ANNUAL REVIEWS

Annual Review of Neuroscience

Language in Brains, Minds, and Machines

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko

Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; email: evelina9@mit.edu

Annu. Rev. Neurosci. 2024. 47:277–301

The *Annual Review of Neuroscience* is online at neuro.annualreviews.org

https://doi.org/10.1146/annurev-neuro-120623-101142

Keywords

Language, artificial language models, natural language processing, neuroimaging, cognitive neuroscience

Abstract

It has long been argued that only humans could produce and understand language. But now, for the first time, artificial language models (LMs) achieve this feat. Here we survey the new purchase LMs are providing on the question of how language is implemented in the brain. We discuss why, a priori, LMs might be expected to share similarities with the human language system. We then summarize evidence that LMs represent linguistic information similarly enough to humans to enable relatively accurate brain encoding and decoding during language processing. Finally, we examine which LM properties—their architecture, task performance, or training—are critical for capturing human neural responses to language and review studies using LMs as in silico model organisms for testing hypotheses about language. These ongoing investigations bring us closer to understanding the representations and processes that underlie our ability to comprehend sentences and express thoughts in language.

Contents

1. THE HUMAN LANGUAGE SYSTEM

1.1. Language as a Distinct Component of the Mind and Brain

Right now, human beings all over the world are translating the thoughts in their minds into sequences of sounds that travel from their mouths into a fellow human's ears, producing thoughts in their fellow's mind similar to those that began in their own. This is the everyday miracle of language, the engine of cumulative human culture and the signature talent of our species. But what is language, and how is it computed in the mind and brain? How can something as abstract as the meaning of a sentence be encoded in the activity of neurons? And what is the relationship between language and thought? These questions, long pondered by philosophers, are suddenly yielding to rigorous empirical investigation with an ever-accelerating pace of synergistic discoveries in cognitive science, neuroscience, and artificial intelligence (AI).

1.1.1. Selectivity of the language system. By language, we refer not to the surface form of speech or text or sign but the more abstract representations common to all these modalities—the representations that allow for the mapping between thoughts and word sequences. In the earliest efforts to identify brain regions engaged in language processing, nineteenth-century neurologists described patients with deficits in speaking and in understanding language that resulted from damage to the frontal and temporal lobes. However, many of these cases reflected deficits in the perception or production of speech rather than language (Luria 1970, Goodglass 1993; for a recent discussion, see E. Fedorenko, S. Piantadosi & E. Gibson, unpublished manuscript), and heated

debates have raged ever since on the question of whether any brain regions are specifically engaged in language per se.

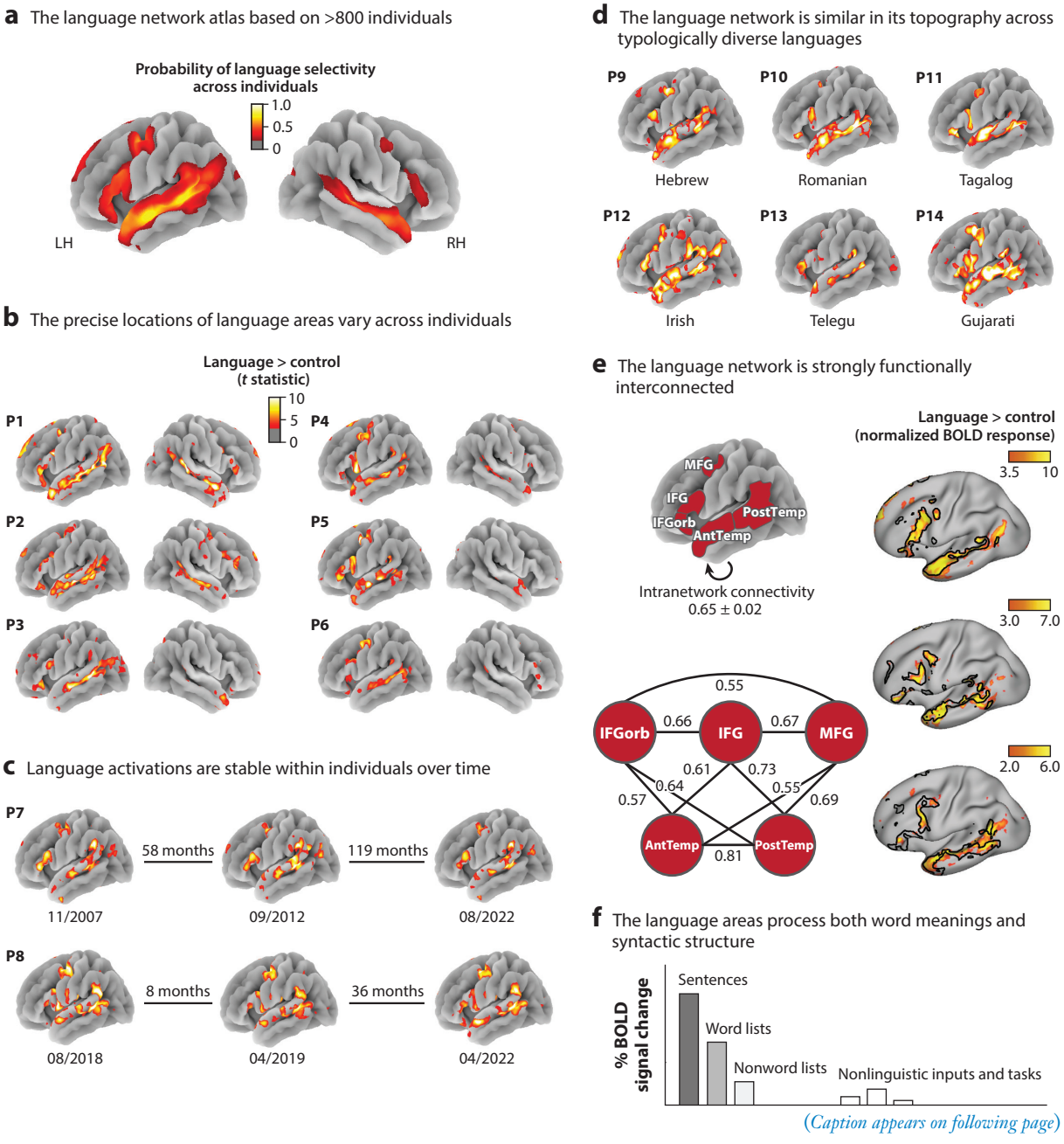
Indeed, when noninvasive neuroimaging methods first became available, many researchers noted that the brain regions in the temporal and frontal lobes that became active in positron emission tomography and functional MRI (fMRI) studies when people understand sentences resembled brain regions engaged in other, nonlinguistic tasks (Dehaene et al. 1999, Levitin & Menon 2003, Novick et al. 2005; for reviews, see Fedorenko & Varley 2016, Fedorenko & Blank 2020). These findings were taken to mean that brain regions engaged in language processing were not specific for language but instead supported a variety of cognitive functions. However, this reasoning suffered from a critical flaw: Because the exact anatomical location of functional regions varies across individuals, analyses that pool data across individuals in a common anatomical brain space necessarily blur functional responses and thus underestimate functional specificity (Saxe et al. 2006). When new methods were developed that functionally identify language regions individually in each participant with localizer tasks contrasting responses to sentences versus strings of nonwords or degraded speech (Fedorenko et al. 2010), it became clear that these regions are highly specific for language and show little response when people perform mental arithmetic, listen to music, hold information in working memory, or exert cognitive control (e.g., Fedorenko et al. 2011, Monti et al. 2012, Amalric et al. 2018, Chen et al. 2023). Other studies tested mental functions even closer to language, including nonverbal semantics (Ivanova et al. 2021), logical reasoning (Monti et al. 2009), understanding computer code (Ivanova et al. 2020, Liu et al. 2020), processing nonverbal communicative signals (Deen et al. 2015, Jouravlev et al. 2019), and reasoning about others' minds (Shain et al. 2023b), and found that even those do not strongly engage the language brain areas.

Complementary evidence for the dissociation of language and nonlinguistic cognition, including thinking and reasoning, comes from studies of patients with global aphasia due to massive left-hemisphere strokes (Luria 1970, Goodglass 1993). These patients, who lack virtually all ability to produce or understand language, are nonetheless able to solve logic and arithmetic problems, appreciate music, hold information in working memory, and think about what other people are thinking (Varley et al. 2005, Apperly et al. 2009). Thus, not only are language and thought dissociable in the brain, but many aspects of thought can proceed in the near absence of language. Taken together, these findings show that the language regions play little role in nonlinguistic tasks, even those that share similarities with language.

1.1.2. Anatomy and the internal structure of the language system. Anatomically, the brain's language system stretches across many square centimeters of cortex and encompasses cortical areas on the lateral surface of the frontal and temporal lobes (**Figure 1**). In most individuals, this system is lateralized to the left hemisphere, with weaker activations in homotopic right-hemisphere regions (**Figure 1*a*,*b***), and the topography is stable within individuals over time (**Figure 1*c***) and similar across typologically diverse languages (**Figure 1*d***).

