org/10.3758/s13415-013-0201-7, PubMed: 23922199

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), Article eaav9436. https://doi.org/10.1126/science.aav9436, PubMed: 31048462

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*, *31*(8), 2801–2814. https://doi.org/10.1523/JNEUROSCI.5003-10.2011, PubMed: 21414902

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bevilacqua, D., Davidesco, I., Wan, L., Chaloner, K., Rowland, J., Ding, M., Poeppel, D., & Dikker, S. (2019). Brain-to-brain synchrony and learning outcomes vary by student–teacher dynamics: Evidence from a real-world classroom electroencephalography study. *Journal of Cognitive Neuroscience*, *31*(3), 401–411. https://doi.org/10.1162/jocn_a_01274, PubMed: 29708820

Bhattasali, S., Brennan, J., Luh, W.-M., Franzluebbers, B., & Hale, J. (2020). The Alice datasets: fMRI & EEG observations of natural language comprehension. In *Proceedings of the 12th language resources and evaluation conference* (pp. 120–125). European Language Resources Association. https://aclanthology.org/2020.lrec-1.15

Bhattasali, S., Fabre, M., Luh, W.-M., Al Saied, H., Constant, M., Pallier, C., Brennan, J. R., Spreng, R. N., & Hale, J. (2019). Localising memory retrieval and syntactic composition: An fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, *34*(4), 491–510. https://doi.org/10.1080/23273798.2018.1518533

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055, PubMed: 19329570

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, *17*(6), 905–917. https://doi.org/10.1162/0898929054021102, PubMed: 16021798

Blank, I. A., & Fedorenko, E. (2017). Domain-general brain regions do not track linguistic input as closely as language-selective regions. *Journal of Neuroscience*, *37*(41), 9999–10011. https://doi.org/10.1523/JNEUROSCI.3642-16.2017, PubMed: 28871034

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, Article 116925. https://doi.org/10.1016/j.neuroimage.2020.116925, PubMed: 32407994

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, *5*(1), Article 12881. https://doi.org/10.1038/srep12881, PubMed: 26242909

Boylan, C., Trueswell, J. C., & Thompson-Schill, S. L. (2015). Compositionality and the angular gyrus: A multi-voxel similarity analysis of the semantic composition of nouns and verbs. *Neuropsychologia*, *78*, 130–141. https://doi.org/10.1016/j.neuropsychologia.2015.10.007, PubMed: 26454087

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173. https://doi.org/10.1016/j.bandl.2010.04.002, PubMed: 20472279

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475, PubMed: 23571845

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *Findings of the Association for Computational Linguisitcs: EMNLP 2021* (pp. 3635–3644). Association for Computational Linguistics. https://hal.archives-ouvertes.fr/hal-03361430. https://doi.org/10.18653/v1/2021.findings-emnlp.308

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), Article 134. https://doi.org/10.1038/s42003-022-03036-1, PubMed: 35173264

Cavanagh, S. E., Hunt, L. T., & Kennerley, S. W. (2020). A diversity of intrinsic timescales underlie neural computations. *Frontiers in Neural Circuits*, *14*, 615626. https://doi.org/10.3389/fncir.2020.615626, PubMed: 33408616

Chan, A. H. D., Luke, K.-K., Li, P., Yip, V., Li, G., Weekes, B., & Tan, L. H. (2008). Neural correlates of nouns and verbs in early bilinguals. *Annals of the New York Academy of Sciences*, *1145*(1), 30–40. https://doi.org/10.1196/annals.1416.000, PubMed: 19076387

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432. https://doi.org/10.1038/nn.2641, PubMed: 20890293

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. https://doi.org/10.1038/nn.4450, PubMed: 27918531

Chen, Z., Chen, S., Wu, Y., Qian, Y., Wang, C., Liu, S., Qian, Y., & Zeng, M. (2022). Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022—IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6147–6151). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9747814

Chien, H.-Y. S., & Honey, C. J. (2020). Constructing and forgetting temporal context in the human cerebral cortex. *Neuron*, *106*(4), 675–686. https://doi.org/10.1016/j.neuron.2020.02.013, PubMed: 32164874

