Neuron

CellPress
OPEN ACCESS

Perspective

Unifying the structures of language in a neural population code

Samuel A. Nastase,[1,*] Zaid Zada,[2,3] Adele Goldberg,[3] and Uri Hasson[2,3]
[1]Department of Psychology and Center for Computational Language Sciences, University of Southern California, Los Angeles, CA, USA
[2]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[3]Department of Psychology, Princeton University, Princeton, NJ, USA
*Correspondence: snastase@usc.edu
https://doi.org/10.1016/j.neuron.2026.07.024

SUMMARY

Large language models (LLMs) have mastered human language in ways that no previous computational system has. While rule-based, symbolic systems sufficed for constrained, well-defined problems, they were not able to accommodate the context-sensitive expressivity of natural language. LLMs instead use statistical learning to encode the diversity of linguistic structures into a unified high-dimensional embedding space. Strikingly, this context-driven, distributed representation closely parallels neural population codes, suggesting that the human language system may have converged on a similar computational strategy. Drawing on a growing body of work at the intersection of artificial intelligence and cognitive neuroscience, we show that LLMs can serve as cognitively plausible models of the neural computations supporting language in the human brain. We conclude that explaining how language can emerge from neural population codes, in both biological and artificial systems, will not be achieved through the incremental refinement of algebraic-symbolic theories but will demand new theoretical paradigms.

INTRODUCTION

Language is a marvelous achievement of the human brain and culture. The remarkable complexity and expressivity of human language have traditionally set it apart as an object of study distinct from other domains of cognition and social behavior. Within linguistics and neurolinguistics, this complexity is typically confronted using a divide-and-conquer approach: researchers have developed rich—yet largely separate—theories of phonology, syntax, semantics, and pragmatics (Figure 1A). Students typically learn about these topics in separate courses, and researchers tend to specialize in only one or two areas. A dominant paradigm since Chomsky[1–3] treats language as a system of categorical, symbolic rules that combine lexical items largely independently of communicative function or context of use (see Box 1). Central to this view is the claim that certain core properties of language—its generative, combinatorial, and recursive productivity, the capacity to "make infinite use of finite means"[2,4]—must arise from a cognitive system of algebraic, rule-based computations operating over discrete symbolic representations.[5–7] The parallel to algebra is illustrated by Fodor and Pylyshyn's[5] example that anyone who understands *John loves Mary* can also understand *Mary loves John* on the assumption that symbols like *Mary* and *John* enter into a constituency relationship homologous to *X loves Y*, where *X* and *Y* are variables that can be filled by any atomic symbol and the *loves* relation applies in a rule-based fashion across any such fillers. (However, we note that the type of filler does matter: *Tyler loves pancakes* versus *Pancakes love Tyler*; *Tyler loves Abraham Lincoln* versus *Lincoln loves Tyler*).

While the algebraic-symbolic tradition has produced a wealth of formal constructs for describing linguistic structure, it has not ultimately produced holistic models capable of processing or producing rich, communicative language in everyday contexts. Early rule-based systems for natural language processing achieved some success on structured problems in highly constrained domains; for example, Winograd's[65] SHRDLU procedural block-world manipulator.[66–69] However, these systems were brittle and scaled poorly: they demanded extensive manual rule construction and failed to handle the ambiguity, variability, and open-endedness of real-world language.[70] These systems were eventually supplanted by statistically oriented models of linguistic structure that were more broadly applicable and easier to scale.[71–74]

Large language models (LLMs) are the first generation of computational models capable of processing and producing fluent, meaningful, and context-sensitive human language.[75–77] What makes these models so effective? In this perspective, we argue that the success of LLMs stems from a foundational computational commitment to statistical learning over a unified, high-dimensional embedding space (Figure 1B). This foundation allows LLMs to absorb and express essentially all of the structures of natural language,[32,78–82] while subsequent training innovations, including instruction-tuning and reinforcement learning from human feedback (RLHF), further refine and align the models to human communicative goals. LLMs scale up the theoretical foundations of connectionism[83] from experimentally motivated, handcrafted models to the full scope of human language (Box 2).[33]

A system of this kind calls for theories and explanations different from those familiar to many researchers in the field.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

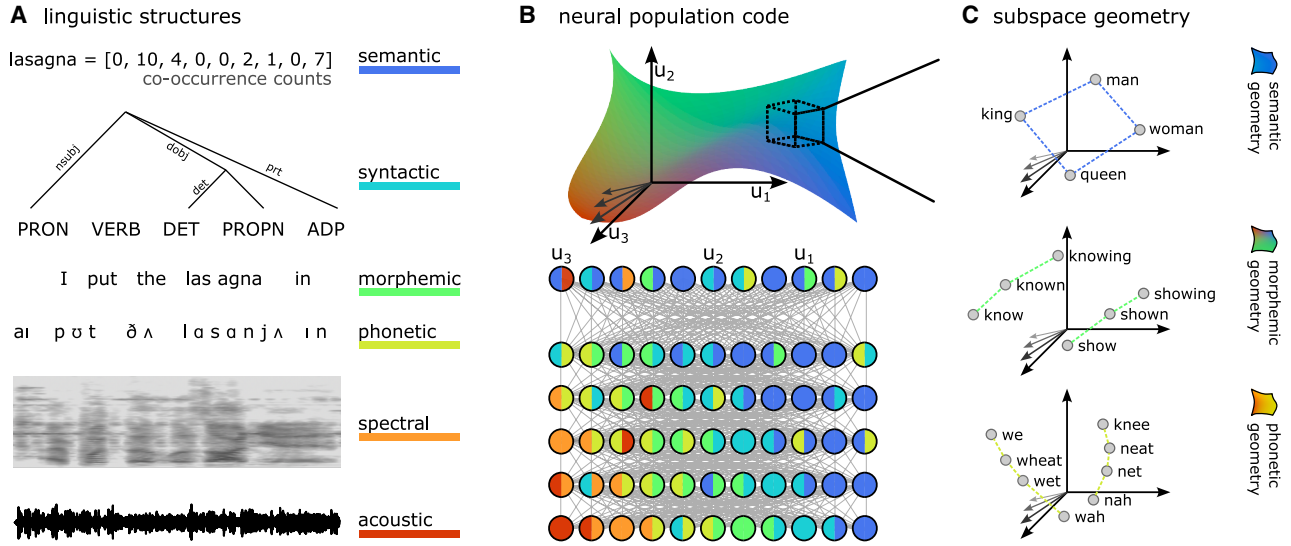


Figure 1. Encoding different kinds of linguistic structure in a unified embedding space

(A) Linguistic formalisms for describing language appeal to different sets of features or rules for different levels of analysis. But the brain must integrate these structures in a lingua franca of neural computation.

(B) Neural population codes, as implemented in connectionist networks such as large language models (LLMs) and in the human brain, provide a unified computational "language" for representing and processing the various structures of language. Linguistic inputs are encoded as distributed patterns of activity—or embeddings—across a population of many simple, neuron-like computing elements. Any given unit may exhibit graded, context-sensitive, heterogeneous tuning over a variety of seemingly distinct linguistic levels. Connections between populations of units transform these representations, and transformations may occur between different populations of units (i.e., various layers of a processing hierarchy) or dynamically within a population from one moment to the next (i.e., recurrent connections).

(C) All structures of language (e.g., phonetic features, morphemic inflections, grammatical constructions, including words, idioms, and more abstract patterns) are encoded in the geometric relationships between embeddings. For example, words with similar meanings are encoded nearer to each other in representational space, and phonetic, morphemic, syntactic, and semantic relations correspond to different directions in representational space. Under this framework, a linguistic utterance unfolds over time as concurrent trajectories of neural activity among interconnected neural populations.

The expressivity of a high-dimensional embedding space allows the model to capture the myriad interrelated patterns of natural language without explicit rules. The geometric structure of the embedding space allows the model to interpolate the meaning and usage of words in novel contexts (Figure 1C). Crucially, from the perspective of a statistical learning model, there is no qualitative distinction between linguistic patterns that behave in a rule-like manner and those that do not. Take verb inflection, for example: the model learns that certain verbs take a regular past tense while others have irregular forms, yet it does not process these two classes in a fundamentally different way. Both are encountered as constituents of surface-level contextual patterns, both are learned through co-occurrence statistics, and both are produced as statistically likely continuations of a preceding token sequence. A subtler example involves argument structure. The verb *make* accepts an open-ended range of adjectives (*He made her happy/sad/worried*), whereas the verb *drive* is conventionally restricted to adjectives conveying a loss of composure (*He drove her crazy/insane/bananas* but not *He drove her sad*). Where a theorist might seek a concise, interpretable rule to capture this contrast, the LLM simply encodes the full distribution of attested patterns—insofar as doing so enables it to produce fluent, contextually appropriate language.

Our focus on LLMs in this review stems from their parallels with another type of neural network that masters language: the human brain. We suspect it is no coincidence that both LLMs and the human brain appear to have found converging solutions to language in neural population codes.[103–108] We hope to express our excitement about what we perceive as a crucial step toward building a unified, ecological theory of language processing—one grounded in real-world communication with mechanistic explanations based on neural computations. Throughout the manuscript, we review evidence for shared computational principles in how LLMs and the human brain process natural language. We focus in particular on the power of statistical learning and the capacity of a high-dimensional embedding space to accommodate the diversity of linguistic structures. Despite the apparent complexity of LLMs, we argue that they provide a more parsimonious—and more neuroscientifically relevant—model of human language processing than alternative models.

HOW LLMs WORK: A NEURAL NETWORK SOLUTION TO NATURAL LANGUAGE

Transformer-based language models[109] now process and produce human language with a fluency and versatility unlike any previous system, matching or exceeding human performance across tasks spanning comprehension, translation, reasoning, and open-ended generation.[75–77,110–115] Autoregressive models, such as the GPT family,[75] are trained on a deceptively simple self-supervised objective: predicting the next token given the context of preceding tokens. In text-based LLMs, tokens

Box 1. Context and diversity in the patterns of everyday language use

Language is fundamentally a communicative act—coordinated between individuals, shaped by context, and oriented toward social goals.[8–11] A long tradition of functional linguistics has emphasized the role that everyday contexts of use play in shaping linguistic structure,[12–20] and a large body of work suggests that the demands of communication exert considerable pressure on the structure of languages themselves.[21–31] We use language in a wide range of contexts for a wide range of purposes: to share information, to organize our thoughts, to question, command, threaten, and signal social affiliation.[18,32–38] More traditional linguistic theories make two commitments that sit uneasily within this ecological framework.

The commitment to linguistic modularity. The divide-and-conquer approach has led to distinct formal systems for distinct levels of linguistic structure (Figure 1A). Phonological units are decomposed into bundles of binary features,[39,40] words into morphological roots and affixes,[41–43] syntactic structures into binary-branching trees annotated with parts of speech and grammatical roles,[2,44] and lexical meanings into hierarchical graphs or semantic decompositions.[45–48] Underlying this approach is a commitment to linguistic modularity, both compartmentalizing language from other domains of cognition and social behavior and subdividing language into separate domains of inquiry. The risk of this approach is that it obscures the fact that many linguistic patterns are inherently multidimensional, simultaneously spanning multiple levels of structure.

Consider the choice between *gave a book to her* versus *gave her a book*. Both patterns are available for common verbs like *give*, but longer, more formal verbs (often borrowed from Latin), like *transfer* or *return*, resist the second pattern: *returned the book to her* sounds more natural than *returned her the book*. Processing factors like constituent length affect the choice: *she gave the book that she found at a second-hand shop to Mary* is more burdensome than *she gave Mary the book she found at a second-hand shop*. Semantic properties, like the animacy of the participants, affect the choice: *gave the book to the library* sounds more natural than *gave the library the book*. Discourse factors affect the choice as well: if *Mary* has already been under discussion, *gave her the book* may be favored. If *the book* is the topic, *gave the book to her* may be more natural. No single level—phonology, syntax, semantics, information structure, or pragmatics alone—predicts the observed behavior, and constraints at multiple levels operate simultaneously.[49] How (and whether) a given linguistic distinction is grammatically marked also varies dramatically across languages. Consider what counts as a subject. In English, an experiencer is encoded as the grammatical subject when it is topical, as in *I like it*. Spanish marks the same relationship differently: *me encanta* ("it delights me") treats the experiencer as an indirect object, paralleling the rarer English construction *it is pleasing to me*. The variation can run deeper still. In Ewe, a Niger-Congo language spoken in Ghana, intentional actions that are semantically intransitive are nonetheless expressed using transitive syntax: "to run" is literally "move.limbs course."[50]

The commitment to universal, unlearned features. The second commitment is to a fixed inventory of symbolic, categorical, and universal features at each level of structure. The Minimalist program, for example, aspires to reduce syntax to a single recursive operation—*Merge*—that combines two formal units into a binary-branching structure of arbitrary depth.[51,52] The primitive units of this system, described as "word-like atomic elements,"[53,54] are presumed to host a range of formal features that determine how atomic concepts are combined in an unlearned Universal Grammar[55,56]: grammatical category features ([N] and [V]), φ-features ([person], [number], and [gender]), case features ([Nom] and [acc]), selectional features ([for types of complements]), positional features, and discourse-related features such as [Top] and [Foc] that may trigger "movement."[48,57,58] The proliferation of such features raises a question: in what sense does the Minimalist program actually minimize the content of Universal Grammar? More fundamentally, treating linguistic categories as fixed, categorical, and universal is at odds with the diversity of the world's roughly 7,000 languages[37,59–62] and the considerable variation within each.[63,64] There are no formal criteria that consistently identify nouns, verbs, or subjects across all languages, except by appeal to semantically prototypical properties, suggesting that these categories may be better understood as emergent from use rather than as primitive or universal.[37]

correspond to words or parts of words (e.g., *running* may be tokenized as *run* and *ning*) and serve as input and output units of the model. Note that different models use different token vocabularies. Speech-based models typically operate on continuous audio or discrete acoustic units rather than text tokens. Despite its simplicity, this objective—applied at scale over vast quantities of text—proves remarkably powerful,[116] effectively pushing the model to learn not just surface statistics but the syntactic, semantic, and pragmatic structure of language.