What division of labor exists across the brain regions that make up the language system? According to one classic proposal, distinct regions support language comprehension versus language production (e.g., Geschwind 1970), based on apparent dissociations in patients with linguistic deficits. However, these dissociations likely pertain to lower-level speech-perception and speech-articulation abilities, which are distinct from higher-level comprehension and production abilities (e.g., see Saussure 1959; for a recent review, see E. Fedorenko, A.A. Ivanova & T.I. Regev, unpublished manuscript). Indeed, fMRI studies that isolate higher-level linguistic components from lower-level speech components find strongly overlapping responses between sentence comprehension and sentence production (Menenti et al. 2011, Hu et al. 2023). Another influential idea

held that the inferior frontal language area is especially important for, and perhaps selectively engaged in, processing syntactic structure (e.g., see Hagoort 2005, Grodzinsky & Santi 2008, Friederici 2012). However, the evidence from patients with syntactic difficulties (so-called agrammatic aphasia) is complex and heterogeneous (e.g., see Badecker & Caramazza 1985, Berndt 1991) and does not support a selective role of the inferior frontal language component in syntactic



(Caption appears on following page)

Figure 1 (*Figure appears on preceding page*)

The neuroanatomy of language processing. (*a*) A probabilistic atlas for the language network based on overlaying activation maps [here and elsewhere obtained with functional MRI (fMRI)] from $n = 806$ participants who performed a language localizer task (Lipkin et al. 2022). The atlas is displayed on lateral views (*left* and *right*) of the brain; yellow areas indicate higher overlap across individuals. The language network includes, most prominently, left-lateralized areas in the frontal and left temporal lobe, and this topography is broadly similar across individuals. Panel *a* adapted from Lipkin et al. (2022). (*b*) Sample language activation maps from $n = 6$ native speakers of English. The variability in the locations, shapes, and sizes of the language areas shows why it is difficult to make inferences about the language system in analyses that average activations across individuals in a common space and assume voxel-wise correspondence (Fedorenko et al. 2010). The color scale reflects the *t* statistic for the language versus control condition contrast (the same color scale is used in panels *c* and *d*). (*c*) Sample language activation maps from the left hemisphere of $n = 2$ native speakers of English each tested three times, including across the span of ∼15 years) (Mahowald & Fedorenko 2016). (*d*) Sample language activation maps from $n = 6$ native speakers of six languages across four language families: Hebrew (language family: Afro-Asiatic), Romanian (language family: Indo-European/Italic), Tagalog (language family: Austroasiatic), Irish (language family: Indo-European/Celtic), Telegu (language family: Dravidian), and Gujarati (language family: Indo-European/Indo-Iranian). Although language activations vary in their precise topography, the variability among the speakers of different languages does not exceed the variability observed among the speakers of the same language (Malik-Moraleda et al. 2022). Panel *d* adapted from Malik-Moraleda et al. (2022). (*e*, *left*) A schematic illustration of the language regions. The red masks correspond to areas within which most individuals show responses during language processing (Fedorenko et al. 2010, Lipkin et al. 2022). The schematic shows the average correlations in neural activity fluctuations (measured with the BOLD signal in fMRI) among the language regions (individually defined) during naturalistic cognition paradigms (data from Malik-Moraleda et al. 2022; $n = 82$ speakers of diverse languages tested in their native language). The average within-network correlation during a story comprehension condition is $r = 0.65$; pairwise region-to-region correlations are shown in the network schematic (*circles* indicate regions; e.g., the average correlation between the PostTemp region and the IFG language functional regions of interest is $r = 0.73$). (*Right*) The topography of the language network identified with a language localizer paradigm (Fedorenko et al. 2010) (*yellow* and *orange*) can be recovered from a large amount of naturalistic resting state (no task) data collected in individual participants by analyzing patterns of voxel cofluctuation (the language network recovered in this way is shown in *black outlines*), as shown in $n = 3$ sample individuals. Right side of panel *e* adapted with permission from Braga et al. (2020) (CC BY 4.0). (*f*) Schematic of BOLD response magnitudes to sentences, word lists, nonword lists, and nonlinguistic conditions in the language areas. Structured stimuli, such as sentences, that convey compositional meanings elicit a stronger response than unstructured ones, such as lists of words, that only express individual lexical meanings; lists of words, in turn, elicit a stronger response than lists of meaningless nonwords (e.g., Fedorenko et al. 2010, Pallier et al. 2011, Shain et al. 2023a). Diverse nonlinguistic inputs and tasks elicit little or no response in the language areas in spite of strongly activating other brain areas (e.g., see Monti et al. 2009, 2012; Fedorenko et al. 2011; Ivanova et al. 2020; Chen et al. 2023). Abbreviations: AntTemp, anterior temporal; BOLD, blood-oxygen-level-dependent; IFG, inferior frontal gyrus; IFGorb, inferior frontal gyrus orbital; LH, left hemisphere; MFG, middle frontal gyrus; PostTemp, posterior temporal; RH, right hemisphere.

processing (Fedorenko et al. 2022). Furthermore, fMRI studies have shown that every language region is strongly sensitive to syntactic structure building (e.g., see Bautista & Wilson 2016, Blank et al. 2016, Shain et al. 2022a), arguing against a focal syntactic hub, and every language region responds at least as strongly to word meaning (Fedorenko et al. 2010, 2020; Shain et al. 2023a) (**Figure 1*f***). Other divisions of labor within the language network have been proposed in the past (e.g., see Hickok & Poeppel 2007, Price 2010, Friederici 2012); however, none of the claims about dissociations among the high-level language regions (cf. those between speech and language regions) withstand empirical scrutiny. Thus, current evidence does not support areal subdivisions within the cortical language system, in line with strong interregional functional connectivity during naturalistic cognition (Blank et al. 2014, Braga et al. 2020, Malik-Moraleda et al. 2022) (**Figure 1*e***). That said, the posterior temporal component may be overall more important for language function based on evidence from aphasia (e.g., see Luria 1970, Wilson et al. 2023), and some heterogeneity exists in the form of spatially interdigitated neural populations (e.g., see Jain et al. 2020, Regev et al. 2023).

The findings reviewed so far indicate that the brain's language system constitutes a distinct component of the mind and brain that is specific for language processing, separable from other cognitive systems, and relatively functionally homogeneous across its regions, with every region supporting computations related to accessing word meanings and combinatorial (syntactic and

semantic) processing. These considerations suggest that the language system is a natural kind—an ontologically meaningful grouping of brain areas—that can be studied in relative isolation (Simon 1962). In saying this, we do not mean to suggest that the language system acts alone. No brain region acts alone. During spoken sentence comprehension, the language system receives input from speech-processing auditory areas (Overath et al. 2015). And during speaking, the language system sends input to speech-articulation areas (Guenther 2016). The language system must also interact with higher-level components of the mind and brain (e.g., to build mental models of the information coming in through language). But the fact that the language system must interact extensively with other brain systems does not undercut the selectivity or distinctness of this system from the rest of the mind and brain (for discussion, see E. Fedorenko, A.A. Ivanova & T.I. Regev, unpublished manuscript). Thus, we turn next to our core question: What representations and computations in the brain's language system allow us to understand the meaning of a sentence or express a thought in language, and can we model these components of human language processing using artificial language models (LMs)?

1.2. What Do We Want from Models of Language Processing?

Before we delve into the utility of artificial LMs to study the language system and its neural basis, let us consider what we want from our scientific models of human language processing. On the one hand, we want the ability to predict behavioral and neural responses to arbitrary stimuli with high accuracy. On the other hand, we want models that are parsimonious and offer some intuitive-level understanding. Most past work in language neuroscience has prioritized parsimony over high predictive accuracy, yielding proposals that coarsely tie cognitive processes (e.g., syntactic processing or lexical access) to particular brain areas (e.g., see Friederici 2002, Hagoort 2005, Hickok & Poeppel 2007, Price 2010, Friederici 2012, Fedorenko et al. 2020, Shain et al. 2023a). Such accounts provide intuitive descriptions of what a given brain region may be doing but leave underspecified much of the detail in its response to language. In particular, what is the nature of the representation of a given sentence? And what algorithms are applied to extract meaning from that sentence?

The field of psycholinguistics has provided traction on these questions, developing sophisticated accounts of linguistic processing based on behavioral experiments and corpus analyses. However, these accounts have typically not attempted to simultaneously address the processing of both linguistic meaning and structure, focusing on context-independent lexical access (e.g., see Dell 1986, Caramazza 1997, Levelt et al. 1999), word-level semantics (e.g., see Landauer et al. 1998, Pennington et al. 2014), or meaning-independent syntactic structure building (e.g., see Clifton & Frazier 1989, Gibson 1998, Lewis & Vasishth 2005). Moreover, these accounts have been difficult to link to neural responses given the lack of appropriate neural data (reliable item-level responses) and the difficulty of deriving quantitative predictions for arbitrary linguistic stimuli at scale.

What language research has long lacked are models whose inner workings can be described mathematically rather than only in words and that (*i*) can build a representation for any arbitrary linguistic stimulus [stimulus computability (Yamins & DiCarlo 2016)]; (*ii*) are data driven, thus avoiding the theoretical precommitments that have so far been needed to operationalize and test hypotheses about human language; and (*iii*) accurately predict behavioral and neural data from humans. Modern LMs have all of these properties (Section 2) and have thus presented language researchers with an exciting opportunity to model human linguistic behavior and neural responses to language with unprecedented quantitative precision (Section 3), albeit at the expense of parsimony (Section 4).

2. LANGUAGE MODELS AS CANDIDATE MODELS OF HUMAN LANGUAGE PROCESSING

2.1. What Are Language Models and What Kinds of Linguistic Knowledge Do They Embody?

LMs are suddenly ubiquitous, rendering many professionals nervous about their future employment, flummoxing professors accustomed to essay-writing assignments, and even convincing some that computer algorithms might be conscious and deserving of moral consideration. What are these things? LMs—sometimes also referred to as artificial neural network language models or large language models—are computer algorithms that are trained to predict the upcoming (or missing) word conditioned on prior (or surrounding) word context (see the **Supplemental Material**, section 1). Early LMs—so-called n-gram models—were based on purely statistical approaches that estimate which word is likely to come next based on how often that word occurs in that context in corpora (Jurafsky & Martin 2008). Then, in the early 2000s, the next-word prediction task was implemented in neural networks, improving performance over earlier approaches (Bengio et al. 2000). The more recent introduction of the transformer architecture (Vaswani et al. 2017) (**Supplemental Figure 1*b***) marked a revolution in next-word prediction. The transformers' training process allows for parallelization on modern computing hardware and thus for efficient use of the abundance of available text data. The key mechanism in transformers, attention, enables the model to focus on diverse aspects of language that matter for which word may come next via multiple attention heads (Bahdanau et al. 2015). In this way, LMs appear to learn about diverse linguistic regularities, ranging from phonological patterns to word forms and meanings and to syntactic structure (for reviews, see Linzen & Baroni 2021, Pavlick 2022, Mahowald et al. 2023).

2.2. A Priori, Why Might We Expect Language Models to Capture Something About Human Language Processing?

We start by stating the obvious: LMs are the first systems apart from the human brain that can generate fluent and coherent text. Indeed, the formal linguistic competence of LMs—knowledge of linguistic rules and regularities (Mahowald et al. 2023)—has been argued to be on par with that of humans (Wang et al. 2020, Brown et al. 2020). Of course, LMs' linguistic prowess in and of itself does not imply that they represent and process language the way humans do (e.g., see Guest & Martin 2023), but similar behavioral outputs are arguably a necessary prerequisite for an artificial model to serve as a candidate model of some biological system.

LMs and humans share several other properties, which makes LMs plausible as candidate models of human language processing (see the sidebar titled Language Learning and Processing in Language Models Versus Humans). First, similar to LMs' core training objective (prediction), abundant evidence indicates that humans predict upcoming linguistic input during comprehension, as measured behaviorally (e.g., see Rayner et al. 2006, Demberg & Keller 2008, Smith & Levy 2013, Brothers & Kuperberg 2021; cf. Huettig & Mani 2016) and neurally (e.g., see Henderson et al. 2016, Willems et al. 2016, Shain et al. 2020, Heilbron et al. 2022; for reviews, see Kuperberg & Jaeger 2016, Ryskin & Nieuwland 2023).