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 276–286). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4828

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 2126–2136). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1198

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770. https://doi.org/10.1038/nn.3381, PubMed: 23603707

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017, PubMed: 28588065

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, *39*(39), 7722–7736. https://doi.org/10.1523/JNEUROSCI.0675-19.2019, PubMed: 31427396

Deniz, F., Tseng, C., Wehbe, L., & Gallant, J. L. (2021). Semantic representations during language comprehension are affected by context. *bioRxiv*. https://doi.org/10.1101/2021.12.15.472839

Ellis, C. T., Baldassano, C., Schapiro, A. C., Cai, M. B., & Cohen, J. D. (2020). Facilitating open-science with realistic fMRI simulation: Validation and application. *PeerJ*, *8*, Article e8564. https://doi.org/10.7717/peerj.8564, PubMed: 32117629

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298

Ettinger, A., Elgohary, A., Phillips, C., & Resnik, P. (2018). Assessing composition in sentence vector representations. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1790–1801). Association for Computational Linguistics. https://aclanthology.org/C18-1152

Evans, S. (2017). What has replication ever done for us? Insights from neuroimaging of speech perception. *Frontiers in Human Neuroscience*, *11*, 41. https://doi.org/10.3389/fnhum.2017.00041, PubMed: 28203154

Fedorenko, E. (2021). The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Current Opinion in Behavioral Sciences*, *40*, 105–112. https://doi.org/10.1016/j.cobeha.2021.02.023

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010, PubMed: 20410363

Friederici, A. D., Opitz, B., & von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: An fMRI investigation of different word types. *Cerebral Cortex*, *10*(7), 698–705. https://doi.org/10.1093/cercor/10.7.698, PubMed: 10906316

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. 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. 32–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–197. https://doi.org/10.1038/72140, PubMed: 10649576

Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLOS ONE*, *10*(3), Article e0120068. https://doi.org/10.1371/journal.pone.0120068, PubMed: 25793289

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, S. C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2021). Thinking ahead: Spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv*. https://doi.org/10.1101/2020.12.02.403477

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)* (pp. 10–18). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0102

Graves, W. W., Binder, J. R., Desai, R. H., Conant, L. L., & Seidenberg, M. S. (2010). Neural correlates of implicit and explicit combinatorial semantic processing. *NeuroImage*, *53*(2), 638–646. https://doi.org/10.1016/j.neuroimage.2010.06.055, PubMed: 20600969

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*, 213–227. https://doi.org/10.1007/s42113-022-00166-x

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (Volume 1: Long papers)* (pp. 1195–1205). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1108

Haber, J., & Poesio, M. (2021). Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2663–2676). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.226

Hamilton, L. S., & Huth, A. G. (2018). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946, PubMed: 32656294

Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, *21*(4), 1639–1651. https://doi.org/10.1016/j.neuroimage.2003.11.029, PubMed: 15050587

Hasson, U., Avidan, G., Gelbard, H., Vallines, I., Harel, M., Minshew, N., & Behrmann, M. (2009). Shared and idiosyncratic cortical activation patterns in autism revealed under continuous real-life viewing conditions. *Autism Research*, *2*(4), 220–231. https://doi.org/10.1002/aur.89, PubMed: 19708061

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008, PubMed: 18322098

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9, PubMed: 14741110

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *ELife*, *9*, Article e56601. https://doi.org/10.7554/eLife.56601, PubMed: 32484439

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1, PubMed: 28726177

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP–IJCNLP)* (pp. 2733–2743). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1275

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. 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. 4129–4138). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1419

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637, PubMed: 27121839

Huth, A. G., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2015). PrAGMATiC: A probabilistic and generative model of areas tiling the cortex. *arXiv*. https://doi.org/10.48550/arXiv.1504.03622

Ince, R. A., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *ELife*, *10*, Article e62461. https://doi.org/10.7554/eLife.62461, PubMed: 34612811

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fMRI. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (10 pp.). NeurIPS.