When we refer to the "statistical structure" of language, we do not simply mean word frequencies or local associations (the fact that *happy* often precedes *birthday*, for instance). We use the term inclusively to encompass the full array of statistical dependencies in natural language: across time (from syllables to words to sentences to discourse) and across levels of description (acoustic, syntactic, semantic, and pragmatic). This includes not only local co-occurrence patterns but also long-distance dependencies: for example, a plural verb (e.g., *were*) must agree with a possibly distant subject in a complex sentence. It also includes context-sensitive meaning: *cold* means something different depending on whether the surrounding discourse concerns the outside temperature, someone's personality, or their flu symptoms. Neural networks, as highly expressive statistical learners, are particularly well-suited to capturing precisely this kind of complex, multiscale statistical structure that resists concise definitions.[117,118] Crucially, a neural network does not need to represent these

Box 2. Scaling up and the specter of connectionism

LLMs are direct descendants of the connectionist models that have played a significant role in neuroscience, linguistics, and cognitive science for over half a century. The principles outlined in work on parallel distributed processing (PDP)[83,84] form the theoretical basis for many recent developments in deep learning. Surprisingly sophisticated behaviors can emerge from a network of simple, interacting units that dynamically settle into a state best satisfying a combination of constraints across units.[85,86] This body of work has made a compelling case that seemingly rule-based linguistic computations can be grounded in more general processes of statistical learning in neural networks, despite the system never explicitly learning the rules.[87–91]

While there have been numerous technical innovations over the past two decades that have led to the development of LLMs, there are two key advances over the previous generation of connectionism. The first is obvious: scale. Massive increases in raw computing capacity have enabled engineers to run larger and deeper networks against massive training corpora. This astronomical scaling up has led to emergent, qualitatively different performance.[92–96] Notably, larger-scale LLMs also appear to better align with human brain activity during natural language comprehension.[97,98] The second advance is, to our reading, underappreciated. The majority of classical connectionist models developed in psychology are "toy" models by modern standards: they were specifically designed to express targeted patterns of behavior, often tied to particular experimental paradigms, and typically trained on handcrafted inputs. The availability of massive volumes of uncurated images and text on the Internet has changed things. The deep learning revolution began in earnest when researchers relinquished handcrafted, easily interpretable features and started developing models that learn whatever features are necessary[99] to perform human-like tasks on largely unconstrained, naturalistic inputs.[100,101] This shift from experimentally constrained boutique models to full-fledged models trained to perform real-world tasks on real-world data has led to qualitatively new capabilities.[33,102]

dependencies in an easily interpretable form to exploit them effectively in processing and producing language.

Each input token is immediately projected into a high-dimensional embedding space—a vector corresponding to a pattern of activity distributed across hundreds or thousands of neuron-like units—and passed through multiple layers of processing. At the final layer, this embedding is used to generate a probability distribution over the entire vocabulary of tokens, indicating which tokens are most likely to come next given all words in the preceding context. In early autoregressive models, this context window included thousands of tokens (e.g., the original GPT-2 had a context window of 1,024 tokens), while in modern models the context window has grown to hundreds of thousands, if not millions of tokens. The output token is sampled from this probability distribution, while the learning signal is derived from the continuous probability distribution (e.g., via cross-entropy loss) rather than the output token itself. Although the input and output tokens are discrete symbols, all of the internal computations are continuous and high-dimensional. As the activity pattern for a given token proceeds from one layer to the next, it is progressively updated by circuit elements called attention heads. These attention heads are the core mechanism for context sensitivity: each attention head can look back over the entire context window and use that context to refine the meaning of the current token. The meaning of *game*, for instance, will correspond to different embeddings depending on whether the preceding text is related to sports or to poaching.

Modern models are further refined and socialized through a number of different methods: for example, RLHF optimizes models toward human-preferred outputs.[119,120] Instruction tuning (a form of supervised fine-tuning; SFT) optimizes models on desirable input-output pairs to bias the model toward helpful responses.[121] Modern models also deploy chain-of-thought reasoning[122] and are trained on tasks with verifiable rewards (e.g., mathematics and programming) to improve their performance on complex, multi-step tasks. These fine-tuning methods build on the foundation of next-word prediction and serve to make models more helpful, reliable, and conversationally appropriate.[123–126] Critically, all of this complexity boils down to relatively simple matrix arithmetic: embedding vectors are transformed via weight matrices from one layer to the next through simple weighted sums and nonlinearities. The same applies to attention heads. All of the weights are *learned*: any meaningful structure in the embeddings (and behavior) is absorbed statistically from linguistic examples and ultimately stored in the learned weights.

Two recurring ideas run through this computing architecture, and we believe they hold the key to why LLMs succeed where rule-based systems have failed. We unpack each in turn below and revisit both later in the paper as candidate principles for understanding language in the human brain.

WHY LLMs WORK: TWO PRINCIPLES OF BIOLOGICAL LEARNING

We contend that the success of LLMs is rooted in two computational principles also observed in biological learning systems: (1) encoding the structures of natural language as distributed representations in a high-dimensional embedding space and (2) replacing rule-based learning with context-driven computations for reproducing the statistical structure of real-world language. In the next two subsections, we unpack how each principle manifests in LLMs and then discuss plausible points of contention.

Principle 1: Encoding the structures of language in a high-dimensional embedding space

How can a neuroscientific theory of language accommodate all of the structures of natural language? Modern artificial neural networks provide one possible solution: LLMs learn to encode the diverse structures of language in a unified, high-dimensional embedding space—a neural population code. The embedding space provides a unified computational *lingua franca*[127] for encoding and integrating all the structures of language. In a neural

CellPress
OPEN ACCESS

population code, the fundamental unit of representation is not a single neuron but a distributed pattern of activity across an ensemble of neuron-like units. These patterns of activity can be thought of as vectors corresponding to specific locations in a high-dimensional, continuous space of neural activity (or "representational space"). No single unit hosts a cleanly interpretable "concept" or "rule" the way one might design a classical computer or cognitive architecture.[128,129] Instead, individual units exhibit "polysemanticity"[130,131]: context-sensitive responses to heterogeneous combinations of linguistic features (related to mixed selectivity observed in the brain).[132,133] Meaning and structure, in the most general sense, are not stored in individual units but encoded in the geometric relationships between patterns of activity—in the distances, angles, and neighborhoods of the representational space (Figure 1C).[134–140]

Surprisingly complex structures can be encoded in the geometry of a vector space of this kind. To invoke a simplistic example, an activity vector for the word *queen* can be approximated via simple vector arithmetic: $v_{king} - v_{male} + v_{female} \approx v_{queen}$.[73,141] In LLMs, all manner of syntactic, morphological, semantic, and contextual structures are encoded along specific directions in this high-dimensional vector space (Figure 1C).[78–81,142–145] For example, Hewitt and Manning[142] demonstrated that hierarchical syntactic structures can be approximately recovered from the representational geometry of an LLM. That is, non-adjacent grammatical relationships between words are encoded in particular subspaces of the embedding space. Diego-Simón and colleagues[146] further demonstrated that the type and direction of syntactic dependencies are oriented along particular directions in the embedding space. While specific internal circuits may approximate familiar syntactic operations,[147–149] the inputs and outputs of these operations are "read out from" and "written into" the embedding space.[150]

This continuous, geometric representational format yields a surprisingly powerful kind of generalization.[151] For example, LLMs can interpolate novel relationships from a handful of previously seen examples without any additional training (i.e., in-context learning).[77,152–154] Even points of the representational space that the system has *never* precisely visited before can be interpreted in a meaningful way, given the overall geometry of the vector space. Consider the combinatorial nature of language. The words you are reading right now may occur in a particular context you have never experienced before. They may express an idea slightly different from any you have previously entertained. How is the LLM able to accommodate such novelty? Not by retrieving a stored match but by triangulating and geometrically interpreting a new location in the embedding space.[155] By training on vast quantities of real-world language, the model constructs a dense, richly structured internal landscape; a geometry of meaning in which related concepts cluster, analogies form parallel lines, and context shifts a word's position in principled ways. A word appearing in a never-before-seen context lands somewhere in this space, and its meaning can be interpolated from its geometric neighborhood. The latent relational structures encoded in this way can be surprisingly abstract, supporting genuine analogical reasoning across domains.[156–158] Novelty, in this framework, is not a problem to be solved by rules—rather, it is a matter of geometric navigation.

While efforts to interpret the circuits and representations of LLMs are valuable for understanding how neural population codes can implement languages, we should resist the temptation to read these models through a symbolic, rule-based lens.[32,33] The fact that we can decode certain linguistic constructs from a model's internal representations does not mean those constructs are driving the model's computations. This is a subtle but important distinction. Just because we can recover something resembling dependency grammar from the attention heads[147,159] does not mean the attention heads are implementing graph-based computations of the kind specified by dependency grammar theory. Certain attention heads may appear to be involved in certain syntactic operations (e.g., identifying direct objects), but these same attention heads are influenced by the semantic plausibility of the relationship they identify.[160] That is, LLMs do not appear to perform purely algebraic, rule-governed operations on symbolic structures like syntax trees or semantic graphs.[140] Instead, they learn to populate a high-dimensional embedding space with representations whose geometric relationships are useful for predicting real-world language. What appear to be qualitatively distinct levels of structure (morphology, syntax, and semantics) are all, from the model's perspective, geometric relationships between points in the same continuous representational space. The boundaries we draw between these levels are our own.

Principle 2: Learning the statistical patterns of language in context

The success of LLMs marks a significant moment in the long-running debate between empiricist and rationalist accounts of language acquisition.[161–163] Where some rule-based theories have argued that the complexity of language requires innate, domain-specific knowledge, LLMs learn to reproduce linguistic behavior directly from exposure to real-world language through a "mindless," algorithmic fitting process.[32] Central to this success is the self-supervised objective of next-word prediction. This deceptively simple task has driven neural network models of language for decades,[84,116,164–167] but its power only became fully apparent at scale: trained on vast quantities of text, next-token prediction proves sufficient to induce rich representations of phonology, morphology, syntax, semantics, and discourse, without specifying any of these as a learning target.[75,77,110]

In LLMs, the combination of the transformer architecture with a large context window of preceding text allows the model to leverage statistical patterns to better capture meaning in context. What *is* "context" in language? For LLMs, context is everything—or, more precisely, *everything is context*. Following the simple, self-supervised objective of next-token prediction, the LLM will seek to leverage *any* patterns in its context window insofar as they are useful for producing fluent, appropriate human language. Some of these patterns are highly regular. These patterns are more easily detected in language behavior, and many correspond to patterns that linguists have richly described in terms of morphological, syntactic, and lexical-semantic structures. Syntactic or structural patterns, for example, appear to crystallize fairly early in a model's learning timeline.[168] Many other patterns are more abstract, graded, multidimensional, or otherwise less regular. These patterns are less obvious and

harder to describe.[169–171] However, for the LLM, all of these structures are simply patterns in context. Syntactic dependencies, for example, are a subset of contextual relations useful to the model but are only theoretically distinct from other contextual relations from the perspective of the linguist or experimenter. Patterns at a particular "level" (Figure 1A) are not privileged or processed in a qualitatively different way than any other pattern.

The primacy of context has profound implications: it renders all types of structures *learnable* by a simple, general-purpose, self-supervised learning algorithm (with some evidence that patterns that are unlearnable by people are also more challenging for LLMs).[172] In pursuing this simple objective, the LLM seeks to compress patterns across prior tokens (i.e., across time) into a single "right now" pattern of activity that captures the current context and is useful for producing upcoming tokens. The power and simplicity of statistical learning challenge us to rethink certain intuitions about language. None of the structures of language are built into or given *a priori* to the model—they are only learned to the extent that they help the model better pursue its objective of reproducing natural language.[162,173] An LLM does not learn a "rule" for the English plural while memorizing certain "exceptions,"[174,175] and an LLM does not seek to learn concise or interpretable rules for language at all—it learns to approximate whatever patterns of language it can absorb from its training data.[33] In the same way, the model does not seek to compartmentalize different levels of linguistic structure: patterns of phonology, morphology, syntax, and semantics are all integrated and accommodated by the same statistical learning mechanism.