Second, LMs acquire rich and detailed syntactic knowledge—the component of language that gives it its generative power and has been emphasized as a human-unique capacity (e.g., see Berwick & Chomsky 2015). Evidence for the similarity of syntactic knowledge and processing between humans and LMs (cf. van Schijndel & Linzen 2021, Zhang et al. 2023) comes from a traditional linguistic methodology of sentence grammaticality/acceptability judgments (e.g., see Linzen et al. 2016, Marvin & Linzen 2018, Futrell et al. 2019, Gauthier et al. 2020, Hu et al. 2020,

LANGUAGE LEARNING AND PROCESSING IN LANGUAGE MODELS VERSUS HUMANS

Although modern LMs can generate human-like language, they fundamentally differ from the human language system in several respects. First, the amount of training data that LMs are exposed to (billions or trillions of words in text corpora) far exceeds human language exposure (20–70 million words by age 10) (Gilkerson et al. 2017; for discussion, see Warstadt & Bowman 2022). Moreover, the type of training data is vastly different: Children learn language from continuous auditory (speech) or visual (sign) signals in the broader context of physically interacting with the environment and engaging in social interactions (e.g., Hoff 2006, Yu & Smith 2012). Second, transformer LMs have equal access to all previous tokens, whereas humans, limited by memory, instead extract the relevant meaning from linguistic input and quickly discard the exact linguistic sequence (Potter 2012; Christiansen & Chater 2016). Finally, LMs implement language in hardware that differs radically from the biological brain (cf. Kozachkov et al. 2023). For example, the human brain is subject to wiring length costs, whereas LMs are not subject to such spatial pressures, and the biological plausibility of backpropagation is a topic of considerable debate (Lillicrap et al. 2020).

Warstadt et al. 2020), as well as a more psycholinguistics-style approach of measuring incremental, word-by-word processing difficulty as a function of changes in syntactic complexity (Wilcox et al. 2020, 2021).

Third, similar to humans (e.g., Jackendoff 2007), in addition to syntax, LMs acquire sensitivity to multiple levels of linguistic structure, ranging from sublexical (sound-level and morphological) regularities to word forms and meanings and to phrase- and sentence-level structure and meaning (e.g., Tenney et al. 2019, Wang et al. 2019, Wiedemann et al. 2019, Manning et al. 2020, Mikhailov et al. 2021). Moreover, these different kinds of regularities, including linguistic structure and meaning, appear to be intertwined in the LMs' internal representations (Bölücü & Can 2022). As discussed in Section 1.1, in humans, knowledge and processing of different aspects of language, including syntax and semantics, are also not spatially segregated: They all draw on the very same set of brain areas (e.g., see Fedorenko et al. 2010, 2020; Bautista & Wilson 2016; Blank et al. 2016; Shain et al. 2023a), including when measured with high spatial and temporal resolution intracranial recordings (Fedorenko et al. 2016, Nelson et al. 2017). This lack of spatial segregation does not undermine strong sensitivity to syntactic structure; it merely suggests that no neural units (in humans or models) selectively support syntactic structure building, presumably because how words combine in natural language strongly depends on the properties of particular words, as all linguistic frameworks now acknowledge. Importantly, the human language system and LMs–two systems that have emerged independently and under different pressures—seem to have both converged on a solution for efficient language processing without the need to compartmentalize syntax and semantics.

Finally, as discussed in Section 1.1, in humans, language does not share machinery with nonlinguistic tasks, including many aspects of knowledge and reasoning (e.g., see Fedorenko et al. 2011), even when the task is presented linguistically (Monti et al. 2012, Amalric et al. 2018, Shain et al. 2023b). Do linguistic and nonlinguistic abilities also dissociate in LMs? Although some LMs (trained on data other than natural language and on objectives beyond text prediction) are becoming increasingly good at solving reasoning tasks (e.g., see Bhargava & Ng 2022, Imani et al. 2023, Yu et al. 2023), the earlier versions of purely text-trained LMs struggle with such tasks. For instance, small text-trained generative pretrained transformer (GPT) models (such as GPT-2) show strong linguistic competence but fail on even two-digit addition/subtraction (Brown et al. 2020)

WHAT DOES IT MEAN FOR A LANGUAGE MODEL TO BE SIMILAR TO THE HUMAN LANGUAGE NETWORK?

In the studies reviewed here, LMs and brains are compared at the level of internal representations (see the **Supplemental Material**, section 1). Consequently, any claims about model-brain similarity pertain to representations, not the algorithms or implementation underlying language. Although algorithms and representations are tightly linked (Marr 1982), they can still be separated. For instance, the representations from an LM trained to predict the next word using only the left context can produce representations similar to those from an LM that predicts a masked word using both left and right context, despite their algorithms being markedly different. Therefore, based on most current evaluation metrics (cf. Khosla & Williams 2023), the brain-model similarity does not entail the similarity of algorithms or implementations.

Despite significant differences between LMs and humans (e.g., see the sidebar titled Language Learning and Processing in Language Models Versus Humans), LMs are currently the best predictive models of human language representations at the resolution of data we have access to. Distilling the model properties that lead to model-brain representational similarity (Section 4) further sets the stage for investigations of algorithmic- and implementation-level correspondence, including the use of new evaluation metrics and tools from mechanistic interpretability in AI (Section 6).

and on generalizing to examples beyond their training data that are solvable with simple logic rules (H. Zhang et al. 2022). Hence, near-human language ability does not entail near-human reasoning ability (for further discussion, see Fedorenko & Varley 2016, Mahowald et al. 2023, Wong et al. 2023).

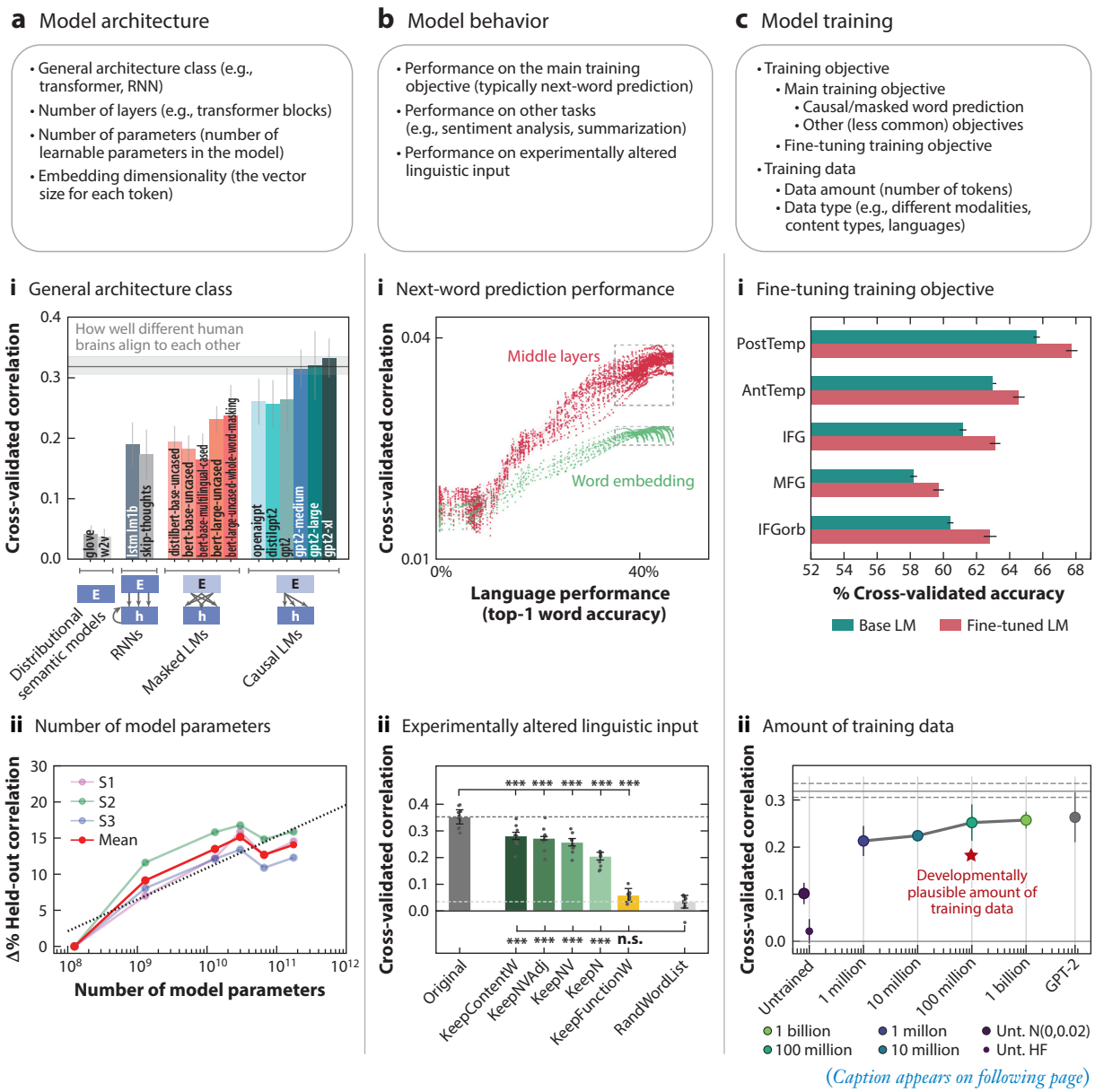
3. LANGUAGE MODELS CAPTURE HUMAN NEURAL RESPONSES TO LANGUAGE

The linguistic success of LMs and their broad similarities to the human language system have made many wonder whether LMs resemble humans at the finer-grained level of the representations they build as they process linguistic input (see the sidebar titled What Does it Mean for a Language Model to Be Similar to the Human Language Network?). Multiple methods have been developed to quantify the representational similarity between LMs and brains, including measuring the ability of LMs to predict the brain activity that will result from a novel linguistic stimulus (encoding methods) or to decode what stimulus elicited a particular brain activity pattern (see the **Supplemental Material**, section 2).