Jain, S., & Huth, A. G. (2023). *Discovering distinct patterns of semantic integration across cortex using natural language encoding models for fMRI* [Manuscript in preparation]. Departments of Computer Science & Neuroscience, University of Texas at Austin.

Jain, S., Vo, V. A., Mahto, S., LeBel, A., Turek, J. S., & Huth, A. (2020). Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33* (pp. 13738–13749). NeurIPS.

Kable, J. W., Lease-Spellmeyer, J., & Chatterjee, A. (2002). Neural substrates of action event knowledge. *Journal of Cognitive Neuroscience*, *14*(5), 795–805. https://doi.org/10.1162/08989290260138681, PubMed: 12167263

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997, PubMed: 9151747

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644. https://doi.org/10.1016/j.neuron.2018.03.044, PubMed: 29681533

Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2018). Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (pp. 284–294). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1027

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2022). Reconstructing the cascade of language processing in the

brain using the internal computations of a transformer-based language model. *bioRxiv*. https://doi.org/10.1101/2022.06.08.495348

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0, PubMed: 6690995

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. 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. 11–20). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1002

LeBel, A., Jain, S., & Huth, A. G. (2021). Voxelwise encoding models show that cerebellar language representations are highly conceptual. *Journal of Neuroscience*, *41*(50), 10341–10355. https://doi.org/10.1523/JNEUROSCI.0118-21.2021, PubMed: 34732520

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., Tang, J., Xu, L., & Huth, A. G. (2022). A natural language fMRI dataset for voxelwise encoding models. *bioRxiv*. https://doi.org/10.1101/2022.09.22.509104

Lerner, Y., Honey, C. J., Katkov, M., & Hasson, U. (2014). Temporal scaling of neural responses to compressed and dilated natural speech. *Journal of Neurophysiology, 111*(12), 2433–2444. https://doi.org/10.1152/jn.00497.2013, PubMed: 24647432

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011, PubMed: 21414912

Levinson, S. C. (2016). Turn-taking in human communication: Origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010, PubMed: 26651245

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (Volume 1: Long papers)* (pp. 1813–1827). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.143

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., Brennan, J. R., Yang, Y., Pallier, C., & Hale, J. (2022). *Le Petit Prince* multilingual naturalistic fMRI corpus. *Scientific Data*, *9*(1), Article 530. https://doi.org/10.1038/s41597-022-01625-7, PubMed: 36038567

Li, Y., Anumanchipalli, G. K., Mohamed, A., Lu, J., Wu, J., & Chang, E. F. (2022). Dissecting neural computations of the human auditory pathway using deep neural networks for speech. *bioRxiv*. https://doi.org/10.1101/2022.03.14.484195

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5210–5217). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.465

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. *arXiv*. https://doi.org/10.48550/arXiv.1807.06882

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: An fMRI study. *Journal of Cognitive Neuroscience*, *22*(1), 203–211. https://doi.org/10.1162/jocn.2009.21203, PubMed: 19302006

Magyari, L., Bastiaansen, M. C. M., de Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*, *26*(11), 2530–2539. https://doi.org/10.1162/jocn_a_00673, PubMed: 24893743

Mahowald, K., Kachergis, G., & Frank, M. C. (2020). What counts as an exemplar model, anyway? A commentary on Ambridge (2020). *First Language*, *40*(5–6), 608–611. https://doi.org/10.1177/0142723720905920

Mahto, S., Vo, V. A., Turek, J. S., & Huth, A. (2020). Multi-timescale representation learning in LSTM language models. *OpenReview.net*. https://openreview.net/forum?id=9lTXiTrAoT

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

Matusz, P. J., Dikker, S., Huth, A. G., & Perrodin, C. (2019). Are we ready for real-world neuroscience? *Journal of Cognitive Neuroscience*, *31*(3), 327–338. https://doi.org/10.1162/jocn_e_01276, PubMed: 29916793

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 12–22). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.2

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv*. https://doi.org/10.48550/arXiv.2206.01685

Millet, J., & King, J.-R. (2021). Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv*. https://doi.org/10.48550/arXiv.2103.01032