LLMs and the mapping problem

The computational architecture of LLMs is fundamentally different from those inherited from formal linguistics and symbolic computing. Consider Poeppel's[176] "mapping problem": can we link the formal structures of linguistics to the neural computations performed by an LLM? In an LLM, any given input or output token corresponds to a collection of activity patterns expressed across the layers of the model. Operations that the linguist may consider as distinct—phase-structure generation, semantic composition, linearization—are all learned through the same general-purpose algorithm, all implemented by self-attention circuits, and all supported by the geometry of the embedding space. Linguistic categories like phonemes and parts of speech are not represented as primitive building blocks but emerge as approximate, continuous structures in the embedding space insofar as they are useful for predicting upcoming inputs.[177–179] The model encodes predictive structure in a graded, continuous, multidimensional fashion and only commits to a discrete output token at the final step, when a single token is sampled from a probability distribution over possible continuations. Producing language that is well-formed, semantically coherent, and contextually appropriate corresponds to probabilistically pursuing a meaningful trajectory through a high-dimensional embedding space. The learned weights of the model propel the internal activity toward appropriate sequences of tokens. If the human brain likewise relies on context-driven computations over a high-dimensional population code—rather than rule-like operations over discrete symbols—this should fundamentally reshape how we study the neural basis of language. Instead of attempting to map the constructs of formal linguistics directly onto the brain, we would be better served by trying to understand how the patterns of language emerge from a neural system that seeks to transform linguistic inputs into fluid, meaningful, context-sensitive outputs.[155]

LLMs do not implement linguistic rules

One plausible objection to our argument so far is that we are conflating learning mechanisms with learning outcomes: while LLMs learn through a "soft" statistical process operating over continuous vector representations, perhaps the outcome of learning is a set of rules for language. After all, a native English speaker, without explicit instruction, eventually commands the complex regularities of the language: they say *the keys to the cabinet were missing*, not *was missing*, correctly matching the verb to the distant plural noun *keys* rather than the closer singular noun *cabinet*. We argue that this framing mischaracterizes what has been learned. The native speaker does not consult a mental rulebook—rather, they have internalized a high-dimensional system of conventions in which the verb agrees in number with the matching subject. Here we distinguish a *rule*, a compact, symbolic abstraction designed to extrapolate beyond experience, from a *convention*, a complex, experience-driven statistical regularity in behavior. LLMs are not rule learners but convention learners: "direct-fit" function approximators that utilize their massive parameter space to approximate the complex manifold of language within the boundaries of their experience.[33] The result is a geometric space of linguistic conventions so dense and richly structured that the model can generalize to an unbounded variety of novel sentences, not by condensing knowledge into context-free symbolic rules but by interpolating within the learned manifold. When we say LLMs do not learn rules, we are not merely commenting on their complexity. We are pointing to a fundamental property of their computing architecture: they solve linguistic problems through dense geometric mapping rather than through the discovery of portable symbolic axioms.

LLMs and Marr's levels of analysis

A related objection is that we are failing to distinguish Marr's[180] computational, algorithmic, and implementational levels of analysis in our treatment of LLMs. That is, one might argue that LLMs are best understood as an implementation of a symbolic computational system.[5,181] However, reducing LLMs to a subordinate level of abstraction within Marr's framework obscures the fundamentally different computational and algorithmic principles separating LLMs from algebraic-symbolic systems. At the computational level, symbolic theories hold that the goal of the language system is to distill a set of grammatical rules and representations to support the productivity of language. LLMs, on the other hand, pursue a statistical objective for producing fluent, context-sensitive probability distributions over token sequences (i.e., to internalize and reproduce a set of conventions). These are not two instantiations of the same computation, but rather fundamentally different computational goals. At an algorithmic level, symbolic systems manipulate discrete representations via rule-governed operations. LLMs, on the other hand, transform high-dimensional vector spaces from one layer to the next via matrix operations and nonlinearities—continuous, geometric computations with no symbolic counterpart. At an implementational level, both symbolic systems and deep learning models enjoy a degree

of hardware independence, but LLMs mirror the continuous representational formats and learning dynamics of biological neural networks far more closely than any rule-based symbolic system (see conspicuous gaps between artificial and biological neural networks). The divergence between these computational systems persists across Marr's levels of analysis.

We should be careful not to invoke Marr's framework to insulate symbolic theories from challenges arising across different levels of analysis, because these levels are not fully independent.[182] Artificial neural networks offer an alternative that satisfies constraints across multiple levels, without privileging one level over the others. Forcing LLMs into a symbolic description to preserve the appearance of independence between levels does not clarify how these models compute and risks obscuring different avenues of explanation.

THE CHALLENGE OF UNITING THE STRUCTURES OF LANGUAGE IN THE HUMAN BRAIN

The earliest theories of how the brain gives rise to language were based on behavioral deficits resulting from localized brain injuries[183,184]: separate, albeit interconnected, left-lateralized systems specialized for speech production and speech comprehension.[185] Although this classical model is now understood to be highly oversimplified,[186–189] the goal of identifying distinct brain areas for different language functions continued to shape later work. With the popularization of noninvasive neuroimaging came an explosion of experimental work aimed at teasing apart specialized systems for different components of language as posited by linguistic theory, for example, localizing neural systems that track syntactic complexity.[190–192] Influential work argued that syntactic combination—*Merge* in Minimalist terms—is localized in a specific subregion of inferior frontal cortex.[193,194] More recent work, however, reports that syntactic and semantic processing are inseparable and distributed across essentially the entire language network.[195,196]

Over the years, researchers have tried to synthesize the sprawling body of experimental results into larger-scale interpretive frameworks; for example, by appealing to distinct processing pathways.[197–200] We cannot provide a thorough treatment of this large body of work here and instead direct readers to other reviews.[199,201–205] Despite these efforts, the results resist a clear synthesis. To borrow a Chomskyan[206] metaphor: we have accumulated many beautiful experimental "butterflies"—but the explanatory principles that would unite them remain elusive.

Why has it been so difficult to piece experimental findings together? Some of the frustration may stem from a reliance on tightly controlled manipulations targeting the neural correlates of very specific linguistic phenomena—a particular syntactic operation, isolated words, decontextualized sentences—that do not generalize well across contexts. Outside the laboratory, language is rarely decontextualized. We use it to tell stories, share experiences, and communicate what matters to us and to our interlocutors.[207,208] Any given word or sentence derives its meaning from the narrative and communicative context surrounding it. Controlled experiments, by design, factor out precisely this kind of contextual situatedness. The result is findings that are internally valid but difficult to generalize to natural language use, a limitation that has led many researchers to advocate for more naturalistic paradigms.[34,209–212] Targeted manipulations yield targeted answers. What we lack are the principles that connect them.

There is a deeper concern about whether experimental work derived from rule-based linguistic models will ever make contact with the neural computations that support natural language processing. Should we expect, for example, dissociable neural systems for syntax and semantics, just because linguistic theory treats them as distinct? Language emerges from human brain activity in some way, so the underlying processes must be interlinked; yet the formal constructs we have developed to describe linguistic structure stubbornly refuse to map onto the neural code. This returns to Poeppel's[176] mapping problem: can we map the descriptive structures in linguistics (e.g., a hierarchical noun phrase or the process of linearization) onto neuroscientific mechanisms (e.g., integrate-and-fire neuronal dynamics or long-term potentiation)? Decades of work suggest the answer is not straightforward. What would a genuinely useful computational model of the neural processes that support language in the brain look like? At minimum, it would need to (1) illuminate the underlying mechanisms that support natural language processing and (2) generate testable predictions about how the brain comprehends and produces language. In what follows, we argue that LLMs offer a glimpse of what such a model might look like.

We do not claim that these models fully resolve the challenges outlined above, nor are they even approximately isomorphic with the human brain (see conspicuous gaps between artificial and biological neural networks below). Nonetheless, following the example of deep learning in visual neuroscience,[213,214] these models offer a new theoretical framework from which to proceed. In the spirit of Dobzhansky's[215] "nothing in biology makes sense except in the light of evolution," we argue that these models can provide a useful explanatory framework for understanding how the diverse range of experimental findings may *emerge* from the interactions of simple neuron-like computing elements running up against an environment of real-world language.[33,216] The diversity of neurolinguistic phenomena, much like the diversity of biological organisms, may be an emergent property of a mindless fitting process operating over a richly structured world—in this case, a world of human language.[33]

In the next two sections, we ask whether Principles 1 and 2, the same principles underlying the success of LLMs, can help resolve these challenges for understanding language in the human brain. We unpack both principles in turn, with examples from a growing body of work that positions LLMs as models of the neural mechanisms underlying language processing in the brain. We direct the reader to other recent reviews for a more global perspective on the relationship between LLMs and the brain.[217–221]

Before turning to the evidence, we want to be explicit about what we can conclude from quantifying the alignment between an LLM's internal representations and human brain activity. A linear encoding model mapping LLM embeddings onto human neural activity demonstrates that the model captures certain features of the stimulus or task (e.g., the relationships between words as they unfold in a narrative) that are also encoded in the neural activity, ideally in a way that generalizes to novel

examples.[222–224] It does not, alone, demonstrate the brain employs a similar architecture or algorithm.[225] LLMs differ substantially from the brain in their circuit architecture, learning rule, objective function, and training diet. More specific claims require theoretically motivated model comparisons and experiments designed to disentangle competing models.[226,227] The claim we defend is narrower and, we think, more defensible. Both systems converge on a similar representational format: a distributed, high-dimensional population code for solving the same underlying learning problem. With that caveat in place, we turn to the evidence.

Principle 1 for the human brain: A unified embedding space

Neural population codes, similar to the kind employed by LLMs, are nearly ubiquitous in theoretical and empirical neuroscience[103–108]: population codes in sensory, motor, and cognitive systems are transformed both moment-to-moment by recurrent connections and by longer-distance connections from one neural population to another. For example, visual inputs are encoded in high-dimensional patterns of neural activity.[228,229] These distributed representations are transformed from one processing stage to another[230–232] to disentangle behaviorally relevant perceptual and conceptual representations.[233–238] In the same vein, neural population codes are thought to support both motor preparation[239,240] and motor execution.[241,242] Similar to LLMs, individual neurons do not selectively encode particular stimulus or task features but rather exhibit graded tuning across a variety of features and nonlinear combinations of features (Figure 1B). This heterogeneous tuning and mixed selectivity, combined with population-level geometric computing, is thought to undergird our capacity for abstract, flexible cognition.[132,133,243]

These ideas have begun to permeate the neuroscience of natural language as well. For example, recent work using high-density intracranial recordings has shown that LLMs can predict population-level activity vectors in cortical language areas (e.g., inferior frontal gyrus; IFG).[244] A simple linear mapping between the LLM embeddings and human neural activity is sufficient to accurately predict activity vectors for left-out words in novel contexts, suggesting that the geometry of the LLM's embedding space is remarkably well-aligned with the geometric relationships between patterns of human brain activity. Work on sentence processing suggests that the language network populates individual words into a high-dimensional representation of the ongoing context.[245] Single-unit recordings in language areas show that individual neurons exhibit mixed tuning for different linguistic features and are highly sensitive to local context[246,247]—a signature of high-dimensional neural population codes.[133] A highly distributed code for language may appear to be at odds with certain empirical results—for example, putative "concept cells"[248–250] or highly selective deficits resulting from localized brain damage.[251,252] We contend that these apparent discrepancies suggest promising directions for future research: for example, introducing sparsity[131,253] and functional-topographic organization[254–256] into language models may help resolve this tension.

In the following, we outline two cases that showcase the explanatory power of Principle 1 in the human neuroscience of natural language. We highlight recent work that directly compares LLMs to human brain activity and suggest that reframing our assumptions about the neural code for language may resolve certain tensions among prior experimental findings.

Case 1: Syntax and semantics in a unified embedding space

Historically, in linguistics, the meaning of words (semantics) and the rules used to combine them (syntax) have been described using very different formal structures, e.g., tree-like syntactic structures[2] and semantic graphs.[45] How are these seemingly different kinds of structure encoded in neural activity? Despite efforts to experimentally disentangle syntax from semantics in the human brain,[257] a growing body of neuroimaging work has suggested that syntactic processing is widely distributed across the language network and closely intertwined with word meanings.[258–269] For example, Shain and colleagues[196] observed a syntactic effect of constituent length throughout the language network and found that the same areas were also modulated by semantic content.

Recent work leveraging LLMs has come to a similar conclusion.[266–270] Kumar, Sumers, and colleagues,[271] for example, deconstructed the circuitry of BERT[76] into the transformations implemented by individual attention heads and mapped these transformations onto fMRI activity during natural story listening. In BERT, these attention heads have been found to exhibit emergent functional specialization for specific syntactic operations.[147,149,272] Kumar, Sumers, and colleagues found that attention heads provided strong predictions across many language areas, and specific features of the transformations, such as look-back distance, captured trends in functional organization across the language network. The context-rich compositional transformations learned by BERT predicted brain activity much better than context-free syntactic dependencies. Despite a long-standing theoretical distinction, mounting evidence shows that the brain does not compartmentalize syntax and semantics.

Modern neural network models for language appear to agree with the brain on this front. As we previously discussed, rather than compartmentalizing syntactic rules from lexical semantics,[174] LLMs fuse these seemingly different kinds of linguistic structure in a high-dimensional embedding space.[78–80,160] The success of LLMs indicates that the structures of both syntax and semantics can be distilled from contextual patterns across words using the same statistical learning mechanisms and encoded in the same geometric format. The theoretical distinction between syntax and semantics may be largely immaterial in terms of neural computation. If the brain follows a similar strategy, syntactic and semantic structures are expected to be widespread, co-localized, and functionally entangled, encoded by overlapping subpopulations of neurons in a mixed and graded fashion. Insofar as a valid cognitive representation must be implementable in a neural system,[176,180] we find it highly relevant that neither human brains[196] nor LLMs[140,160] appear to strictly distinguish syntax and lexical semantics.