Building on early efforts to relate human brain responses to decontextualized fixed-vector representations of word meanings (Mitchell et al. 2008, Palatucci et al. 2009, Pereira et al. 2011, Fyshe et al. 2014, Huth et al. 2016) (see the **Supplemental Material**, section 3), language researchers can now use modern LMs to study correspondences between the brain and multifaceted contextualized language representations. In particular, LMs provide a way to represent any arbitrary linguistic stimulus, including, critically, compositional stimuli such as phrases and sentences. Some studies using recurrent neural network (RNN) LMs showed that internal representations that contain information about prior context better predict brain responses compared to representations of only the words themselves or the output probabilities of next words (Wehbe et al. 2014, Qian et al. 2016, Jain & Huth 2018). Transformer LMs (Vaswani et al. 2017) provided further support for the importance of representing words in context for capturing brain responses during language processing. Several studies reported greater similarity between transformer LM representations and those extracted from human brains compared to decontextualized word-embedding models

in both fMRI (Toneva & Wehbe 2019, Anderson et al. 2021, Schrimpf et al. 2021, Caucheteux & King 2022, Pasquiou et al. 2022) (**Figure 2*a***) and intracranial recordings (Schrimpf et al. 2021, Goldstein et al. 2022, Goldstein et al. 2023a). Whereas all of the above studies use encoding models (see the **Supplemental Material**, section 2), LMs have also proven useful for decoding stimuli based on brain activity while participants process sentences (Gauthier & Levy 2019, Abdou et al. 2021, Zou et al. 2022) or stories (Abdou et al. 2021, Tang et al. 2023).

Most studies have focused on modeling brain responses during language understanding, but some recent studies have shown that LMs can also predict activity during language generation,



(*Caption appears on following page*)

Figure 2 (*Figure appears on preceding page*)

Distilling the properties of LMs that are important for capturing human neural responses to language. (*a*) Model architecture. (*i*) Schrimpf et al. (2021) investigated several general architecture classes and showed that causal LMs (transformer LMs that have access to only the left context when predicting the next word) best predicted human brain responses, followed by masked LMs (transformer LMs with access to left and right context) and RNNs (recurrence-based LMs), and finally, distributional semantic models (in which each sentence is represented as the average of decontextualized distributional semantic word vectors) (see the **Supplemental Material**, section 3). (*ii*) Antonello et al. (2023) showed that brain encoding performance increases as a logarithmic function of the number of model parameters [for a class of causal LMs (S. Zhang et al. 2022)], with a plateau for models with greater than 30 billion parameters (the plateau is likely to be partly attributable to the limitations of human brain data). Panel *a*, *i* adapted from Schrimpf et al. (2021), and panel *a*, *ii* adapted with permission from Antonello et al. (2023). (*b*) Model behavior. (*i*) Caucheteux & King (2022) demonstrated that models that are better at predicting the next (or missing) word are also better at predicting brain responses (see also Schrimpf et al. 2021). (*ii*) Kauf et al. (2023) altered the input that is passed to the LM (Radford et al. 2018) and evaluated encoding performance on brain responses to the original, unaltered input. They showed that when content words were kept (*green bars*), predictivity remained high, in strong contrast to manipulations where only function words were kept (*yellow bar*). Panel *b*, *i*,*ii* adapted with permission from Caucheteux & King (2022) (CC BY 4.0) and Kauf et al. (2023) (CC BY 4.0), respectively. (*c*) Model training. (*i*) Aw & Toneva (2023) fine-tuned a masked LM on summarization of book chapters and showed improvements in brain encoding performance for several anatomically defined language regions. (*ii*) Hosseini et al. (2024) showed that a causal LM trained on a developmentally plausible amount of data (*red star*) exhibited similar brain encoding performance as a model trained on 1 billion tokens (*light green point*) or a fully trained LM (*gray point*) (Radford et al. 2018). Panel *c*, *i* adapted with permission from Aw & Toneva (2023), and panel *c*, *ii* adapted from Hosseini et al. (2024) (CC BY 4.0). Panels *a*, *i* and *c*, *ii* were adapted to show raw rather than ceiling-normalized correlations. Abbreviations: E, embedding; h, hidden state; LM, language model; n.s., nonsignificant; RNN, recurrent neural network.

before articulation begins, by using intracranially recorded brain activity during spontaneous conversation (Goldstein et al. 2023b, Zada et al. 2023). These results show that the language system relies on abstract representations of linguistic structure and meaning, which support both comprehension and production and resemble the representations learned by transformer LMs (Zada et al. 2023). Importantly, LMs are not trained to predict human data or account for human responses to language; they simply arrive at similar representations of language by virtue of being trained to predict words in text.

4. HOW CAN WE USE LANGUAGE MODELS TO STUDY LANGUAGE PROCESSING IN THE HUMAN BRAIN?

To summarize Sections 2 and 3, for the first time in the history of language research, we have LMs that not only produce and understand language—a feat long argued to only be achievable by humans (Chomsky 1965)—but also represent linguistic information in a sufficiently similar way to humans, allowing relatively accurate brain encoding and decoding during language processing (see the sidebar titled What Does it Mean for a Language Model to Be Similar to the Human Language Network?). These models can now be systematically probed in order to identify properties that are critical for model-to-brain alignment (Section 4.1), and they can be used as in silico model organisms to evaluate hypotheses about language at an unprecedented granularity and scale (Section 4.2).

4.1. Which Properties of Language Models Enable Them to Capture Human Responses to Language?

LMs vary along many dimensions (**Figure 2**), including (*i*) model architecture, which includes intrinsic properties such as the number of layers and parameters; (*ii*) model behavior, which includes a model's performance on natural language processing (NLP) tasks; and (*iii*) model training, which includes the training task (objective) and the training data. The importance of many model properties for capturing human language responses has now been examined, including comparisons of off-the-shelf LMs (studying LMs in the wild) and tighter comparisons of minimally differing LMs (controlled experimental investigations).

4.1.1. Model architecture. All neural networks are defined by their architecture—the arrangement of neuron-like units and the mathematical operations that define the ways in which they are connected. The architecture shapes what kinds of representations can be learned during training. One broad distinction pertains to the general architecture class, with two dominant LM architectures being RNNs and transformers. A few studies have found that transformer LMs predict brain data better than the RNN LMs (Toneva & Wehbe 2019, Schrimpf et al. 2021) (**Figure 2*a***), but other studies have reported comparable performance between a particular class of RNN LMs called long short-term memory models (LSTMs) (Hochreiter & Schmidhuber 1997) and transformer LMs (Hollenstein et al. 2019, Anderson et al. 2021, Oota et al. 2022a, Pasquiou et al. 2022), and Abnar et al. (2019) found that LSTM representations align better with human brain data compared to transformers. However, most past studies have used off-the-shelf models (cf. Pasquiou et al. 2022), which vary not only in their architecture but also in the amount of training they receive, with transformers typically being trained on vastly more data.

More generally, based on the evidence so far, no particular architectural property appears critical for brain alignment: Many instantiations of LM architectures fit brain data well (Schrimpf et al. 2021, Caucheteux & King 2022, Pasquiou et al. 2022, Antonello et al. 2023; for similar findings in vision and audition, see Conwell et al. 2023, Tuckute et al. 2023). That said, at least for transformer models, larger models predict brain data better (Schrimpf et al. 2021, Caucheteux & King 2022, Antonello et al. 2023) (**Figure 2*a***). Some have also begun to develop approaches that aim to better differentiate high-performing models, for example, through the use of so-called controversial stimuli, for which different models make distinct predictions (e.g., see Golan et al. 2023, Hosseini et al. 2023). Such approaches may help uncover architectural motifs that critically modulate model-brain similarity.

4.1.2. Model behavior. One powerful idea is that neural representations (in biological or artificial systems) are shaped by behavioral demands (e.g., Khaligh-Razavi & Kriegeskorte 2014, Yamins et al. 2014, Kell et al. 2018). Indeed, artificial networks that perform better on a target behavior, be it object recognition (for visual neural networks) or next-word prediction (for LMs), appear to develop representations that show greater similarity to brains. In the domain of language, a few studies found that models that are better at next-word prediction also better capture brain responses (Schrimpf et al. 2021, Caucheteux & King 2022, Hosseini et al. 2024) (**Figure 2*a***). In contrast, model performance on other linguistic tasks, including judgments about syntactic or semantic sentence properties, did not significantly explain model-to-brain similarity (Schrimpf et al. 2021). These findings led to the claim that the ability of LMs to predict upcoming linguistic input is a critical factor in explaining human-like representations. However, later work has challenged this claim. Although evidence for the predictive nature of human language processing abounds (e.g., see Smith & Levy 2013, Willems et al. 2016, Shain et al. 2020, Heilbron et al. 2022), a correlation between a model's performance on some task and its similarity to the brain need not imply that the model and the brain are performing the same task. Antonello & Huth (2023) suggested that the critical factor may instead be the generalizability of the representations. They approximated how well representations from a given LM transfer to the representations from a large set of LMs and showed that this metric also positively correlates with model-to-brain similarity scores. This finding suggests that the next-word prediction objective may simply be a powerful way to obtain generalizable representations; it remains unknown (*a*) whether other training objectives can lead to similarly generalizable representations, and if so, (*b*) whether such representations would be able to explain human-like representations, or whether something is special about the next-word prediction objective.

The effect of model behavior on brain predictivity can also be investigated by examining how the model's representations change in response to manipulations of the linguistic input. For

example, how does altering a sentence's structure or meaning affect the ability of the model representations to predict brain responses to the original sentence? This approach allows isolating the aspects of the representation that critically mediate model-to-brain similarity. Kauf et al. (2023) performed a series of experiments to systematically perturb a sentence's structure (e.g., local word swaps, removal of function words) or meaning (e.g., removal of nouns or verbs, paraphrasing to retain a close meaning or only the general topic) (**Figure 2*b***). They found that word meanings were the main contributor to model-to-brain similarity; in contrast, a sentence's syntactic form did not carry much brain-relevant information. The limited importance of syntactic features for model-to-brain similarity aligns with the findings of another input-perturbation investigation: Caucheteux et al. (2021a) obtained a representation of a syntactic frame by averaging the embeddings of ten sentences that share syntactic structure but differ in meaning, and treated the residuals of the syntactic embedding as semantic representations. They showed that the semantic representations have overall higher predictive power than the syntactic ones. Future work should investigate the extent to which the detection of structure effects on model-to-brain similarity is limited by the low temporal resolution of fMRI, the use of materials where word order is not critical for interpretation, and the ways that encoding of syntactic structure differs between LMs and humans.