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195. https://doi.org/10.1126/science.1152876, PubMed: 18511683

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., ... Hasson, U. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1), Article 250. https://doi.org/10.1038/s41597-021-01033-3, PubMed: 34584100

Nayebi, A., Attinger, A., Campbell, M. G., Hardcastle, K., Low, I. I. C., Mallory, C. S., Mel, G. C., Sorscher, B., Williams, A. H., Ganguli, S., Giocomo, L. M., & Yamins, D. L. K. (2021). Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *bioRxiv*. https://doi.org/10.1101/2021.10.30.466617

Noppeney, U., Josephs, O., Kiebel, S., Friston, K. J., & Price, C. J. (2005). Action selectivity in parietal and temporal cortex. *Cognitive Brain Research*, *25*(3), 641–649. https://doi.org/10.1016/j.cogbrainres.2005.08.017, PubMed: 16242924

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, *197*, 482–492. https://doi.org/10.1016/j.neuroimage.2019.04.012, PubMed: 31075394

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911. https://doi.org/10.1038/nn.4021, PubMed: 25984889

Pandia, L., & Ettinger, A. (2021). Sorting through the noise: Testing robustness of information processing in pre-trained language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1583–1596). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.119

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401, PubMed: 26168109

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Poldrack, R. A., Whitaker, K., & Kennedy, D. (2020). Introduction to the special issue on reproducibility in neuroimaging. *NeuroImage*, *218*, Article 116357. https://doi.org/10.1016/j.neuroimage.2019.116357, PubMed: 31733374

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, *24*(11), 1628–1636. https://doi.org/10.1038/s41593-021-00921-6, PubMed: 34711960

Prasad, G., van Schijndel, M., & Linzen, T. (2019). Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)* (pp. 66–76). Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1007

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* [Preprint]. Papers With Code.

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, *12*(1), 5540. https://doi.org/10.1038/s41467-021-25409-6, PubMed: 34545079

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 7237–7256). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.647

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, Article 116392. https://doi.org/10.1016/j.neuroimage.2019.116392, PubMed: 31770637

Reddy, A. J., & Wehbe, L. (2020). Can fMRI reveal the representation of syntactic structure in the brain? *bioRxiv*. https://doi.org/10.1101/2020.06.16.155499

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, *33*(40), 15978–15988. https://doi.org/10.1523/JNEUROSCI.1580-13.2013, PubMed: 24089502

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), Article e2105646118. https://doi.org/10.1073/pnas.2105646118, PubMed: 34737231

Scott, S. K. (2019). From speech and talkers to the social world: The neural processing of human spoken language. *Science*, *366*(6461), 58–62. https://doi.org/10.1126/science.aax0288, PubMed: 31604302

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466, PubMed: 27386919

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain: A Journal of Neurology*, *115*(1), 15–36. https://doi.org/10.1093/brain/115.1.15, PubMed: 1559150

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, Article 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307, PubMed: 31874149

Sievers, B., Welker, C., Hasson, U., Kleinbaum, A. M., & Wheatley, T. (2020). How consensus-building conversation changes our minds and aligns our brains. *PsyArXiv*. https://doi.org/10.31234/osf.io/562z7

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632, PubMed: 22006061

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning (Volume 70)* (pp. 3319–3328). Association for Computing Machinery.

Suzanne Scherf, K., Behrmann, M., Minshew, N., & Luna, B. (2008). Atypical development of face and greeble recognition in autism. *Journal of Child Psychology and Psychiatry*, *49*(8), 838–847. https://doi.org/10.1111/j.1469-7610.2008.01903.x, PubMed: 18422548

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv*. https://doi.org/10.48550/arXiv.1905.05950

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., & Pavlick, E. (2018). *What do you learn from context? Probing for sentence structure in contextualized word representations*. International Conference on Learning Representations. https://openreview.net/forum?id=SJzSgnRcKX

Toneva, M., Mitchell, T. M., & Wehbe, L. (2020). Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv*. https://doi.org/10.1101/2020.09.28.316935

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*. NeurIPS.