Case 2: Transforming the acoustics of speech into meaningful language

How does the brain extract meaning from a continuous speech signal? Traditionally, this process has been divided into an explicit hierarchy: acoustic signals are transformed into phonemes or syllables, indexed to lexical representations, and

CellPress
OPEN ACCESS

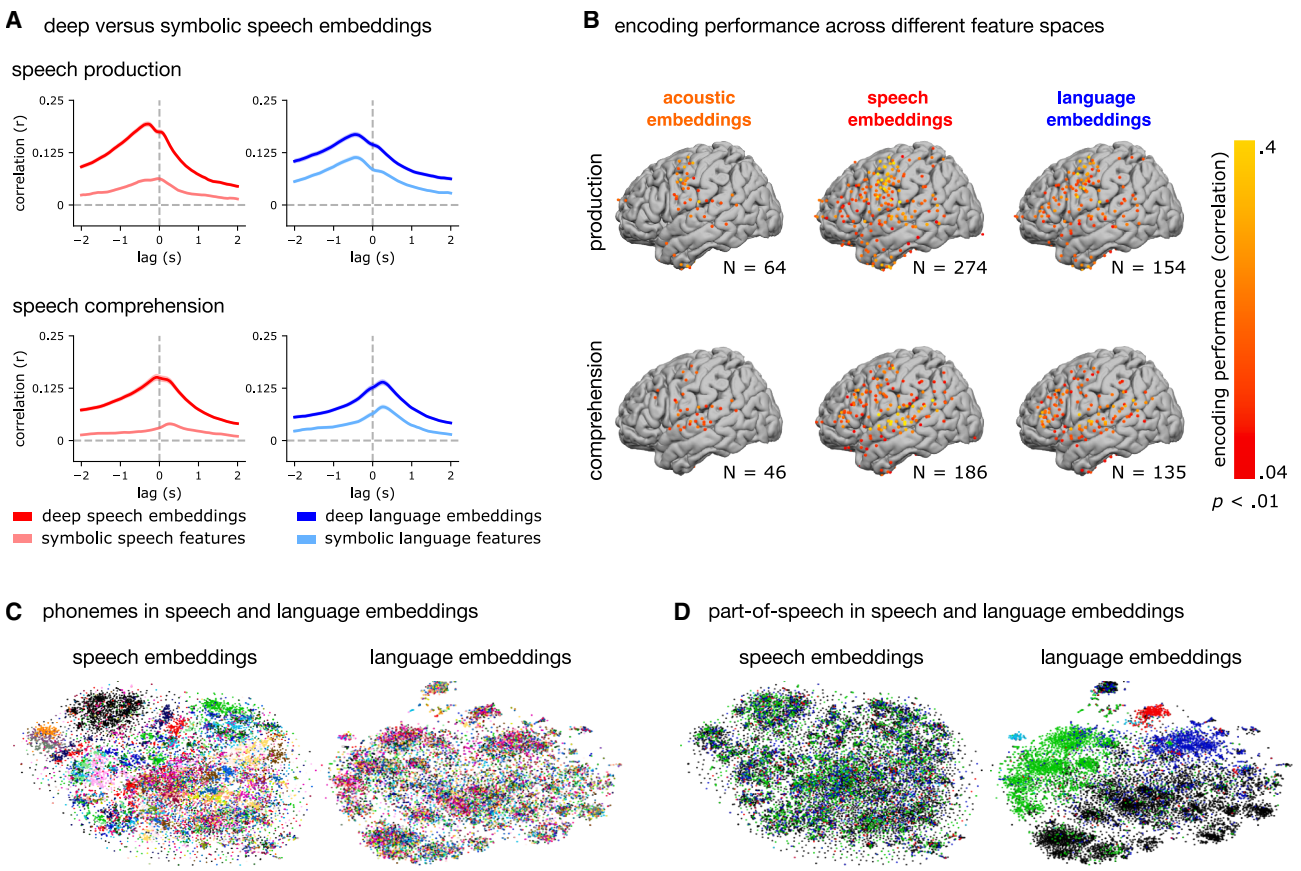


Figure 2. A unified model for speech and language

The Whisper model learns to transform acoustic inputs into text via a joint encoder-decoder architecture. We extracted both speech (encoder) embeddings and language (decoder) embeddings from the same model. We evaluated their alignment to human neural activity for both speech production and comprehension across ~100 h of natural conversations.

(A) Both speech embeddings and language embeddings dramatically outperform symbolic speech features (e.g., phonemes) and symbolic language features (e.g., part-of-speech) during both production and comprehension. Error bands indicate standard error of the mean encoding performance across left-out test conversations.

(B) Acoustic embeddings extracted before the transformer block in the speech encoder best predict electrode activity in articulatory areas during production and in perceptual areas during comprehension. Speech embeddings yield more widespread encoding performance, with strong predictions in areas associated with both speech articulation and perception. Language embeddings yield the strongest predictions in higher-level language areas such as the IFG and the angular gyrus. Electrodes shown are significant at $p < 0.01$ after Bonferroni correction.

(C) Phoneme categories can be recovered from the speech embedding space but are not as clearly encoded in the language embedding space.

(D) On the other hand, part-of-speech categories can be recovered from the language embedding space but are not as evident in the speech embedding space. In both cases, the model does not explicitly use symbolic representations like phonemes or parts of speech in its internal computations. Rather, these structures are learned from surface-level language inputs and emerge from the geometry of the embedding space. Adapted with permission from Goldstein and colleagues.[178]

assembled via syntactic operations into higher-level meaning (Figure 1A).[199,273] Early models of neural speech representations relied on handcrafted features such as spectrotemporal filters or phonetic categories.[274–277] Kell and colleagues[278] took a significant step forward, showing that a deep neural network trained simply to map sound waveforms onto words reproduces human word-recognition behavior and learns internal representations that predict neural activity better than any handcrafted model.[177,279–281]

Goldstein and colleagues[178] extracted three internal representations from a transformer-based speech-to-text model called Whisper:[282] (1) low-level acoustic embeddings from an early layer of the audio encoder, (2) mid-level speech embeddings from the final layer of the audio encoder, and (3) high-level contextual word embeddings from a late intermediate layer of the text decoder. All three types of embeddings predicted brain activity across many electrodes during hours of held-out natural conversations, and all three dramatically outperformed symbolic features like phonemes and parts of speech (Figure 2A). Acoustic embeddings predicted a relatively focal set of electrodes in sensory and motor regions (Figure 2B). Speech embeddings predicted a much broader set of electrodes, with the strongest alignment in superior temporal gyrus (STG) during comprehension and in motor areas during production, suggesting a strong overlap in features underlying perception and articulation. Finally, contextual word embeddings best predicted higher-level language areas, including IFG and the angular gyrus, during both comprehension and production.

Critically, although phoneme categories can be partially recovered from Whisper's speech embeddings and part-of-speech categories can be recovered from the model's language embeddings (Figures 2C and 2D), the model does not encode discrete, symbolic representations of these categories internally. Rather, they emerge as approximate byproducts of the embedding geometry shaped entirely by the task of predicting speech and text. Different levels of linguistic structure—acoustic, phonetic, syntactic, semantic, and contextual patterns—are fused into a continuous representational space. Goldstein and colleagues[178] show, for instance, that higher-level contextual embeddings (from the decoder) predict brain activity more accurately across many electrodes when they incorporate information from mid-level speech embeddings (from the encoder). A variance partitioning analysis shows that many electrodes exhibit mixed tuning for mid-level speech and high-level contextual-semantic features. This "soft hierarchy" differs from the compartmentalized picture of classical linguistics but offers a unified account of how the brain might integrate information across levels in a way that resonates with the mixed tuning observed in individual human neurons.[247,250,283,284]

Principle 2 for the human brain: Context-driven prediction

From infancy, humans are exquisitely sensitive to the statistical structure of their environment, leveraging a continuous stream of experience to learn patterns that predict upcoming events.[285,286] This predictive capacity operates on conditional probabilities and supports language comprehension across levels of representation, modalities, and timescales.[287–289] Predictive coding theories of brain function have gained considerable traction on this basis: for example, the N400 event-related potential scales with the predictability of a word given its preceding context.[290,291]

Modern deep neural networks provide a proof of principle that linguistic structure can be acquired through statistical learning alone.[162] These models are trained on a simple self-supervised objective, predicting the next sample in a sequence, where the upcoming input itself serves as an imminently available teaching signal. In speech recognition, self-supervised models trained in this way learn internal representations that predict human brain activity more accurately than handcrafted acoustic features and supervised models.[162,177,179,279–281] The same pattern holds for text: across numerous studies, text-based LLMs that learn to contextualize their internal representations for next-token prediction capture human neural activity far better than handcrafted linguistic features or context-free semantic representations.[267,271,292–298] Recent work has begun to chart out how the higher-level contextual features learned by LLMs emerge over time in the developing human brain.[299]

In the following, we discuss two cases where Principle 2 can shed light on how humans learn to efficiently process language in context and ultimately converge on shared meaning for communication.

Case 3: Context-driven predictive processing

A number of studies have explored parallels between predictive processing in LLMs and the brain,[294,295,297,300] suggesting, for example, that as LLMs become better at predicting upcoming tokens, they also become more closely aligned with human brain activity.[227,301] Researchers have also used LLMs to show that the brain likely deploys predictions at multiple levels of granularity[302] and across multiple words.[303]

Goldstein and colleagues[296] capitalized on the high temporal resolution of human intracranial recordings to more directly ascertain whether features of upcoming words are encoded in neural activity before those words are actually heard. They found that encoding performance for LLM embeddings begins to rise well before word onset. However, this could be partly because these models, by design, encode contextual information from previous words into the current embedding. Follow-up analyses ruled out this confound: pre-onset encoding persists even with non-contextual embeddings, after removing highly predictable bigrams, and after regressing out the embeddings of preceding words.[296,304] The nature of this pre-onset encoding, and to what extent it can be dissociated from stimulus-driven dependencies, remains a subject of active debate.[305] One clue may be the observation that pre-onset encoding does not appear to be a simple pre-activation of the post-onset word representation, since pre-onset predictions do not transfer directly to post-onset neural activity.[304,305] This pattern is consistent with the framework we propose here. We would not expect the brain to pre-activate a discrete representation of a specific upcoming word—that would require committing to a single prediction in advance. Instead, the brain likely maintains a continuous, distributed representation that simultaneously encodes the current context and probabilistically anticipates a range of features consistent with the likely upcoming words, without committing to one of them until the word actually arrives.

These studies leverage the predictive capacity of LLMs to lend new neuroscientific evidence to longstanding arguments that the brain deploys predictive processing to facilitate language comprehension. However, LLMs also suggest a deeper connection: moment-by-moment predictions furnish the model with a feedback signal that drives language acquisition. LLMs not only learn to predict but also use prediction to learn.[306] Recent work has suggested that LLMs, driven by the objective of next-token prediction, follow a trajectory similar to that of children in acquiring the structures of speech[307] and language.[308] While the degree to which prediction drives language learning in human development is an active area of research,[309–314] the success of LLMs highlights the power of prediction in learning and provides a plausible explanation for how multiple, seemingly different kinds of linguistic features (e.g., syntactic and semantic features) could be predicted jointly by a generic algorithm operating over a high-dimensional embedding space.

Case 4: Learning a shared geometry for communicating our thoughts

To understand one another, we must come to agree on the meaning of words in context. Language is typically a highly interactive process between two or more people, in which individuals dynamically alternate between speaker and listener (or reader, writer, signer, and so on). Conversation is the most basic form of language use, and speakers interactively align with one another according to a shared understanding of the world.[8–10] In science, for example, we frequently use a scaffold of familiar words in novel combinations to inspire new ideas in the reader's mind. Modern

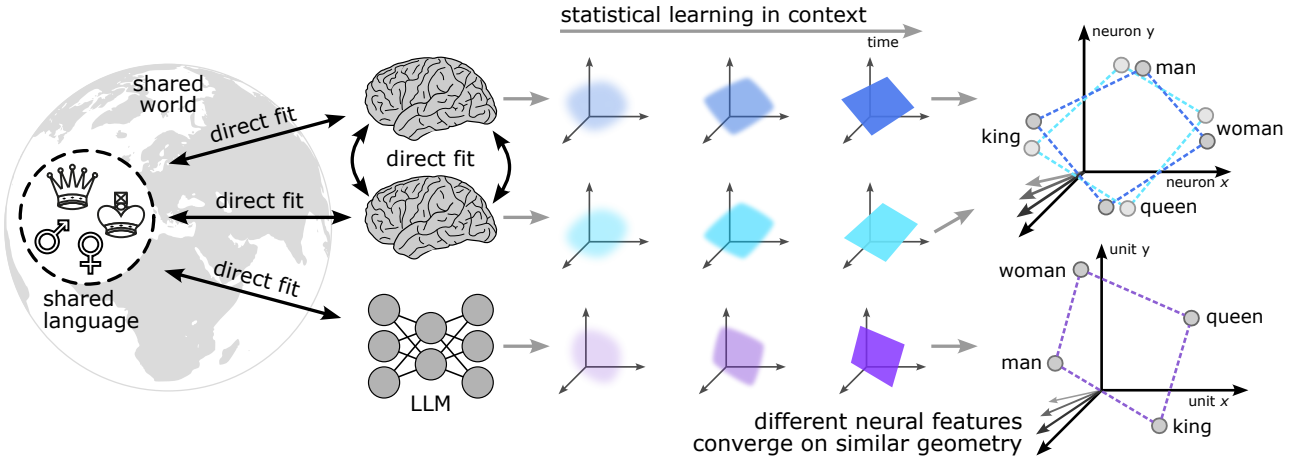


Figure 3. Different neural systems converge on a shared geometry for language

No two humans have exactly the same neuronal tuning for language. Nonetheless, we each ultimately converge on a geometry of internal neural states that is similar enough across language users to allow us to communicate with one another. We learn this geometry of internal states by interacting with other language users in a shared world. Even native speakers of different languages converge on a remarkably similar neural population code.[327] In the same vein, LLMs learn a neural population code with representational geometry sufficiently similar to our own for them to communicate with us meaningfully and usefully. They learn this geometry by encoding the structures of language in a neural population code using self-supervised statistical learning. Note that the parallelogram depicted here is a gross oversimplification for the purpose of visualization: the internal geometries learned by both humans and LLMs are inflected by context to differentiate all manner of constructions such that "man" differs in its role in "man and woman" and "you're the man!", and "queen" differs meaningfully from "the queen of shoes" and "Freddie Mercury's Queen."

language models, like ChatGPT, have become increasingly conversational in recent years. While these advances were built on a foundation of next-word prediction, conversational models appear to require intensive training on conversational data[315] and human feedback[119,121,125,316,317] to align with the conventions of human dialogue. Human neuroscience generally agrees that both language production and comprehension rely on shared neural systems. Likewise, language comprehension and language production in LLMs rely fundamentally on the same neural machinery. The use of special tokens (like the "end-of-text" delimiter token) allows the model to learn when to relinquish the "speaking" role and start "listening."