4.1.3. Model training. One critical aspect of a model's training procedure is the amount and kind of training data. For example, does an LM's ability to predict brain data critically depend on being trained on massive amounts of text, far exceeding the amount of data a human is exposed to (Gilkerson et al. 2017, Warstadt & Bowman 2022) (see the sidebar titled Language Learning and Processing in Language Models Versus Humans)? Hosseini et al. (2024) showed that this is not the case: LMs trained on a developmentally plausible amount of data already capture brain responses well (**Figure 2*c***). However, some training is necessary (Schrimpf et al. 2021, Caucheteux & King 2022, Pasquiou et al. 2022; for a discussion, see Hosseini et al. 2024). The kinds of data an LM is trained on can also be manipulated to ask which properties are critical for capturing brain responses. Pasquiou et al. (2023) trained LMs on two variants of a text corpus: a semantic version, which retained only the content words, and a syntactic version, which retained only the morphosyntactic features (e.g., part of speech tags and agreement information). They found that models trained on these data sets could each predict brain data to some extent, although semantic features predicted responses across a larger portion of the brain (see also the sidebar titled What Are We Modeling?).

What about the training task? Most models that have been analyzed for their similarity to the brain have been trained on predicting the next word (causal LMs) or a missing word (masked

WHAT ARE WE MODELING?

Some researchers target the language system precisely with functional localizers (e.g., Saxe et al. 2006, Fedorenko et al. 2010), others use predefined anatomical areas, and yet others examine responses across the whole brain. By specifically targeting language cortex via functional localization, we can test the degree to which LMs capture language processing in the brain, unconfounded from the processing of perceptual inputs to the language system (e.g., Overath et al. 2015), or the downstream cognitive processes that operate on linguistic meaning. Of course, the representation of meaning in the brain is not restricted to language-specific cortex (e.g., Huth et al. 2016), so researchers interested in meaning representations more broadly may choose to examine neural activity across the brain or in brain areas known to represent specific aspects of meaning (e.g., Jain & Huth 2018, Anderson et al. 2021, Toneva et al. 2022). Whether different components of LM representations capture variance in the core language areas versus other brain areas remains an important open question.

LMs), and causal LMs outperform masked LMs at predicting brain data (Schrimpf et al. 2021, Caucheteux & King 2022; cf. Pasquiou et al. 2022). Some studies have investigated how fine-tuning affects the LMs' ability to predict brain responses. In particular, after being trained on the next-word prediction objective, an LM can be fine-tuned on a specific data set or task, which forces the model to pay attention to particular task-relevant information. The results have not been consistent. Gauthier & Levy (2019) fine-tuned a transformer LM on four standard NLP tasks (e.g., question answering) and found that all four tasks impaired brain decoding relative to the non-fine-tuned model. In contrast, Oota et al. (2022b) fine-tuned an LM on 10 NLP tasks and found improvements in brain encoding performance for many of these tasks, which suggests that the representational features that these tasks emphasize may also be important dimensions in how humans represent linguistic information. Gauthier & Levy (2019) also fine-tuned an LM on two custom tasks, which consisted of predicting missing words in a sentence or predicting which sentence is likely to come next in a corpus where words were shuffled within a sentence or paragraph, thus selecting against word-order information. Fine-tuning these tasks actually improved decoding performance, which again suggests a limited importance of syntactic information for capturing language responses (at least in fMRI sentence-level data), similar to Kauf et al. (2023) and Caucheteux et al. (2021a). Finally, some studies have found that fine-tuning LMs for deeper understanding—via training on narrative summarization (Aw & Toneva 2023) or via next-word prediction on texts similar to the ones that the brain recordings correspond to (Merlin & Toneva 2022)—improves brain encoding performance (**Figure 2*c***).

4.2. Using Encoding Models as In Silico Language Networks

LMs can also be used to simulate and design neuroscience experiments. These applications rely on accurate encoding models, which are mappings from LM representations to brain responses (see the **Supplemental Material**, section 2). Modern LMs' ability to represent any linguistic input enables predictions about brain responses to arbitrary new stimuli. Encoding models can therefore be used as a virtual language network to simulate the language areas' responses. Such in silico experiments can be used to (*i*) validate prior empirical findings and (*ii*) test new manipulations, for which no human brain data exist (Wehbe et al. 2018; Jain et al. 2020, 2023; Ratan Murty et al. 2021). For example, using an LSTM-based encoding model, Jain et al. (2020) (see also Caucheteux et al. 2021b) were able to recapitulate prior empirical findings of shorter temporal integration windows in auditory areas and longer ones in the language areas (Lerner et al. 2011, Blank & Fedorenko 2020).

Encoding models can also be used to identify supernormal stimuli for a particular system (Barrett 2010), that is, stimuli that elicit the strongest possible response. These predictions can then be evaluated empirically in a closed-loop design, and the critical stimuli can be analyzed to better understand the computations that the relevant brain region supports. This approach has proven successful in systems neuroscience (e.g., Bashivan et al. 2019, Ponce et al. 2019), and Tuckute et al. (2024) applied a similar strategy to the language network. They first built an encoding model based on brain responses to 1,000 diverse sentences, then derived predictions for millions of new sentences, and finally, collected brain responses in new participants to the sentences that were predicted to elicit the strongest response (drive sentences). The drive sentences indeed elicited a very strong response in the language areas, which suggests that an LM-based encoding model was sufficiently accurate for model-guided experiments in brain areas implicated in higher-level cognition. The analysis of the drive sentences further revealed that the language areas respond most strongly to surprising sentences with unusual grammar and/or meaning. Critically, using model predictions to obtain experimental stimuli effectively expands the hypothesis space beyond the experimenters' preconceived notions.

5. THE CHALLENGES OF USING LANGUAGE MODELS TO UNDERSTAND LANGUAGE IN THE BRAIN

5.1. General Methodological Challenges

The use of LMs to understand the human language system faces numerous challenges. One challenge is that transformer LMs—the most widely used architecture—are incredibly expressive (Yun et al. 2020), allowing them to find patterns in any sequential input data (text, audio, amino acids, etc.). Some have described transformers as universal computational engines (Lu et al. 2021). This power makes it critical to include rigorous controls to ensure that the obtained results are not due to trivial reasons (e.g., encoding of some low-level sentence features) and reflect what the experimenter thinks they reflect (for a discussion, see Kauf et al. 2023).

Other challenges are inherent to how LM representations are compared to neural data (see the **Supplemental Material**, section 2). For example, LMs can provide representations of linguistic input with long windows of preceding context, meaning that in a narrative, the model considers the entire preceding story when generating a representation for each word. Depending on how the data are divided into train and test splits, contextual representations may inflate model-to-brain similarity if the representations were derived by taking the full sequence into account (Antonello et al. 2023, Kauf et al. 2023). Furthermore, most studies that compare LM representations to brains do not test generalization to held-out participants (Jain & Huth 2018, Toneva & Wehbe 2019, Jain et al. 2020, Schrimpf et al. 2021, Merlin & Toneva 2022, Oota et al. 2022b, Aw & Toneva 2023, Hosseini et al. 2024; cf. Toneva et al. 2022, Tang et al. 2023, Tuckute et al. 2024), which can lead to reliance on participant-specific idiosyncrasies. Such overfitting may not matter for some medical applications, such as individualized brain computer interfaces (geared toward the language/semantic system of a particular patient) (e.g., Tang et al. 2023), but it discourages the discovery of general models of language. Lastly, consensus is currently lacking on how to define the theoretical maximum similarity between an LM representation and neural recordings, that is, the noise ceiling. In perceptual domains, noise ceilings are typically estimated using stimulus repetitions under the assumption that repeated presentations of the same stimulus elicit the same neural response (e.g., see Allen et al. 2022). However, this assumption may not hold for language (or other cognitive domains). And from a practical standpoint, it is challenging to even collect brain responses for multiple repetitions of language stimuli because language processing requires attentional engagement (e.g., see Cohen et al. 2021), which is harder to sustain over stimulus repetitions. Consequently, studies vary widely in how they collect language neuroscience data and how they assess reliability, posing challenges for cross-study comparison.

5.2. Challenges Related to the Increasing Divergence Between Neuroscience and Engineering Goals

AI and neuroscience share a deeply intertwined history (e.g., see Zador et al. 2023). In recent years, neuroscientists have benefited from engineering advances in AI, repurposing models developed by engineers as hypotheses for neural processes in biological brains. For language, the goals of engineers/computer scientists and neuroscientists were aligned for a long time: The former were working to create models that understand and produce language, and the latter were seeking such models to understand human language processing. After formal linguistic competence rose to a human-like level with the advent of transformer LMs (Brown et al. 2020, Wang et al. 2020), the goals of the two communities started to diverge.

The primary goal of AI has now shifted toward the development of artificial general intelligence (AGI) models: next-word-prediction-trained models that are subsequently adapted to

perform diverse downstream tasks, including those outside of the language domain (e.g., solving math proofs) (Bommasani et al. 2022, Imani et al. 2023). The ability to support a wide range of tasks leads to models that increasingly differ from the human language system, which is highly selective for linguistic tasks (Section 1.1). The goal of developing general-purpose models has led to models that continue to increase in size, guided by scaling laws that predict improvements in model performance with increased scale (model size and amount of training data and compute) (Kaplan et al. 2020, cf. McKenzie et al. 2023). Additionally, to support a wider range of tasks, models are expanding their training data to include nonlinguistic input (e.g., images, computer code bases) (Achiam et al. 2023) and incorporating additional objectives such as reinforcement-based human feedback (Ziegler et al. 2019). In contrast to the larger and more diverse AGI models, neuroscientists seek models that offer both high predictive power and parsimony (Section 1.1). Larger models are inherently less parsimonious and do not always provide gains in the predictive accuracy of behavioral and brain responses. For example, larger models trained on text prediction are actually worse at predicting human behavioral data, such as reading times (Shain et al. 2022b, Oh & Schuler 2023, Steuer et al. 2023). They also appear to get worse on some language tasks, struggling with negation (Jang et al. 2023, McKenzie et al. 2023) and quantifiers (Gupta 2023, Michaelov & Bergen 2023) and tending to memorize more (Carlini et al. 2023, McKenzie et al. 2023). The ability to predict brain responses does increase with model size, but this relationship appears to level off at approximately 30 billion parameters (Antonello et al. 2023). Overall, larger LMs provide diminishing returns for predictive accuracy, at least for the kinds of data that the field is currently trying to model (this may change for higher-dimensional data, such as single-unit recordings). They also pose greater interpretive challenges because they have more parameters, are trained on larger and more diverse data sets, and are often additionally fine-tuned on other training objectives.

Larger models developed in engineering contexts present two additional challenges for neuroscientists. First, they are often proprietary, providing no access to model internals or even knowledge about architecture or training (Achiam et al. 2023). And second, because of this lack of information and/or compute resources, it is not possible for scientists to perform controlled experimental model comparisons, which often require, for example, retraining a model from scratch.