Toneva, M., Williams, J., Bollu, A., Dann, C., & Wehbe, L. (2022). Same cause; different effects in the brain. *Proceedings of the First Conference on Causal Learning and Reasoning*, *177*, 787–825.

Vaidya, A. R., Jain, S., & Huth, A. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. *Proceedings of the 39th International Conference on Machine*

Learning, *162*, 21927–21944. https://proceedings.mlr.press/v162/vaidya22a.html

van der Wees, M., Bisazza, A., & Monz, C. (2017). Dynamic data selection for neural machine translation. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1400–1410). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1147

Vo, V. A., Jain, S., Beckage, N., Chien, H.-Y. S., Obinwa, C., & Huth, A. G. (2023). *A unifying computational account of temporal processing in natural speech across cortex* [Manuscript in preparation]. Departments of Computer Science & Neuroscience, University of Texas at Austin.

Wallentin, M., Østergaard, S., Lund, T. E., Østergaard, L., & Roepstorff, A. (2005). Concrete spatial language: See what I mean? *Brain and Language*, *92*(3), 221–233. https://doi.org/10.1016/j.bandl.2004.06.106, PubMed: 15721955

Wang, A., Tarr, M., & Wehbe, L. (2019). Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*. NeurIPS. https://proceedings.neurips.cc/paper/2019/hash/f490c742cd8318b8ee6dca10af2a163f-Abstract.html

Wang, S., Zhang, J., Lin, N., & Zong, C. (2020). Probing brain activation patterns by dissociating semantics and syntax in sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5), 9201–9208. https://doi.org/10.1609/aaai.v34i05.6457

Wang, S., Zhang, X., Zhang, J., & Zong, C. (2022). A synchronized multimodal neuroimaging dataset for studying brain language processing. *Scientific Data*, *9*(1), Article 590. https://doi.org/10.1038/s41597-022-01708-5, PubMed: 36180444

Wehbe, L., Huth, A. G., Deniz, F., Gao, J., Kieseler, M.-L., & Gallant, J. L. (2018). BOLD predictions: Automated simulation of fMRI experiments [Poster]. 2018 Conference on Cognitive Computational Neuroscience, Philadelphia, Pennsylvania.

Wehbe, L., Huth, A. G., Deniz, F., Gao, J., Kieseler, M.-L., & Gallant, J. L. (2021). *BOLDpredictions* [Software]. https://github.com/boldprediction

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, *9*(11), Article e112575. https://doi.org/10.1371/journal.pone.0112575, PubMed: 25426840

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 233–243). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1030

Westfall, J., Nichols, T. E., & Yarkoni, T. (2017). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*, *1*, 23. https://doi.org/10.12688/wellcomeopenres.10298.2, PubMed: 28503664

Wilcox, E., Vani, P., & Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long papers)* (pp. 939–952). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.76

Wu, A., Wang, C., Pino, J., & Gu, J. (2020). Self-supervised representations improve end-to-end speech translation. *Interspeech 2020*, 1491–1495. https://doi.org/10.21437/Interspeech.2020-3094

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024, PubMed: 16776594

Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. *Cerebral Cortex*, *15*(8), 1234–1242. https://doi.org/10.1093/cercor/bhi006, PubMed: 15677350

Yamins, D. L., & DiCarlo, J. J. (2016). Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, *37*, 114–120. https://doi.org/10.1016/j.conb.2016.02.001, PubMed: 26921828

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685, PubMed: 33342451

Yeshurun, Y., Nguyen, M., & Hasson, U. (2017). Amplification of local changes along the timescale processing hierarchy. *Proceedings of the National Academy of Sciences*, *114*(35), 9475–9480. https://doi.org/10.1073/pnas.1701652114, PubMed: 28811367

Zhang, X., Wang, S., Lin, N., Zhang, J., & Zong, C. (2022). Probing word syntactic representations in the brain by a feature elimination method [Poster]. *Proceedings of the 36th AAAI conference on artificial intelligence*. Association for the Advancement of Artifical Intelligence. https://aaai-2022.virtualchair.net/poster_aaai7935