Generative LLMs offer a robust framework for modeling the neural basis of language production in natural conversations.[178,318,319] Zada and colleagues[298] analyzed a rare dataset of intracranial recordings acquired simultaneously in dyadic pairs of epilepsy patients while they engaged in free-form, face-to-face conversations. Using GPT-2 embeddings to predict neural activity simultaneously in both brains during speaking and listening, the authors were able to model how thoughts are transmitted from one brain to another: linguistic content emerges in the speaker's brain just before word articulation and re-emerges in the listener's brain after word onset, word by word, in natural conversations. They show that contextual embeddings from GPT-2 better capture the shared content of speaker–listener coupling in natural conversations than lexical, syntactic, or articulatory features do. In a second hyperscanning study, Zada and colleagues[320] used the superior spatial coverage of fMRI to map model-based brain-to-brain coupling. They found that contextual embeddings also capture coupling beyond the language network, extending into default-mode areas associated with theory of mind.[321–324] These findings suggest that production and comprehension rely on shared neural mechanisms for language processing and that we interactively align our internal neural representations during conversations.

How do we ultimately align our internal representations of linguistic structure well enough to understand one another? We are not born with knowledge of the context-specific meaning of words or how to convey our thoughts to one another. We do not have direct access to each other's brain activity, and no two people have direct neuron-to-neuron correspondence anyway: even if we *could* transmit our brain activity directly to another person, there would be no way to map it onto their unique configuration of neurons.[325] We ultimately align through interactive conversations, in which we listen to and speak with one another in a shared social world (Figure 3).[326] Through this interactive process, we *learn* a configuration of internal neural representations for language (a representational geometry) that roughly matches the internal geometry of other speakers in our community. Amazingly, LLMs seem to learn an internal geometry similar enough to ours, despite their radically different neural architectures (see conspicuous gaps between artificial and biological neural networks), that they can produce conversational, human-like language. These models demonstrate that shared linguistic input, shared objectives, statistical learning, and an expressive embedding space are sufficient for neural systems to converge on a shared geometry for language across a community of language users.

CONSPICUOUS GAPS BETWEEN ARTIFICIAL AND BIOLOGICAL NEURAL NETWORKS

All models make simplifying assumptions. Artificial neural networks, including LLMs, embody several commitments about what features of neural information processing are most important. From a neuroscientific perspective, artificial neural

networks do not respect the diversity and complexity of individual neurons;[328,329] they often simplify neuronal spiking to a rate-like code;[330] they do not strictly adhere to biological learning algorithms;[331] and they do not reflect well-known biological circuit motifs.[332] Instead, LLMs make the following commitments: the structures of language are represented across simple, interconnected computing elements (akin to neurons); these units are neuron-like in the sense that they aggregate their inputs and generate an output according to a nonlinear activation function (akin to a voltage threshold); the memory of the network is encoded in the connection weights between units (akin to the synaptic strength between neurons); and these weights are iteratively updated according to a relatively simple learning rule (akin to synaptic plasticity). Like ants in an ant colony, no individual neuron is in control or "knows" what's going on in a global sense—instead, cognition and complex behavior *emerge* from the dynamic interactions between units.

Certain features of LLMs are at odds with human language learning and processing. Modern LLMs are trained on orders of magnitude more language input than humans receive (it would take thousands of years for a human to read the amount of text used to train newer GPT models).[333] That said, recent results from the 1,000 First Days Project[307,334] and the BabyLM Challenge[335–337] suggest that certain linguistic skills, such as syntactic competence, may be achievable with developmentally plausible input volumes.[338] Transformer-based models also retain veridical access to thousands of tokens of prior context at any given time. Humans, on the other hand, must rapidly condense the present input to be ready for the next input.[339] Most text-based LLMs also lack fine-grained temporal dynamics, let alone the rich, endogenous oscillatory dynamics observed in human brain activity.[340–343] Recurrent neural network architectures[84,166,344] are more biologically plausible and may ultimately better capture these temporal dynamics.

Some have argued that LLMs fail to recapitulate human grammaticality judgments or comprehension,[345,346] but other work has shown strikingly high correlations between humans and LLMs when accounting for task design, evaluation metrics, and pragmatic sensitivity.[114,347–349] LLMs must be taught to be helpful[350] while humans may be more inherently cooperative.[11] Most LLMs learn from text alone, whereas children receive much richer, multimodal inputs. LLMs trained to perform next-word prediction are passively yoked to their linguistic inputs. They cannot actively explore or perturb their linguistic environment and cannot "choose their own adventure." Infants, on the other hand, are active agents who perturb their environment, including the behavior of social others, and observe the consequences of their actions. Moreover, while next-word prediction is a surprisingly powerful objective, an infant's learning objectives are surely richer, including social affiliation, capturing their caretaker's attention, and ensuring their needs are met.

Recent LLMs are becoming surprisingly good at answering novel questions in one- or few-shot tasks,[77,152,349] but they remain highly data-inefficient learners. What will it take for neural networks to become more efficient and human-like in their learning and behavior? One possibility is that these capabilities may emerge from different network architectures. For instance, introducing complementary memory systems[351,352] and/or control systems[353,354] may yield qualitative improvements in learning. Subjecting models to more ecological architectural constraints and learning objectives may also help.[33,355] We suspect that as models become more embodied, agentic, and social, they will learn even deeper, more human-like representations of the world.

CONCLUSION

In this perspective, we have argued that the recent success of LLMs is rooted in the same computational principles that govern biological brains. LLMs are not just feats of engineering; they are a working demonstration that context-based statistical learning is sufficient to implement natural language at scale in a neural population code. Rather than factorizing linguistic structure into separate representational formats, the way a linguist or engineer might decompose phonology, syntax, and semantics into separate modules, LLMs encode the full diversity of language in a unified, high-dimensional embedding space. Rather than relying on handcrafted features or rule-based grammars, LLMs learn to reproduce the patterns of language use directly from real-world contexts. The result is a fully mechanistic model that processes natural language and generates fluent, context-sensitive outputs, achieving this without any of the computational commitments that linguistic theory has long assumed were necessary. The primacy of context and the commitment to neural population codes in both LLMs and the human brain present a challenge to traditional theories of language. If LLMs can master language without "building in" any of the constructs of formal linguistics, perhaps the brain does not need them either.

Deep neural networks are sometimes dismissed as "black boxes," and there would indeed be little scientific value in replacing one black box with another. It might be tempting to abandon such models altogether or simply resign ourselves to their complexity. We resoundingly disagree with these sentiments. Despite their complexity, LLMs are not black boxes: they are fully transparent, and we can "open the box" to explore their inner workings and their alignment with human brain activity. Work toward a deeper, more mechanistic understanding of these models is already well underway,[131,150,153,169–171,356] with promising parallels to emerging theories in neuroscience.[133,240,357,358] LLMs can serve as "animal models" or "digital twins" for studying human language,[359,360] but we should resist projecting rule-based linguistic theories onto them.[32] To quote Rumelhart and McClelland,[87] "lawful behavior and judgments may be produced by a mechanism in which there is no explicit representation of the rule." Rather than search for old theories inside these models, we should let them inspire new ways of thinking.

Finally, despite their apparent complexity, there is something remarkably elegant about these models. What appear to be complex examples of computational machinery, like the objective function for next-word prediction or the self-attention circuit, can be summarized in one-line equations. Their power (and apparent complexity) stems from running a simple direct-fitting algorithm up against the rich complexity of their linguistic environment.[33] One might be tempted to disparage these models as "just a bunch of matrix multiplications," but herein lies their

simplicity and their utility for understanding language in the brain. Even simple computational principles, implemented at scale, can give rise to the incredible complexity and beauty of natural language.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) Collaborative Research in Computational Neuroscience (CRCNS) grant 2605721 (to S.A.N.) and the National Institute on Deafness and Other Communication Disorders (NIDCD), National Institutes of Health (NIH), CRCNS grant R01DC022534 (to U.H.).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Chomsky, N. (1957). Syntactic Structures (De Gruyter Mouton).
2. Chomsky, N. (1965). Aspects of the Theory of Syntax (MIT Press).
3. Chomsky, N., Roberts, I., and Watumull, J. (2023). Noam Chomsky: The False Promise of ChatGPT (The New York Times). https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.
4. Humboldt W. von On Language: On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species. In: Losonsky M., editor. Cambridge University Press; 1836.
5. Fodor, J.A., and Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: a critical analysis. Cognition *28*, 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.
6. Pinker, S., and Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. Cognition *28*, 73–193. https://doi.org/10.1016/0010-0277(88)90032-7.
7. Marcus, G.F. (2001). The Algebraic Mind: Integrating Connectionism and Cognitive Science (MIT Press). https://doi.org/10.7551/mitpress/1187.001.0001.
8. Grice, P. (1991). Studies in the Way of Words (Harvard University Press).
9. Clark, H.H., and Brennan, S.E. (1991). Grounding in communication. In Perspectives on Socially Shared Cognition, L.B. Resnick, J.M. Levine, and S.D. Teasley, eds. (American Psychological Association), pp. 127–149. https://doi.org/10.1037/10096-006.
10. Pickering, M.J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behav. Brain Sci. *27*, 169–190. 90; discussion. https://doi.org/10.1017/s0140525x04000056.
11. Tomasello, M. (2008). Origins of Human Communication (MIT Press). https://doi.org/10.7551/mitpress/7551.001.0001.
12. Wittgenstein, L. (1953). Philosophical Investigations. Fourth Edition, P. Hacker and J. Schulte, eds. (Wiley-Blackwell).
13. Hopper, P.J., and Thompson, S.A. (1984). The discourse basis for lexical categories in universal grammar. Language *60*, 703–752. https://doi.org/10.1353/lan.1984.0020.
14. Langacker, R. (1986). An introduction to cognitive grammar. Cogn. Sci. *10*, 1–40. https://doi.org/10.1016/s0364-0213(86)80007-6.
15. Givón, T. (1995). Functionalism and Grammar (John Benjamins Publishing Company publishing). https://doi.org/10.1075/z.74.
16. Tomasello, M. (2005). Constructing a Language: A Usage-Based Theory of Language Acquisition (Harvard University Press). https://doi.org/10.4159/9780674044395.
17. Goldberg, A.E. (2006). Constructions at Work: The Nature of Generalization in Language (Oxford University Press).
18. Goldberg, A.E. (2019). Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions (Princeton University Press). https://doi.org/10.2307/j.ctvc772nn.
19. Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T.; The "Five Graces Group" (2009). Language is a complex adaptive system: position paper. Lang. Learn. *59*, 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x.
20. Bybee, J. (2010). Language, Usage and Cognition (Cambridge University Press). https://doi.org/10.1017/CBO9780511750526.
21. Zipf, G.K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology (Addison-Wesley press).
22. Hockett, C.F. (1960). The origin of speech. Sci. Am. *203*, 89–96. https://doi.org/10.1038/scientificamerican0960-88.
23. Hawkins, J.A. (2004). Efficiency and Complexity in Grammars (Oxford University Press). https://doi.org/10.1093/acprof:oso/9780199252695.001.0001.
24. Piantadosi, S.T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. Proc. Natl. Acad. Sci. USA *108*, 3526–3529. https://doi.org/10.1073/pnas.1012551108.
25. Piantadosi, S.T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. Cognition *122*, 280–291. https://doi.org/10.1016/j.cognition.2011.10.004.
26. Kemp, C., and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. Science *336*, 1049–1054. https://doi.org/10.1126/science.1218811.
27. Du Bois, J.W. (2014). Towards a dialogic syntax. Cogn. Linguist. *25*, 359–410. https://doi.org/10.1515/cog-2014-0024.
28. Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. Proc. Natl. Acad. Sci. USA *112*, 10336–10341. https://doi.org/10.1073/pnas.1502134112.
29. Gibson, E., Futrell, R., Piantadosi, S.P., Dautriche, I., Mahowald, K., Bergen, L., and Levy, R. (2019). How efficiency shapes human language. Trends Cogn. Sci. *23*, 389–407. https://doi.org/10.1016/j.tics.2019.02.003.
30. Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., and Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. Proc. Natl. Acad. Sci. USA *118*, e2025993118. https://doi.org/10.1073/pnas.2025993118.
31. Fedorenko, E., Piantadosi, S.T., and Gibson, E.A.F. (2024). Language is primarily a tool for communication rather than thought. Nature *630*, 575–586. https://doi.org/10.1038/s41586-024-07522-w.
32. Weissweiler, L., Mahowald, K., and Goldberg, A.E. (2025). Linguistic generalizations are not rules: impacts on evaluation of LMs. In Proceedings of the Second International Workshop on Construction Grammars and NLP, C. Bonial, M. Torgbi, L. Weissweiler, A. Blodgett, K. Beuls, P. Van Eecke, and H.T. Madabushi, eds. (Association for Computational Linguistics), pp. 61–74. https://aclanthology.org/2025.cxgsnlp-1.7/.
33. Hasson, U., Nastase, S.A., and Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. Neuron *105*, 416–434. https://doi.org/10.1016/j.neuron.2019.12.002.
34. Nastase, S.A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. Neuroimage *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254.
35. Austin, J.L. (1975). How to Do Things with Words. Second Edition, J.O. Urmson and M. Sbisà, eds. (Harvard University Press). https://doi.org/10.1093/acprof:oso/9780198245537.001.0001.
36. Bates, E., and MacWhinney, B. (1989). Functionalism and the Competition Model. In The Crosslinguistic Study of Sentence Processing, B. MacWhinney and E. Bates, eds. (Cambridge University Press), pp. 3–76.

37. Croft, W. (2001). Radical Construction Grammar: Syntactic Theory in Typological Perspective (Oxford University Press). https://doi.org/10.1093/acprof:oso/9780198299554.001.0001.