In summary, the latest models from AI appear to be worse accounts of human language processing than some earlier models, and practical limitations prevent neuroscientists from even using these models in a rigorous (transparent and replicable) way in their investigations.

6. WHAT'S NEXT?

The field of computational cognitive neuroscience is still in its infancy (Naselaris et al. 2018), especially for language. Nevertheless, the initial successes summarized here lay a promising foundation for future efforts. A core goal of the field is to build increasingly accurate models of the human language system. One step toward this goal will be to build developmentally plausible models that learn linguistic computations directly from speech signals and from realistic amounts and kinds of data (e.g., Beguš 2021, Warstadt et al. 2023). Another step toward this goal will be the development of LMs that interact with both lower-level mechanisms (e.g., speech perception mechanisms) and higher-level systems of knowledge and reasoning. Although language is distinct from both, the human language system must interact with perception, motor control, and cognition to enable the full array of human abilities. Building such multicomponent models may also provide critical clues as to how representations get transformed from lower-level perception to language to downstream reasoning. Finally, another important goal is to develop models that can

explain implementation/algorithmic-level processes, not only representational similarity (Blank 2023). The growing field of mechanistic interpretability in AI provides an increasing number of tools to dissect the inner workings of models (e.g., see Wang et al. 2022, Hosseini & Fedorenko 2023, Meng et al. 2023). Neuroscientists can use these tools to formulate hypotheses about implementation and algorithms that underlie language behavior. These hypothesized mechanisms can then be manipulated or ablated, and the effects on both (*i*) downstream linguistic task performance and (*ii*) model-to-brain similarity can be evaluated, allowing for a virtuous cycle of hypothesis generation and testing.

In conclusion, artificial LMs have provided language researchers with a powerful new tool for understanding human language processing by providing computationally explicit hypotheses of how language might work in the brain. Like all methodological approaches, LMs have limitations and pose numerous challenges. Nonetheless, these models have lifted critical barriers on our path to understanding the neural, cognitive, and computational architecture of language processing, providing exciting opportunities to understand the human language system with unprecedented computational precision.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Colton Casto for help with **Figure 1**. We are grateful to Cory Shain for feedback on the manuscript, to Klemen Kotar and Kyle Mahowald for input on particular sections, and to Roger Levy for helpful discussions of predictive processing. G.T. was supported by the Amazon Fellowship from the Science Hub (administered by the MIT Schwarzman College of Computing), the International Doctoral Fellowship from the American Association of University Women, and the K. Lisa Yang Integrative Computational Neuroscience Center Graduate Fellowship. E.F. was supported by National Institutes of Health award U01-NS121471 and by research funds from the McGovern Institute for Brain Research, the Department of Brain and Cognitive Sciences, the Simons Center for the Social Brain, and the MIT Quest for Intelligence.

LITERATURE CITED

Abdou M, Gonzalez AV, Toneva M, Hershcovich D, Søgaard A. 2021. Does injecting linguistic structure into language models lead to better alignment with brain recordings? arXiv:2101.12608 [cs.CL]

Abnar S, Beinborn L, Choenni R, Zuidema W. 2019. Blackbox meets blackbox: representational similarity and stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, ed. T Linzen, G Chrupala, Y Belinkov, D Hupkes, pp. 191–203. Florence, Italy: Assoc. Comput. Linguist.

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. 2023. GPT-4 technical report. arXiv:2303.08774 [cs.CL]

Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* 25(1):116–26

Amalric M, Denghien I, Dehaene S. 2018. On the role of visual experience in mathematical development: evidence from blind mathematicians. *Dev. Cogn. Neurosci.* 30:314–23

Anderson AJ, Kiela D, Binder JR, Fernandino L, Humphries CJ, et al. 2021. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *J. Neurosci.* 41(18):4100–19

Antonello R, Huth A. 2023. Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiol. Lang.* **https://doi.org/10.1162/nol_a_00087**

Antonello R, Vaidya A, Huth AG. 2023. Scaling laws for language encoding models in fMRI. arXiv:2305.11863 [cs.CL]

Apperly IA, Samson D, Humphreys GW. 2009. Studies of adults can inform accounts of theory of mind development. *Dev. Psychol.* 45(1):190–201

Aw KL, Toneva M. 2023. *Training language models to summarize narratives improves brain alignment*. Paper presented at the Eleventh International Conference on Learning Representations, Kigali, Rwanda, May 1–5. **https://openreview.net/forum?id=KzkLAE49H9b**

Badecker W, Caramazza A. 1985. On considerations of method and theory governing the use of clinical categories in neurolinguistics and cognitive neuropsychology: the case against agrammatism. *Cognition* 20(2):97–125

Bahdanau D, Cho K, Bengio Y. 2015. *Neural machine translation by jointly learning to align and translate*. Paper presented at the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, May 7–9

Barrett D. 2010. *Supernormal Stimuli: How Primal Urges Overran Their Evolutionary Purpose*. New York: W. W. Norton & Company

Bashivan P, Kar K, DiCarlo JJ. 2019. Neural population control via deep image synthesis. *Science* 364(6439):eaav9436

Bautista A, Wilson SM. 2016. Neural responses to grammatically and lexically degraded speech. *Lang. Cogn. Neurosci.* 31(4):567–74

Beguš G. 2021. CiwGAN and fiwGAN: encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. *Neural Netw.* 139:305–25

Bengio Y, Ducharme R, Vincent P. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, ed. T Leen, T Dietterich, V Tresp. San Diego, CA: NeurIPS. **https://papers.nips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html**

Berndt RS. 1991. Sentence processing in aphasia. In *Acquired Aphasias*, ed. M Sarno, pp. 223–70. Orlando, FL: Academic Press

Berwick RC, Chomsky N. 2015. *Why Only Us: Language and Evolution*. Cambridge, MA: MIT Press

Bhargava P, Ng V. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: a survey. arXiv:2201.12438 [cs.CL]

Blank IA. 2023. What are large language models supposed to model? *Trends Cogn. Sci.* 27(11):987–89

Blank IA, Balewski Z, Mahowald K, Fedorenko E. 2016. Syntactic processing is distributed across the language system. *NeuroImage* 127:307–23

Blank IA, Fedorenko E. 2020. No evidence for differences among language regions in their temporal receptive windows. *NeuroImage* 219:116925

Blank IA, Kanwisher N, Fedorenko E. 2014. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J. Neurophysiol.* 112(5):1105–18

Bölücü N, Can B. 2022. Analysing syntactic and semantic features in pre-trained language models in a fully unsupervised setting. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, ed. Md. S Akhtar, T Chakraborty, pp. 19–31. New Delhi, India: Assoc. Comput. Linguist.

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, et al. 2022. On the opportunities and risks of foundation models. arXiv:2108.07258 [cs.LG]

Braga RM, DiNicola LM, Becker HC, Buckner RL. 2020. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *J. Neurophysiol.* 124(5):1415–48

Brothers T, Kuperberg GR. 2021. Word predictability effects are linear, not logarithmic: implications for probabilistic models of sentence comprehension. *J. Mem. Lang.* 116:104174

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin. San Diego, CA: NeurIPS. **https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html**

Caramazza A. 1997. How many levels of processing are there in lexical access? *Cogn. Neuropsychol.* 14(1):177–208

Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F, Zhang C. 2023. Quantifying memorization across neural language models. arXiv:2202.07646 [cs.LG]

Caucheteux C, Gramfort A, King J-R. 2021a. Disentangling syntax and semantics in the brain with deep networks. arXiv:2103.01620 [cs.CL]

Caucheteux C, Gramfort A, King J-R. 2021b. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. arXiv:2110.06078 [q-bio.NC]

Caucheteux C, King J-R. 2022. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* 5(1):134

Chen X, Affourtit J, Ryskin R, Regev TI, Norman-Haignere S, et al. 2023. The human language system, including its inferior frontal component in "Broca's area," does not support music perception. *Cereb. Cortex* 33(12):7904–29

Chomsky N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press

Christiansen MH, Chater N. 2016. The now-or-never bottleneck: a fundamental constraint on language. *Behav. Brain Sci.* 39:e62

Clifton C, Frazier L. 1989. Comprehending sentences with long-distance dependencies. In *Linguistic Structure in Language Processing*, ed. GN Carlson, MK Tanenhaus, pp. 273–317. Dordrecht, Neth.: Springer

Cohen L, Salondy P, Pallier C, Dehaene S. 2021. How does inattention affect written and spoken language processing? *Cortex* 138:212–27

Conwell C, Prince JS, Kay KN, Alvarez GA, Konkle T. 2023. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? bioRxiv 2022.03.28.485868. **https://doi.org/10.1101/2022.03.28.485868**

Deen B, Koldewyn K, Kanwisher N, Saxe R. 2015. Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb. Cortex* 25(11):4596–609

Dehaene S, Spelke E, Pinel P, Stanescu R, Tsivkin S. 1999. Sources of mathematical thinking: behavioral and brain-imaging evidence. *Science* 284(5416):970–74

Dell GS. 1986. A spreading-activation theory of retrieval in sentence production. *Psychol. Rev.* 93(3):283–321

Demberg V, Keller F. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2):193–210

Fedorenko E, Behr MK, Kanwisher N. 2011. Functional specificity for high-level linguistic processing in the human brain. *PNAS* 108(39):16428–33

Fedorenko E, Blank IA. 2020. Broca's area is not a natural kind. *Trends Cogn. Sci.* 24(4):270–84

Fedorenko E, Blank IA, Siegelman M, Mineroff Z. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition* 203:104348

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. *J. Neurophysiol.* 104(2):1177–94

Fedorenko E, Ryskin R, Gibson E. 2022. Agrammatic output in non-fluent, including Broca's, aphasia as a rational behavior. *Aphasiology* 37(12):1981–2000

Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, et al. 2016. Neural correlate of the construction of sentence meaning. *PNAS* 113(41):E6256–62

Fedorenko E, Varley R. 2016. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Ann. N. Y. Acad. Sci.* 1369(1):132–53

Friederici AD. 2002. Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* 6(2):78–84

Friederici AD. 2012. The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* 16(5):262–68

Futrell R, Wilcox E, Morita T, Qian P, Ballesteros M, Levy R. 2019. Neural language models as psycholinguistic subjects: representations of syntactic state. arXiv:1903.03260 [cs.CL]

Fyshe A, Talukdar PP, Murphy B, Mitchell TM. 2014. Interpretable semantic vectors from a joint model of brain- and text-based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ed. K Toutanova, H Wu, pp. 489–99. Baltimore, MD: Assoc. Comput. Linguist.