38. M. Haspelmath, M.S. Dryer, D. Gil, and B. Comrie, eds. (2005). The World Atlas of Language Structures (Oxford University Press).

39. Chomsky, N., and Halle, M. (1968). The Sound Pattern of English. https://doi.org/10.1037/031431.

40. Kenstowicz, M., and Kisseberth, C. (1979). Generative Phonology: Description and Theory (Academic Press).

41. Hockett, C.F. (1954). Two models of grammatical description. Word *10*, 210–234. https://doi.org/10.1080/00437956.1954.11659524.

42. Halle, M., and Marantz, A. (1994). Some key features of distributed morphology. MIT Working Pap. Linguist. *21*, 275–288. https://web.mit.edu/morrishalle/pubworks/papers/1994_Halle_Marantz_Key_Features_of_Distributed_Morphology.pdf.

43. Bobaljik, J.D. (2017). Distributed Morphology. In Oxford Research Encyclopedia of Linguistics https://doi.org/10.1093/acrefore/9780199384655.013.131.

44. Kayne, R.S. (2016). Connectedness and Binary Branching (Walter de Gruyter). https://doi.org/10.1515/9783111682228.

45. Quillian, M.R. (1967). Word concepts: a theory and simulation of some basic semantic capabilities. Behav. Sci. *12*, 410–430. https://doi.org/10.1002/bs.3830120511.

46. Jackendoff, R. (1990). Semantic Structures (MIT Press).

47. Wierzbicka, A. (1996). Semantics: Primes and Universals (Oxford University Press). https://doi.org/10.1093/oso/9780198700029.001.0001.

48. Kratzer, A., and Selkirk, E. (2020). Deconstructing information structure. Glossa *5*, 113. https://doi.org/10.5334/gjgl.968.

49. Bresnan, J., Deo, A., and Sharma, D. (2007). Typology in variation: a probabilistic approach to be and n't in the Survey of English Dialects. Engl. Lang. Linguist. *11*, 301–346. https://doi.org/10.1017/s1360674307002274.

50. Essegbey, J. (2008.). Intransitive verbs in ewe and the unaccusativity hypothesis. In Crosslinguistic Perspectives on Argument Structure: Implications for, M. Bowerman and P. Brown, eds. (Learnability Lawrence Erlbaum Associates), pp. 213–230.

51. Chomsky, N. (1995). The Minimalist Program (MIT Press).

52. Hauser, M.D., Chomsky, N., and Fitch, W.T. (2002). The faculty of language: what is it, who has it, and how did it evolve? Science *298*, 1569–1579. https://doi.org/10.1126/science.298.5598.1569.

53. Berwick, R.C., and Chomsky, N. (2016). Why Only Us: Language and Evolution (MIT Press). https://doi.org/10.7551/mitpress/9780262034241.001.0001.

54. Berwick, R.C., and Chomsky, N. (2017). Why only us: recent questions and answers. J. Neurolinguist. *43*, 166–177. https://doi.org/10.1016/j.jneuroling.2016.12.002.

55. Chomsky, N., and Smith, N. (2000). New Horizons in the Study of Language and Mind (Cambridge University Press). https://doi.org/10.1017/CBO9780511811937.

56. Chomsky, N. (2007). Approaching UG from Below. In Interfaces + Recursion = Language? Chomsky's Minimalism and the View from Syntax-Semantics, U. Sauerland and H.-M. Gärtner, eds. (Mouton de Gruyter), pp. 1–29. https://doi.org/10.1515/9783110207552-001.

57. Rizzi, L. (1997). The fine structure of the left periphery. In Kluwer International Handbooks of Linguistics, L. Haegeman, ed. (Springer Netherlands), pp. 281–337. https://doi.org/10.1007/978-94-011-5420-8_7.

58. Rizzi, L., and Cinque, G. (2016). Functional categories and syntactic theory. Annu. Rev. Linguist. *2*, 139–163. https://doi.org/10.1146/annurev-linguistics-011415-040827.

59. Evans, N., and Levinson, S.C. (2009). The myth of language universals: language diversity and its importance for cognitive science. Behav. Brain Sci. *32*, 429–448. https://doi.org/10.1017/S0140525X0999094X.

60. Majid, A., Roberts, S.G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., Woll, B., LeLan, B., de Sousa, H., Cansler, B.L., et al. (2018). Differential coding of perception in the world's languages. Proc. Natl. Acad. Sci. USA *115*, 11369–11376. https://doi.org/10.1073/pnas.1720419115.

61. Bellingham, E., Evers, S., Kawachi, K., Mitchell, A., Park, S.-H., Stepanova, A., and Bohnemeyer, J. (2020). Exploring the representation of causality across languages: integrating production, comprehension and conceptualization perspectives. In Perspectives on Causation, E.A. Bar-Asher Siegal and N. Boneh, eds. (Springer International Publishing), pp. 75–119. https://doi.org/10.1007/978-3-030-34308-8_3.

62. Thompson, B., Roberts, S.G., and Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. Nat. Hum. Behav. *4*, 1029–1038. https://doi.org/10.1038/s41562-020-0924-8.

63. Sapir, E. (1939). Language: An Introduction to the Study of Speech (Harcourt Publishers, Brace and Company).

64. Lakoff, G. (1965). On the Nature of Syntactic Irregularity (Indiana University).

65. Winograd, T. (1972). Understanding natural language. Cogn. Psychol. *3*, 1–191. https://doi.org/10.1016/0010-0285(72)90002-3.

66. Bobrow, D. (1964). In Natural language input for a computer problem solving system (Massachusetts Institute of Technology) http://hdl.handle.net/1721.1/5922.

67. Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM *9*, 36–45. https://doi.org/10.1145/365153.365168.

68. Schank, R.C., and Tesler, L. (1969). A conceptual dependency parser for natural language. In Proceedings of the 1969 Conference on Computational Linguistics COLING '69 (Association for Computational Linguistics), pp. 1–3. https://doi.org/10.3115/990403.990405.

69. Woods, W.A. (1970). Transition network grammars for natural language analysis. Commun. ACM *13*, 591–606. https://doi.org/10.1145/355598.362773.

70. Winograd, T. (1980). What does it mean to understand language? Cogn. Sci. *4*, 209–241. https://doi.org/10.1207/s15516709cog0403_1.

71. Landauer, T.K., and Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. *104*, 211–240. https://doi.org/10.1037/0033-295X.104.2.211.

72. Manning, C.D., and Schütze, H. (1999). Foundations of Statistical Natural Language Processing (MIT Press).

73. Mikolov, T., Yih, W.-T., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746–751. https://aclanthology.org/N13-1090/.

74. Pennington, J., Socher, R., and Manning, C. (2014). GloVe: global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), A. Moschitti, B. Pang, and W. Daelemans, eds. (Association for Computational Linguistics), pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162.

75. Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training (OpenAI). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

76. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, *1 (Long and Short Papers)* (Association for Computational Linguistics), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

77. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds. (Curran Associates, Inc.), pp. 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

78. Manning, C.D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. Proc. Natl. Acad. Sci. USA *117*, 30046–30054. https://doi.org/10.1073/pnas.1907367117.

79. Linzen, T., and Baroni, M. (2021). Syntactic structure from deep learning. Annu. Rev. Linguist. *7*, 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035.

80. Pavlick, E. (2022). Semantic structure in deep learning. Annu. Rev. Linguist. *8*, 447–471. https://doi.org/10.1146/annurev-linguistics-031120-122924.

81. Misra, K., and Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: the case of the missing AANNs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 913–929. https://doi.org/10.18653/v1/2024.emnlp-main.53.

82. Scivetti, W., Wilcox, E., Schneider, N., Misra, K., and Weissweiler, L. (2026). Language models learn constructional semantics, not to mention syntax: investigating LM understanding of Paired-Focus constructions. In Proceedings of the 30th Conference on Computational Natural Language Learning (Association for Computational Linguistics), 250–267. https://doi.org/10.18653/v1/2026.conll-main.15.

83. Rumelhart, D.E., and McClelland, J.L.; The PDP Research Group (1986). Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations (The MIT Press). https://doi.org/10.7551/mitpress/5236.001.0001.

84. Elman, J.L. (1990). Finding structure in time. Cogn. Sci. *14*, 179–211. https://doi.org/10.1207/s15516709cog1402_1.

85. McClelland, J.L., and Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychol. Rev. *88*, 375–407. https://doi.org/10.1037/0033-295X.88.5.375.

86. Smolensky, P. (1988). On the proper treatment of connectionism. Behav. Brain Sci. *11*, 1–23. https://doi.org/10.1017/s0140525x00052432.

87. Rumelhart, D.E., and McClelland, J.L. (1986). On learning the past tenses of English verbs. In Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models (The MIT Press), pp. 216–271. https://doi.org/10.7551/mitpress/5236.003.0008.

88. Smolensky, P. (1991). The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn. In Connectionism and the Philosophy of Mind, T. Horgan and J. Tienson, eds. (Springer Netherlands), pp. 281–308. https://doi.org/10.1007/978-94-011-3524-5_13.

89. Christiansen, M.H., and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. Cogn. Sci. *23*, 157–205. https://doi.org/10.1207/s15516709cog2302_2.

90. McClelland, J.L., and Plaut, D.C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? Trends Cogn. Sci. *3*, 166–168. https://doi.org/10.1016/s1364-6613(99)01320-0.

91. Seidenberg, M.S., and Elman, J.L. (1999). Networks are not "hidden rules". Trends Cogn. Sci. *3*, 288–289. https://doi.org/10.1016/s1364-6613(99)01355-8.

92. Sutton, R. (2019). The Bitter Lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

93. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2001.08361.

94. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2108.07258.

95. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the Imitation Game: quantifying and extrapolating the capabilities of language models. Trans. J. Mach. Learn. Res. *2023*. https://openreview.net/forum?id=uyTL5Bvosj.

96. Cantlon, J.F., and Piantadosi, S.T. (2024). Uniquely human intelligence arose from expanded information capacity. Nat. Rev. Psychol. *3*, 275–293. https://doi.org/10.1038/s44159-024-00283-3.

97. Antonello, R., Vaidya, A., and Huth, A. (2023). Scaling laws for language encoding models in fMRI. In Advances in Neural Information Processing Systems, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds. (Curran Associates, Inc.), pp. 21895–21907. https://proceedings.neurips.cc/paper_files/paper/2023/hash/4533e4a352440a32558c1c227602c323-Abstract-Conference.html.

98. Hong, Z., Wang, H., Zada, Z., Gazula, H., Turner, D., Aubrey, B., Niekerken, L., Doyle, W., Devore, S., Dugan, P., et al. (2024). Scale matters: large language models with billions (rather than millions) of parameters better match neural representations of natural language. eLife. https://doi.org/10.7554/elife.101204.1.

99. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, eds. (Curran Associates, Inc.) https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

100. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848. ieeexplore.ieee.org.

101. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

102. Carvalho, W., and Lampinen, A. (2025). Naturalistic computational cognitive science: towards generalizable models and theories that capture the full range of natural behavior. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.20349.

103. Yuste, R. (2015). From the neuron doctrine to neural networks. Nat. Rev. Neurosci. *16*, 487–497. https://doi.org/10.1038/nrn3962.

104. Eichenbaum, H. (2018). Barlow versus Hebb: When is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? Neurosci. Lett. *680*, 88–93. https://doi.org/10.1016/j.neulet.2017.04.006.

105. Saxena, S., and Cunningham, J.P. (2019). Towards the neural population doctrine. Curr. Opin. Neurobiol. *55*, 103–111. https://doi.org/10.1016/j.conb.2019.02.002.

106. Vyas, S., Golub, M.D., Sussillo, D., and Shenoy, K.V. (2020). Computation through neural population dynamics. Annu. Rev. Neurosci. *43*, 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115.

107. Barack, D.L., and Krakauer, J.W. (2021). Two views on the cognitive brain. Nat. Rev. Neurosci. *22*, 359–371. https://doi.org/10.1038/s41583-021-00448-6.

108. Ebitz, R.B., and Hayden, B.Y. (2021). The population doctrine in cognitive neuroscience. Neuron *109*, 3055–3068. https://doi.org/10.1016/j.neuron.2021.07.011.

109. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł.U., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.) https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

110. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners (OpenAI).

https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

111. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds. (Curran Associates, Inc.) https://papers.nips.cc/paper_files/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

112. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N.A. (2021). All that's 'Human' is not gold: evaluating human evaluation of generated text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), C. Zong, F. Xia, W. Li, and R. Navigli, eds. (Association for Computational Linguistics), pp. 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565.

113. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.12712.

114. Cuneo, N., Graves, N., Rakshit, S., and Goldberg, A.E. (2025). For GPT-4 as with humans: information structure predicts acceptability of long-distance dependencies. In Proceedings of the Annual Meeting of the Cognitive Science Society. https://doi.org/10.48550/arXiv.2505.09005.

115. Jones, C.R., and Bergen, B.K. (2026). Large language models pass a standard three-party Turing test. Proc. Natl. Acad. Sci. USA *123*, e2524472123. https://doi.org/10.1073/pnas.2524472123.

116. Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. J. Mach. Learn. Res. *3*, 932–938. https://www.jmlr.org/papers/v3/bengio03a.html.

117. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. *2*, 303–314. https://doi.org/10.1007/bf02551274.

118. Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Netw. *2*, 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

119. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. Preprint at arXiv. https://doi.org/10.48550/arXiv.1909.08593.

120. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., and Finn, C. (2023). Direct preference optimization: your language model is secretly a reward model. In Advances in Neural Information Processing Systems, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds. (Curran Associates, Inc), pp. 53728–53741. https://doi.org/10.52202/075280-2338.

121. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds. (Curran Associates, Inc), pp. 27730–27744. https://doi.org/10.52202/068431-2011.

122. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V., and Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds. (Curran Associates, Inc.), pp. 24824–24837. https://doi.org/10.52202/068431-1800.

123. Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.; Gemini Team (2023). Gemini: a family of highly capable multimodal models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.11805.

124. OpenAI (2023). GPT-4 Technical Report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.08774.

125. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: open foundation and fine-tuned chat models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.09288.

126. Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku (Anthropic). https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

127. Lieto, A., Chella, A., and Frixione, M. (2017). Conceptual spaces for cognitive architectures: a lingua franca for different levels of representation. Biol. Inspired Cogn. Archit. *19*, 1–9. https://doi.org/10.1016/j.bica.2016.10.005.

128. Newell, A., and Simon, H. (1956). The logic theory machine—a complex information processing system. IEEE Trans. Inf. Theory *2*, 61–79. https://doi.org/10.1109/tit.1956.1056797.

129. Anderson, J.R. (1983). The Architecture of Cognition (Harvard University Press).

130. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom in: an introduction to circuits. Distill *5*. https://doi.org/10.23915/distill.00024.001.

131. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. (2022). Toy models of superposition. Transformer Circuits Thread. https://transformer-circuits.pub/2022/toy_model/index.html.

132. Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590. https://doi.org/10.1038/nature12160.

133. Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Curr. Opin. Neurobiol. *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010.

134. Osgood, C.E., Suci, G.J., and Tannenbaum, P. (1967). The Measurement of Meaning (University of Illinois Press).

135. Shepard, R.N., and Chipman, S. (1970). Second-order isomorphism of internal representations: shapes of states. Cogn. Psychol. *1*, 1–17. https://doi.org/10.1016/0010-0285(70)90002-2.

136. Edelman, S. (1998). Representation is representation of similarities. Behav. Brain Sci. *21*, 449–467. https://doi.org/10.1017/s0140525x98001253.

137. Gärdenfors, P. (2000). Conceptual Spaces: The Geometry of Thought (MIT Press). https://doi.org/10.7551/mitpress/2076.001.0001.

138. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4. https://doi.org/10.3389/neuro.06.004.2008.

139. Piantadosi, S.T., Muller, D.C.Y., Rule, J.S., Kaushik, K., Gorenstein, M., Leib, E.R., and Sanford, E. (2024). Why concepts are (probably) vectors. Trends Cogn. Sci. *28*, 844–856. https://doi.org/10.1016/j.tics.2024.06.011.

140. Rakshit, S., and Goldberg, A.E. (2025). Meaning-infused grammar: gradient acceptability shapes the geometric representations of constructions in LLMs. In Proceedings of the Second International Workshop on Construction Grammars and NLP, C. Bonial, M. Torgbi, L. Weissweiler, A. Blodgett, K. Beuls, P. Van Eecke, and H.T. Madabushi, eds. (Association for Computational Linguistics), pp. 151–157. https://aclanthology.org/2025.cxgsnlp-1.15/.

141. Rumelhart, D.E., and Abrahamson, A.A. (1973). A model for analogical reasoning. Cogn. Psychol. *5*, 1–28. https://doi.org/10.1016/0010-0285(73)90023-6.

142. Hewitt, J., and Manning, C.D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers), pp. 4129–4138. https://doi.org/10.18653/v1/N19-1419.

143. Zhao, H., Panigrahi, A., Ge, R., and Arora, S. (2023). Do transformers parse while predicting the masked word? In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, H.

Bouamor, J. Pino, and K. Bali, eds. (Association for Computational Linguistics), pp. 16513–16542. https://doi.org/10.18653/v1/2023.emnlp-main.1029.

144. Hofmann, V., Weissweiler, L., Mortensen, D.R., Schütze, H., and Pierrehumbert, J.B. (2025). Derivational morphology reveals analogical generalization in large language models. Proc. Natl. Acad. Sci. USA *122*, e2423232122. https://doi.org/10.1073/pnas.2423232122.

145. Potts, C. (2024). Characterizing English Preposing in PP constructions. J. Linguist. 1–39. https://doi.org/10.1017/s0022226724000227.

146. Diego-Simón, P., D'Ascoli, S., Chemla, E., Lakretz, Y., and King, J.-R. (2024). A polar coordinate system represents syntax in large language models. In Advances in Neural Information Processing Systems, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds. (Curran Associates, Inc.), pp. 105375–105396. https://doi.org/10.52202/079017-3344.

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

148. Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics) (Association for Computational Linguistics), pp. 3651–3657. https://doi.org/10.18653/v1/p19-1356.

149. Tenney, I., Das, D., and Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 4593–4601. https://doi.org/10.18653/v1/P19-1452.

150. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread. https://transformer-circuits.pub/2021/framework/index.html.

151. Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. Science *237*, 1317–1323. https://doi.org/10.1126/science.3629243.

152. Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J., and Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. In Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds. (Curran Associates, Inc.), pp. 18878–18891. https://doi.org/10.52202/068431-1371.

153. Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. (2022). In-context learning and induction heads. Transformer Circuits Thread. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

154. Raventós, A., Paul, M., Chen, F., and Ganguli, S. (2023). Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In Advances in Neural Information Processing Systems, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds. (Curran Associates, Inc.), pp. 14228–14246. https://doi.org/10.52202/075280-0626.

155. Elman, J.L. (2004). An alternative view of the mental lexicon. Trends Cogn. Sci. *8*, 301–306. https://doi.org/10.1016/j.tics.2004.05.003.

156. Webb, T., Holyoak, K.J., and Lu, H. (2023). Emergent analogical reasoning in large language models. Nat. Hum. Behav. *7*, 1526–1541. https://doi.org/10.1038/s41562-023-01659-w.

157. Park, C.F., Lee, A., Lubana, E.S., Yang, Y., Okawa, M., Nishi, K., Wattenberg, M., and Tanaka, H. (2025). ICLR: in-context learning of representations. In International Conference on Learning Representations, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, eds., pp. 53258–53284. https://proceedings.iclr.cc/paper_files/paper/2025/hash/83fe5a77502e3d4cfab5960aed0ee6c3-Abstract-Conference.html.

158. Liu, Q.E., Marjieh, R., Zhu, J.-Q., Goldberg, A.E., and Griffiths, T.L. (2026). Parallelograms strike back: LLMs generate better analogies than people. Preprint at arXiv. https://doi.org/10.48550/arXiv.2603.19066.

159. Htut, P.M., Phang, J., Bordia, S., and Bowman, S.R. (2019). Do attention heads in BERT track syntactic dependencies?. Preprint at arXiv. https://doi.org/10.48550/arXiv.1911.12246.

160. McGee, T.A., Zhang, Y., and Blank, I.A. (2026). Evidence against syntactic encapsulation in large language models. Cogn. Sci. *50*, e70187. https://doi.org/10.1111/cogs.70187.

161. Goldberg, A.E. (2024). Usage-based constructionist approaches and large language models. Constr. Fram *16*, 220–254. https://doi.org/10.1075/cf.23017.gol.

162. Piantadosi, S.T. (2024). Modern language models refute Chomsky's approach to language. In From Linguistic Fieldwork to Linguistic Theory: A Tribute to Dan Everett, E. Gibson and M. Poliak, eds. (Language Science Press), pp. 353–414.

163. Futrell, R., and Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models. Behav. Brain Sci. *49*, e198. https://doi.org/10.1017/S0140525X2510112X.

164. Jordan, M.I. (1986). Serial order: a parallel distributed processing approach (University of California, San Diego). https://cseweb.ucsd.edu/~gary/PAPER-SUGGESTIONS/Jordan-TR-8604-OCRed.pdf.

165. Elman, J.L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. Mach. Learn. *7*, 195–225. https://doi.org/10.1007/bf00114844.

166. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

167. Smith, L.B., and Thelen, E. (2003). Development as a dynamic system. Trends Cogn. Sci. *7*, 343–348. https://doi.org/10.1016/s1364-6613(03)00156-6.

168. Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M., and Saphra, N. (2024). Sudden drops in the loss: syntax acquisition, phase transitions, and simplicity bias in MLMs. In International Conference on Learning Representations, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, eds., pp. 14479–14510. https://proceedings.iclr.cc/paper_files/paper/2024/hash/3e8df255a03b383f30abe09203770213-Abstract-Conference.html.

169. Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. (2024). Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity/.

170. Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N.L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., et al. (2025). Circuit tracing: revealing computational graphs in language models. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

171. Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N.L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., et al. (2025). On the biology of a large language model. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

172. Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., and Potts, C. (2024). Mission: impossible language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), A. Martins and V. Srikumar, eds. (Association for Computational Linguistics), pp. 14691–14714. https://doi.org/10.18653/v1/2024.acl-long.787.

173. Contreras Kallens, P., Kristensen-McLachlan, R.D., and Christiansen, M.H. (2023). Large language models demonstrate the potential of statistical learning in language. Cogn. Sci. *47*, e13256. https://doi.org/10.1111/cogs.13256.

174. Pinker, S. (1999). Words and Rules: The Ingredients of Language (Basic Books).

175. Yang, C. (2016). The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language (MIT Press). https://doi.org/10.7551/mitpress/9780262035323.001.0001.

176. Poeppel, D. (2012). The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language.

Cogn. Neuropsychol. *29*, 34–55. https://doi.org/10.1080/02643294.2012.710600.

177. Vaidya, A.R., Jain, S., and Huth, A. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. In Proceedings of the 39th International Conference on Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds. (PMLR), pp. 21927–21944. https://proceedings.mlr.press/v162/vaidya22a.html.

178. Goldstein, A., Wang, H., Niekerken, L., Schain, M., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S.A., Gazula, H., Singh, A., et al. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. Nat. Hum. Behav. *9*, 1041–1055. https://doi.org/10.1038/s41562-025-02105-9.

179. Orhan, P., Boubenec, Y., and King, J.-R. (2025). The detection of algebraic auditory structures emerges with self-supervised learning. PLOS Comput. Biol. *21*, e1013271. https://doi.org/10.1371/journal.pcbi.1013271.

180. Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (MIT Press).

181. Griffiths, T.L., Lake, B.M., McCoy, R.T., Pavlick, E., and Webb, T.W. (2026). Whither symbols in the era of advanced neural networks? Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2026.02.003.

182. Pillow, J.W. (2024). Cross Talk opposing view: Marr's three levels of analysis are not useful as a framework for neuroscience. J. Physiol. *602*, 1915–1917. https://doi.org/10.1113/JP279550.

183. Broca, P. (1861). Remarques sur le Siège de la Faculté du Langage Articulé, Suivies d'une Observation d'Aphémie (Perte de la Parole). Bulletin et Memoires de la Societe Anatomique de Paris 6, 330–357.

184. Wernicke, C. (1969). The symptom complex of aphasia. In Boston Stud. Philos. Sci. (D. Reidel publishing company), pp. 34–97. https://doi.org/10.1007/978-94-010-3378-7_2.

185. Geschwind, N. (1970). The organization of language and the brain. Science *170*, 940–944. https://doi.org/10.1126/science.170.3961.940.

186. Bogen, J.E., and Bogen, G.M. (1976). Wernicke's region—where is it? Ann. N. Y. Acad. Sci. *280*, 834–843. https://doi.org/10.1111/j.1749-6632.1976.tb25546.x.

187. Amunts, K., Lenzen, M., Friederici, A.D., Schleicher, A., Morosan, P., Palomero-Gallagher, N., and Zilles, K. (2010). Broca's region: novel organizational principles and multiple receptor mapping. PLOS Biol. *8*, e1000489. https://doi.org/10.1371/journal.pbio.1000489.

188. Tremblay, P., and Dick, A.S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. Brain Lang. *162*, 60–71. https://doi.org/10.1016/j.bandl.2016.08.004.

189. Fedorenko, E., and Blank, I.A. (2020). Broca's area is not a natural kind. Trends Cogn. Sci. *24*, 270–284. https://doi.org/10.1016/j.tics.2020.01.001.

190. Friederici, A.D., Meyer, M., and von Cramon, D.Y. (2000). Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. Brain Lang. *74*, 289–300. https://doi.org/10.1006/brln.2000.2313.

191. Vandenberghe, R., Nobre, A.C., and Price, C.J. (2002). The response of left temporal cortex to sentences. J. Cogn. Neurosci. *14*, 550–560. https://doi.org/10.1162/08989290260045800.

192. Santi, A., and Grodzinsky, Y. (2010). fMRI adaptation dissociates syntactic complexity dimensions. Neuroimage *51*, 1285–1293. https://doi.org/10.1016/j.neuroimage.2010.03.034.

193. Grodzinsky, Y., and Friederici, A.D. (2006). Neuroimaging of syntax and syntactic processing. Curr. Opin. Neurobiol. *16*, 240–246. https://doi.org/10.1016/j.conb.2006.03.007.

194. Zaccarella, E., and Friederici, A.D. (2015). Merge in the human brain: a sub-region based functional investigation in the left pars opercularis. Front. Psychol. *6*, 1818. https://doi.org/10.3389/fpsyg.2015.01818.

195. Fedorenko, E., Blank, I.A., Siegelman, M., and Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. Cognition *203*, 104348. https://doi.org/10.1016/j.cognition.2020.104348.

196. Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., and Fedorenko, E. (2024). Distributed sensitivity to syntax and semantics throughout the language network. J. Cogn. Neurosci. *36*, 1427–1471. https://doi.org/10.1162/jocn_a_02164.

197. Rauschecker, J.P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc. Natl. Acad. Sci. USA *97*, 11800–11806. https://doi.org/10.1073/pnas.97.22.11800.

198. Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition *92*, 67–99. https://doi.org/10.1016/j.cognition.2003.10.011.

199. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402. https://doi.org/10.1038/nrn2113.

200. Saur, D., Kreher, B.W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., et al. (2008). Ventral and dorsal pathways for language. Proc. Natl. Acad. Sci. USA *105*, 18035–18040. https://doi.org/10.1073/pnas.0805234105.