Gauthier J, Hu J, Wilcox E, Qian P, Levy R. 2020. SyntaxGym: an online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ed. A Celikyilmaz, T-H Wen, pp. 70–76. Assoc. Comput. Linguist. **https://doi.org/10.18653/v1/2020.acl-demos.10**

Gauthier J, Levy R. 2019. Linking artificial and human neural representations of language. 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)*, ed. K Inui, J Jiang, V Ng, X Wan, pp. 529–539. Hong Kong: Assoc. Comput. Linguist.

Geschwind N. 1970. The organization of language and the brain. *Science* 170(3961):940–44

Gibson E. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68(1):1–76

Gilkerson J, Richards JA, Warren SF, Montgomery JK, Greenwood CR, et al. 2017. Mapping the early language environment using all-day recordings and automated analysis. *Am. J. Speech Lang. Pathol.* 26(2):248–65

Golan T, Siegelman M, Kriegeskorte N, Baldassano C. 2023. Testing the limits of natural language models for predicting human language judgements. *Nat. Mach. Intel.* 5(9):952–64

Goldstein A, Ham E, Nastase SA, Zada Z, Grinstein-Dabus A, et al. 2023a. Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. bioRxiv 2022.07.11.499562. **https://doi.org/10.1101/2022.07.11.499562**

Goldstein A, Wang H, Niekerken L, Zada Z, Aubrey B, et al. 2023b. Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations. bioRxiv 2023.06.26.546557. **https://doi.org/10.1101/2023.06.26.546557**

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* 25(3):369–80

Goodglass H. 1993. *Understanding Aphasia*. San Diego, CA: Academic Press

Grodzinsky Y, Santi A. 2008. The battle for Broca's region. *Trends Cogn. Sci.* 12(12):474–80

Guenther FH. 2016. *Neural Control of Speech*. Cambridge, MA: MIT Press

Guest O, Martin AE. 2023. On logical inference over brains, behaviour, and artificial neural networks. *Comput. Behav.* 6(2):213–27

Gupta A. 2023. Probing quantifier comprehension in large language models: another example of inverse scaling. arXiv:2306.07384 [cs.CL]

Hagoort P. 2005. On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9(9):416–23

Heilbron M, Armeni K, Schoffelen J-M, Hagoort P, de Lange FP. 2022. A hierarchy of linguistic predictions during natural language comprehension. *PNAS* 119(32):e2201968119

Henderson JM, Choi W, Lowder MW, Ferreira F. 2016. Language structure in the brain: a fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage* 132:293–300

Hickok G, Poeppel D. 2007. The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8(5):393–402

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–80

Hoff E. 2006. How social contexts support and shape language development. *Dev. Rev.* 26(1):55–88

Hollenstein N, de la Torre A, Langer N, Zhang C. 2019. CogniVal: a framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, ed. M Bansal, A Villavicencio, pp. 538–49. Hong Kong: Assoc. Comput. Linguist.

Hosseini EA, Fedorenko E. 2023. *Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language*. Paper presented at the 37th Annual Conference on Neural Information Processing Systems, New Orleans, LA, Dec. 10–16

Hosseini EA, Schrimpf M, Zhang Y, Bowman S, Zaslavsky N, Fedorenko E. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiol. Lang.* In press. **https://doi.org/10.1162/nol_a_00137**

Hosseini EA, Zaslavsky N, Casto C, Fedorenko E. 2023. *Teasing apart the representational spaces of ANN language models to discover key axes of model-to-brain alignment*. Paper presented at the 2023 Conference on Cognitive Computational Neuroscience (CCN 2023), Oxford, UK, Aug. 24–27

Hu J, Gauthier J, Qian P, Wilcox E, Levy RP. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, ed. D Jurafsky, J Chai, N Schluter, J Tetreault, pp. 1725–44. Assoc. Comput. Linguist. **https://aclanthology.org/2020.acl-main.158/**

Hu J, Small H, Kean H, Takahashi A, Zekelman L, et al. 2023. Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cereb. Cortex* 33:4384–404

Huettig F, Mani N. 2016. Is prediction necessary to understand language? Probably not. *Lang. Cogn. Neurosci.* 31(1):19–31

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532:453–58

Imani S, Du L, Shrivastava H. 2023. MathPrompter: mathematical reasoning using large language models. arXiv:2303.05398 [cs.CL]

Ivanova AA, Mineroff Z, Zimmerer V, Kanwisher N, Varley R, Fedorenko E. 2021. The language network is recruited but not required for nonverbal event semantics. *Neurobiol. Lang.* 2(2):176–201

Ivanova AA, Srikant S, Sueoka Y, Kean HH, Dhamala R, et al. 2020. Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife* 9:e58906

Jackendoff R. 2007. A parallel architecture perspective on language processing. *Brain Res.* 1146:2–22

Jain S, Huth AG. 2018. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett. San Diego, CA: NeurIPS. **https://proceedings.neurips.cc/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html**

Jain S, Vo VA, Mahto S, LeBel A, Turek JS, Huth A. 2020. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. In *In Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin. San Diego, CA: NeurIPS. **https://proceedings.neurips.cc/paper/2020/hash/9e9a30b74c49d07d8150c8c83b1ccf07-Abstract.html**

Jang J, Ye S, Seo M. 2023. Can large language models truly understand prompts? A case study with negated prompts. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, ed. A Albalak, C Zhou, C Raffel, D Ramachandran, S Ruder, X Ma, pp. 52–62. PMLR. **https://proceedings.mlr.press/v203/jang23a.html**

Jain S, Vo VA, Wehbe L, Huth AG. 2023. Computational language modeling and the promise of in silico experimentation. *Neurobiol. Lang*. In press. **https://doi.org/10.1162/nol_a_00101**

Jouravlev O, Schwartz R, Ayyash D, Mineroff Z, Gibson E, Fedorenko E. 2019. Tracking colisteners' knowledge states during language comprehension. *Psychol. Sci.* 30(1):3–19

Jurafsky D, Martin J. 2008. *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall. 2nd ed.

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, et al. 2020. Scaling laws for neural language models. arXiv:2001.08361 [cs.LG]

Kauf C, Tuckute G, Levy R, Andreas J, Fedorenko E. 2023. Lexical-semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network. *Neurobiol. Lang.* In press. **https://doi.org/10.1162/nol_a_00116**

Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98(3):630–44.e16

Khaligh-Razavi S-M, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput. Biol.* 10(11):e1003915

Khosla M, Williams AH. 2023. Soft matching distance: a metric on neural representations that captures single-neuron tuning. arXiv:2311.09466 [cs.LG]

Kozachkov L, Kastanenka KV, Krotov D. 2023. Building transformers from neurons and astrocytes. *PNAS* 120(34):e2219150120

Kuperberg GR, Jaeger TF. 2016. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31(1):32–59

Landauer TK, Foltz PW, Laham D. 1998. An introduction to latent semantic analysis. *Discourse Process* 25(2–3):259–84

Lerner Y, Honey CJ, Silbert LJ, Hasson U. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31(8):2906–15

Levelt WJM, Roelofs A, Meyer AS. 1999. A theory of lexical access in speech production. *Behav. Brain Sci.* 22(1):1–38

Levitin DJ, Menon V. 2003. Musical structure is processed in "language" areas of the brain: a possible role for Brodmann area 47 in temporal coherence. *NeuroImage* 20(4):2142–52

Lewis RL, Vasishth S. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29(3):375–419

Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. 2020. Backpropagation and the brain. *Nat. Rev. Neurosci.* 21(6):335–46

Linzen T, Baroni M. 2021. Syntactic structure from deep learning. *Annu. Rev. Linguist.* 7:195–212

Linzen T, Dupoux E, Goldberg Y. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguist.* 4:521–35

Lipkin B, Tuckute G, Affourtit J, Small H, Mineroff Z, et al. 2022. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Sci. Data* 9(1):529

Liu Y-F, Kim J, Wilson C, Bedny M. 2020. Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *eLife* 9:e59340

Lu K, Grover A, Abbeel P, Mordatch I. 2021. Pretrained transformers as universal computation engines. arXiv:2103.05247 [cs.LG]

Luria AR. 1970. The functional organization of the brain. *Sci. Am.* 222(3):66–72

Mahowald K, Fedorenko E. 2016. Reliable individual-level neural markers of high-level language processing: a necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage* 139:74–93

Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, Fedorenko E. 2023. Dissociating language and thought in large language models. arXiv:2301.06627 [cs.CL]

Malik-Moraleda S, Ayyash D, Gallée J, Affourtit J, Hoffmann M, et al. 2022. An investigation across 45 languages and 12 language families reveals a universal language network. *Nat. Neurosci.* 25(8):1014–19

Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *PNAS* 117(48):30046–54

Marr D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman

Marvin R, Linzen T. 2018. Targeted syntactic evaluation of language models. arXiv:1808.09031 [cs.CL]

McKenzie IR, Lyzhov A, Pieler M, Parrish A, Mueller A, et al. 2023. Inverse scaling: when bigger isn't better. arXiv:2306.09479 [cs.CL]

Menenti L, Gierhan SME, Segaert K, Hagoort P. 2011. Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychol. Sci.* 22(9):1173–82

Meng K, Bau D, Andonian A, Belinkov Y. 2023. Locating and editing factual associations in GPT. arXiv:2202.05262 [cs.CL]

Merlin G, Toneva M. 2022. Language models and brain alignment: beyond word-level semantics and prediction. arXiv:2212.00596 [cs.CL]

Michaelov JA, Bergen BK. 2023. Rarely a problem? Language models exhibit inverse scaling in their predictions following few-type quantifiers. arXiv:2212.08700 [cs.CL]

Mikhailov V, Serikov O, Artemova E. 2021. Morph call: probing morphosyntactic content of multilingual transformers. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, ed. E Vylomova, E Salesky, S Mielke, G Lapesa, R Kumar, et al., pp. 97–121. Assoc. Comput. Linguist. **https://aclanthology.org/2021.sigtyp-1.10/**

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191–95

Monti MM, Parsons LM, Osherson DN. 2009. The boundaries of language and thought in deductive inference. *PNAS* 106(30):12554–59

Monti MM, Parsons LM, Osherson DN. 2012. Thought beyond language: neural dissociation of algebra and natural language. *Psychol. Sci.* 23(8):914–22

Naselaris T, Bassett DS, Fletcher AK, Kording K, Kriegeskorte N, et al. 2018. Cognitive computational neuroscience: a new conference for an emerging discipline. *Trends Cogn. Sci.* 22(5):365–67

Nelson MJ, El Karoui I, Giber K, Yang X, Cohen L, et al. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *PNAS* 114(18):E3669–78

Novick JM, Trueswell JC, Thompson-Schill SL. 2005. Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. *Cogn. Affect. Behav. Neurosci.* 5(3):263–81

Oh B-D, Schuler W. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Trans. Assoc. Comput. Linguist.* 11:336–50

Oota SR, Alexandre F, Hinaut X. 2022a. Long-term plausibility of language models and neural dynamics during narrative listening. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, ed. J Culbertson, A Perfors, H Rabagliati, V Ramenzoni, pp. 2462–69. Univ. Calif. **https://escholarship.org/uc/item/7r95j62c**

Oota SR, Arora J, Agarwal V, Marreddy M, Gupta M, Surampudi BR. 2022b. Neural language taskonomy: which NLP tasks are the most predictive of fMRI brain activity? arXiv:2205.01404 [cs.CL]

Overath T, McDermott JH, Zarate JM, Poeppel D. 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* 18(6):903–11

Palatucci M, Pomerleau D, Hinton GE, Mitchell TM. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, ed. Y Bengio, D Schuurmans, J Lafferty, C Williams, A Culotta. San Diego, CA: NeurIPS. **https://papers.nips.cc/paper_files/paper/2009/hash/1543843a4723ed2ab08e18053ae6dc5b-Abstract.html**

Pallier C, Devauchelle A-D, Dehaene S. 2011. Cortical representation of the constituent structure of sentences. *PNAS* 108(6):2522–27

Pasquiou A, Lakretz Y, Hale J, Thirion B, Pallier C. 2022. Neural language models are not born equal to fit brain data, but training helps. arXiv:2207.03380 [cs.AI]

Pasquiou A, Lakretz Y, Thirion B, Pallier C. 2023. Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax and context. arXiv:2302.14389 [cs.CL]

Pavlick E. 2022. Semantic structure in deep learning. *Annu. Rev. Linguist.* 8:447–71

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*, ed. A Moschitti, B Pang, W Daelemans, pp. 1532–43. Doha, Qatar: Assoc. Comput. Linguist.

Pereira F, Detre G, Botvinick M. 2011. Generating text from functional brain images. *Front. Hum. Neurosci.* 5:72

Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* 177(4):999–1009.e10

Potter MC. 2012. Recognition and memory for briefly presented scenes. *Front. Psychol.* 3:32

Price CJ. 2010. The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* 1191(1):62–88

Qian P, Qiu X, Huang X. 2016. Bridging LSTM architecture and the neural dynamics during reading. arXiv:1604.06635 [cs.CL]

Radford A, Narasimhan K, Salimans T, Sutskever I. 2018. *Improving language understanding by generative pre-training*. Work. Pap., OpenAI, San Francisco, CA. **https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf**

Ratan Murty NA, Bashivan P, Abate A, DiCarlo JJ, Kanwisher N. 2021. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* 12(1):5540

Rayner K, Reichle ED, Stroud MJ, Williams CC, Pollatsek A. 2006. The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychol. Aging* 21(3):448–65

Regev TI, Casto C, Hosseini EA, Adamek M, Ritaccio AL, et al. 2023. Neural populations in the language network differ in the size of their temporal receptive windows. bioRxiv 2022.12.30.522216. **https://doi.org/10.1101/2022.12.30.522216**

Ryskin R, Nieuwland MS. 2023. Prediction during language comprehension: What is next? *Trends Cogn. Sci.* 27(11):1032–52

Saussure F. 1959. *Course in General Linguistics*. New York: Columbia Univ. Press

Saxe R, Brett M, Kanwisher N. 2006. Divide and conquer: a defense of functional localizers. *NeuroImage* 30(4):1088–96

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, et al. 2021. The neural architecture of language: integrative modeling converges on predictive processing. *PNAS* 118(45):e2105646118

Shain C, Blank IA, Fedorenko E, Gibson E, Schuler W. 2022a. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *J. Neurosci.* 42(39):7412–30

Shain C, Blank IA, van Schijndel M, Schuler W, Fedorenko E. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* 138:107307

Shain C, Kean H, Lipkin B, Affourtit J, Siegelman M, et al. 2023a. 'Constituent length' effects in fMRI do not provide evidence for abstract syntactic processing. bioRxiv 2021.11.12.467812. **https://doi.org/10.1101/2021.11.12.467812**

Shain C, Meister C, Pimentel T, Cotterell R, Levy RP. 2022b. Large-scale evidence for logarithmic effects of word predictability on reading time. PsyArXiv. **https://doi.org/10.31234/osf.io/4hyna**

Shain C, Paunov A, Chen X, Lipkin B, Fedorenko E. 2023b. No evidence of theory of mind reasoning in the human language network. *Cereb. Cortex* 33(10):6299–319

Simon HA. 1962. The architecture of complexity. *Proc. Am. Philos. Soc.* 106(6):467–82

Smith NJ, Levy R. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3):302–19

Steuer J, Mosbach M, Klakow D. 2023. Large GPT-like models are bad babies: a closer look at the relationship between linguistic competence and psycholinguistic measures. arXiv:2311.04547 [cs.CL]

Tang J, LeBel A, Jain S, Huth AG. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.* 26:858–66

Tenney I, Das D, Pavlick E. 2019. BERT rediscovers the classical NLP pipeline. arXiv:1905.05950 [cs.CL]

Toneva M, Mitchell TM, Wehbe L. 2022. Combining computational controls with natural text reveals aspects of meaning composition. *Nat. Comput. Sci.* 2(11):745–57

Toneva M, Wehbe L. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 14954–64. San Diego, CA: NeurIPS

Tuckute G, Feather J, Boebinger D, McDermott JH. 2023. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biol.* 21(12):e3002366

Tuckute G, Sathe A, Srikant S, Taliaferro M, Wang M, et al. 2024. Driving and suppressing the human language network using large language models. *Nat. Hum. Behav.* In press. **https://doi.org/10.1038/s41562-023-01783-7**

van Schijndel M, Linzen T. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cogn. Sci.* 45(6):e12988

Varley RA, Klessinger NJC, Romanowski CAJ, Siegal M. 2005. Agrammatic but numerate. *PNAS* 102(9):3519–24

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, ed. I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, et al. San Diego, CA: NeurIPS. **https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html**

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, et al. 2020. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. arXiv:1905.00537 [cs.CL]

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. 2019. GLUE: a multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461 [cs.CL]

Wang K, Variengien A, Conmy A, Shlegeris B, Steinhardt J. 2022. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. arXiv:2211.00593 [cs.LG]

Warstadt A, Bowman SR. 2022. What artificial neural networks can tell us about human language acquisition. arXiv:2208.07998 [cs.CL]

Warstadt A, Choshen L, Mueller A, Williams A, Wilcox E, Zhuang C. 2023. Call for papers—the BabyLM Challenge: sample-efficient pretraining on a developmentally plausible corpus. arXiv:2301.11796 [cs.CL]

Warstadt A, Parrish A, Liu H, Mohananey A, Peng W, et al. 2020. BLiMP: the benchmark of linguistic minimal pairs for English. *Trans. Assoc. Comput. Linguist.* 8:377–92

Wehbe L, Huth AG, Deniz F, Gao J, Kieseler M-L, Gallant JL. 2018. *BOLD predictions: automated simulation of fMRI experiments*. Poster presented at the 2018 Conference on Cognitive Computational Neuroscience, Philadelphia, PA, Sept. 6

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):e112575

Wiedemann G, Remus S, Chawla A, Biemann C. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. arXiv:1909.10430 [cs.CL]

Wilcox EG, Gauthier J, Hu J, Qian P, Levy R. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, ed. S Denison, M Mack, Y Xu, BC Armstrong, pp. 1707–13. Seattle, WA: Cogn. Sci. Soc.

Wilcox EG, 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*, ed. C Zong, F Xia, W Li, R Navigli, pp. 939–52. Assoc. Comput. Linguist. **https://doi.org/10.18653/v1/2021.acl-long.76**

Willems RM, Frank SL, Nijhof AD, Hagoort P, van den Bosch A. 2016. Prediction during natural language comprehension. *Cereb. Cortex* 26(6):2506–16

Wilson SM, Entrup JL, Schneck SM, Onuscheck CF, Levy DF, et al. 2023. Recovery from aphasia in the first year after stroke. *Brain* 146(3):1021–39

Wong L, Grand G, Lew AK, Goodman ND, Mansinghka VK, Andreas J, Tenenbaum JB. 2023. From word models to world models: translating from natural language to the probabilistic language of thought. arXiv:2306.12672 [cs.CL]

Yamins DLK, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19(3):356–65

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111(23):8619–24

Yu C, Smith LB. 2012. Embodied attention and word learning by toddlers. *Cognition* 125(2):244–62

Yu J, Wang X, Tu S, Cao S, Zhang-Li D, et al. 2023. KoLA: carefully benchmarking world knowledge of large language models. arXiv:2306.09296 [cs.CL]

Yun C, Bhojanapalli S, Rawat A, Reddi SJ, Kumar S. 2020. *Are transformers universal approximators of sequence-to-sequence functions?* Paper presented at the International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, Apr. 30

Zada Z, Goldstein A, Michelmann S, Simony E, Price A, et al. 2023. A shared linguistic space for transmitting our thoughts from brain to brain in natural conversations. bioRxiv 2023.06.27.546708. **https://doi.org/10.1101/2023.06.27.546708**

Zador A, Escola S, Richards B, Ölveczky B, Bengio Y, et al. 2023. Catalyzing next-generation Artificial Intelligence through NeuroAI. *Nat. Commun.* 14(1):1597

Zhang H, Li LH, Meng T, Chang K-W, den Broeck GV. 2022. On the paradox of learning to reason from data arXiv:2205.11502 [cs.CL]

Zhang S, Roller S, Goyal N, Artetxe M, Chen M, et al. 2022. OPT: open pre-trained transformer language models. arXiv:2205.01068 [cs.CL]

Zhang Y, Gibson E, Davis F. 2023. Can language models be tricked by language illusions? Easier with syntax, harder with semantics. arXiv:2311.01386 [cs.CL]

Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, et al. 2019. Fine-tuning language models from human preferences. arXiv:1909.08593 [cs.CL]

Zou S, Wang S, Zhang J, Zong C. 2022. Cross-modal cloze task: a new task to brain-to-word decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, ed. S Muresan, P Nakov, A Villavicencio, pp. 648–57. Dublin, Irel.: Assoc. Comput. Linguist. **https://doi.org/10.18653/v1/2022.findings-acl.54**