201. Bookheimer, S. (2002). Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. Annu. Rev. Neurosci. *25*, 151–188. https://doi.org/10.1146/annurev.neuro.25.112701.142946.

202. Vigneau, M., Beaucousin, V., Hervé, P.Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., and Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage *30*, 1414–1432. https://doi.org/10.1016/j.neuroimage.2005.11.002.

203. Friederici, A.D. (2011). The brain basis of language processing: from structure to function. Physiol. Rev. *91*, 1357–1392. https://doi.org/10.1152/physrev.00006.2011.

204. Price, C.J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage *62*, 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062.

205. Hagoort, P., and Indefrey, P. (2014). The neurobiology of language beyond single words. Annu. Rev. Neurosci. *37*, 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847.

206. Chomsky, N. (1979). Language and Responsibility: Based on Conversations with Mitsou Ronat (Pantheon Books).

207. Bruner, J.S. (1985). Actual Minds, Possible Worlds (Harvard University Press).

208. Willems, R.M., Nastase, S.A., and Milivojevic, B. (2020). Narratives for neuroscience. Trends Neurosci. *43*, 271–273. https://doi.org/10.1016/j.tins.2020.03.003.

209. Willems, R.M. (2015). Cognitive Neuroscience of Natural Language Use (Cambridge University Press).

210. Brennan, J. (2016). Naturalistic sentence comprehension in the brain. Lang. Linguist. Compass *10*, 299–313. https://doi.org/10.1111/lnc3.12198.

211. Hasson, U., Egidi, G., Marelli, M., and Willems, R.M. (2018). Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. Cognition *180*, 135–157. https://doi.org/10.1016/j.cognition.2018.06.018.

212. Hamilton, L.S., and Huth, A.G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. Lang. Cogn. Neurosci. *35*, 573–582. https://doi.org/10.1080/23273798.2018.1499946.

213. Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447.

214. Yamins, D.L.K., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365. https://doi.org/10.1038/nn.4244.

215. Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. Am. Biol. Teach. *35*, 125–129. https://doi.org/10.2307/4444260.

216. Kanwisher, N., Khosla, M., and Dobs, K. (2023). Using artificial neural networks to ask "why" questions of minds and brains. Trends Neurosci. *46*, 240–254. https://doi.org/10.1016/j.tins.2022.12.008.

217. Abdou, M. (2022). Connecting neural response measurements and computational models of language: a non-comprehensive guide. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.05300.

218. Arana, S., Pesnot Lerousseau, J., and Hagoort, P. (2024). Deep learning models to study sentence comprehension in the human brain. Lang. Cogn. Neurosci. *39*, 972–990. https://doi.org/10.1080/23273798.2023.2198245.

219. Jain, S., Vo, V.A., Wehbe, L., and Huth, A.G. (2024). Computational language modeling and the promise of in silico experimentation. Neurobiol. Lang. *5*, 80–106. https://doi.org/10.1162/nol_a_00101.

220. Lopopolo, A., Fedorenko, E., Levy, R., and Rabovsky, M. (2024). Cognitive computational neuroscience of language: using computational models to investigate language processing in the brain. Neurobiol. Lang. *5*, 1–6. https://doi.org/10.1162/nol_e_00131.

221. Tuckute, G., Kanwisher, N., and Fedorenko, E. (2024). Language in brains, minds, and machines. Annu. Rev. Neurosci. *47*, 277–301. https://doi.org/10.1146/annurev-neuro-120623-101142.

222. Wu, M.C.-K., David, S.V., and Gallant, J.L. (2006). Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. *29*, 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024.

223. Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage *56*, 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073.

224. Dupré la Tour, T., Visconti di Oleggio Castello, M., and Gallant, J.L. (2025). The Voxelwise Encoding Model framework: a tutorial introduction to fitting encoding models to fMRI data. Imaging Neurosci. *3*, imag_a_00575. https://doi.org/10.1162/imag_a_00575.

225. Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat. Neurosci. *22*, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2.

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

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

228. Paradiso, M.A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. Biol. Cybern. *58*, 35–49. https://doi.org/10.1007/BF00363954.

229. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K.D. (2019). High-dimensional geometry of population responses in visual cortex. Nature *571*, 361–365. https://doi.org/10.1038/s41586-019-1346-5.

230. DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341. https://doi.org/10.1016/j.tics.2007.06.010.

231. DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434. https://doi.org/10.1016/j.neuron.2012.01.010.

232. Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412. https://doi.org/10.1016/j.tics.2013.06.007.

233. Rolls, E.T., and Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J. Neurophysiol. *73*, 713–726. https://doi.org/10.1152/jn.1995.73.2.713.

234. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science *293*, 2425–2430. https://doi.org/10.1126/science.1063736.

235. Pasupathy, A., and Connor, C.E. (2002). Population coding of shape in area V4. Nat. Neurosci. *5*, 1332–1338. https://doi.org/10.1038/nn972.

236. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863–866. https://doi.org/10.1126/science.1117593.

237. Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043.

238. Freiwald, W.A., and Tsao, D.Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science *330*, 845–851. https://doi.org/10.1126/science.1194908.

239. Kaufman, M.T., Churchland, M.M., Ryu, S.I., and Shenoy, K.V. (2014). Cortical activity in the null space: permitting preparation without movement. Nat. Neurosci. *17*, 440–448. https://doi.org/10.1038/nn.3643.

240. Churchland, M.M., and Shenoy, K.V. (2024). Preparatory activity and the expansive null-space. Nat. Rev. Neurosci. *25*, 213–236. https://doi.org/10.1038/s41583-024-00796-z.

241. Georgopoulos, A.P., Schwartz, A.B., and Kettner, R.E. (1986). Neuronal population coding of movement direction. Science *233*, 1416–1419. https://doi.org/10.1126/science.3749885.

242. Russo, A.A., Bittner, S.R., Perkins, S.M., Seely, J.S., London, B.M., Lara, A.H., Miri, A., Marshall, N.J., Kohn, A., Jessell, T.M., et al. (2018). Motor cortex embeds muscle-like commands in an untangled population response. Neuron *97*, 953–966.e8. https://doi.org/10.1016/j.neuron.2018.01.004.

243. Sorscher, B., Ganguli, S., and Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. Proc. Natl. Acad. Sci. USA *119*, e2200800119. https://doi.org/10.1073/pnas.2200800119.

244. Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Nastase, S.A., Zada, Z., Ham, E., Feder, A., et al. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. Nat. Commun. *15*, 2768. https://doi.org/10.1038/s41467-024-46631-y.

245. Desbordes, T., Lakretz, Y., Chanoine, V., Oquab, M., Badier, J.-M., Trébuchon, A., Carron, R., Bénar, C.-G., Dehaene, S., and King, J.-R. (2023). Dimensionality and ramping: signatures of sentence integration in the dynamics of brains and deep language models. J. Neurosci. *43*, 5350–5364. https://doi.org/10.1523/JNEUROSCI.1163-22.2023.

246. Jamali, M., Grannan, B., Cai, J., Khanna, A.R., Muñoz, W., Caprara, I., Paulk, A.C., Cash, S.S., Fedorenko, E., and Williams, Z.M. (2024). Semantic encoding during language comprehension at single-cell resolution. Nature *631*, 610–616. https://doi.org/10.1038/s41586-024-07643-2.

247. Cai, J., Kfir, Y., Jamali, M., Huang, H., Kim, Y.J., Cash, S.S., and Williams, Z.M. (2026). Mapping the neuronal building blocks of human language with language models. Nature. https://doi.org/10.1038/s41586-026-10691-5.

248. Quiroga, R.Q. (2012). Concept cells: the building blocks of declarative memory functions. Nat. Rev. Neurosci. *13*, 587–597. https://doi.org/10.1038/nrn3251.

249. Plaut, D.C., and McClelland, J.L. (2010). Locating object knowledge in the brain: comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. Psychol. Rev. *117*, 284–288. https://doi.org/10.1037/a0017101.

250. Yan, X., Li, J.-A., Franch, M., Zhu, H., Cowan, R.L., Belanger, J., Chavez, A.G.L., Chericoni, A., Ismail, T., Katlowitz, K., et al. (2026).

Polysemanticity in human hippocampal neurons. Preprint at bioRxiv, 2026.05.02.722435. https://doi.org/10.64898/2026.05.02.722435.

251. Mahon, B.Z., and Caramazza, A. (2009). Concepts and categories: a cognitive neuropsychological perspective. Annu. Rev. Psychol. *60*, 27–51. https://doi.org/10.1146/annurev.psych.60.110707.163532.

252. Plaut, D.C. (1995). Double dissociation without modularity: evidence from connectionist neuropsychology. J. Clin. Exp. Neuropsychol. *17*, 291–321. https://doi.org/10.1080/01688639508405124.

253. Li, Z., You, C., Bhojanapalli, S., Li, D., Rawat, A.S., Reddi, S.J., Ye, K., Chern, F., Yu, F., Guo, R., et al. (2022). The lazy neuron phenomenon: on emergence of activation sparsity in transformers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.06313.

254. Margalit, E., Lee, H., Finzi, D., DiCarlo, J.J., Grill-Spector, K., and Yamins, D.L.K. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. Neuron *112*, 2435–2451.e7. https://doi.org/10.1016/j.neuron.2024.04.018.

255. Rathi, N., Mehrer, J., AlKhamissi, B., Binhuraib, T., Blauch, N., and Schrimpf, M. (2025). TopoLM: brain-like spatio-functional organization in a topographic language model. In International Conference on Learning Representations, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, eds., pp. 32087–32112. https://proceedings.iclr.cc/paper_files/paper/2025/hash/4f7f521c08b5e0cd9931e74fa353b59b-Abstract-Conference.html.

256. AlKhamissi, B., Mehrer, J., Marinov, L., Abdelaal, A., Gokce, A., and Schrimpf, M. (2026). Discovering functionally selective brain regions with a deep topographic multimodal model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2606.09770.

257. Friederici, A.D., Chomsky, N., Berwick, R.C., Moro, A., and Bolhuis, J.J. (2017). Language, mind and brain. Nat. Hum. Behav. *1*, 713–722. https://doi.org/10.1038/s41562-017-0184-4.

258. Allen, K., Pereira, F., Botvinick, M., and Goldberg, A.E. (2012). Distinguishing grammatical constructions with fMRI pattern analysis. Brain Lang. *123*, 174–182. https://doi.org/10.1016/j.bandl.2012.08.005.

259. Fedorenko, E., Nieto-Castañon, A., and Kanwisher, N. (2012). Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. Neuropsychologia *50*, 499–513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014.

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

261. Bozic, M., Fonteneau, E., Su, L., and Marslen-Wilson, W.D. (2015). Grammatical analysis as a distributed neurobiological function. Hum. Brain Mapp. *36*, 1190–1201. https://doi.org/10.1002/hbm.22696.

262. Rodd, J.M., Vitello, S., Woollams, A.M., and Adank, P. (2015). Localising semantic and syntactic processing in spoken and written language comprehension: an Activation Likelihood Estimation meta-analysis. Brain Lang. *141*, 89–102. https://doi.org/10.1016/j.bandl.2014.11.012.

263. Blank, I., Balewski, Z., Mahowald, K., and Fedorenko, E. (2016). Syntactic processing is distributed across the language system. Neuroimage *127*, 307–323. https://doi.org/10.1016/j.neuroimage.2015.11.069.

264. Fedorenko, E., Scott, T.L., Brunner, P., Coon, W.G., Pritchett, B., Schalk, G., and Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. Proc. Natl. Acad. Sci. USA *113*, E6256–E6262. https://doi.org/10.1073/pnas.1612132113.

265. Nelson, M.J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S.S., Naccache, L., Hale, J.T., Pallier, C., et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. Proc. Natl. Acad. Sci. USA *114*, E3669–E3678. https://doi.org/10.1073/pnas.1701590114.

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

267. Toneva, M., Mitchell, T.M., and Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. Nat. Comput. Sci. *2*, 745–757. https://doi.org/10.1038/s43588-022-00354-6.

268. Lyu, B., Marslen-Wilson, W.D., Fang, Y., and Tyler, L.K. (2024). Finding structure during incremental speech comprehension. eLife *12*, RP89311. https://doi.org/10.7554/eLife.89311.

269. Tu, Z., Dai, L., Zhang, B., Chen, S., Yang, Y., Meng, D., Gong, Y., and Sun, J. (2025). Revealing human brain syntactic processing: insights from voxel-wise models and network representation. Brain Lang. *265*, 105569. https://doi.org/10.1016/j.bandl.2025.105569.

270. Caucheteux, C., Gramfort, A., and King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. In Proceedings of the 38th International Conference on Machine Learning Research, M. Meila and T. Zhang, eds. (PMLR), pp. 1336–1348. https://proceedings.mlr.press/v139/caucheteux21a.html.

271. Kumar, S., Sumers, T.R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K.A., Griffiths, T.L., Hawkins, R.D., and Nastase, S.A. (2024). Shared functional specialization in transformer-based language models and the human brain. Nat. Commun. *15*, 5523. https://doi.org/10.1038/s41467-024-49173-5.

272. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned. Preprint at arXiv. https://doi.org/10.48550/arXiv.1905.09418.

273. Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. *111*, 1872–1891. https://doi.org/10.1121/1.1458026.

274. Chi, T., Ru, P., and Shamma, S.A. (2005). Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. *118*, 887–906. https://doi.org/10.1121/1.1945807.

275. Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010. https://doi.org/10.1126/science.1245994.

276. Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLOS Comput. Biol. *10*, e1003412. https://doi.org/10.1371/journal.pcbi.1003412.

277. de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The hierarchical cortical organization of human speech processing. J. Neurosci. *37*, 6539–6557. https://doi.org/10.1523/jneurosci.3267-16.2017.

278. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630–644.e16. https://doi.org/10.1016/j.neuron.2018.03.044.

279. Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., and King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. In Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds. (Curran Associates, Inc.), pp. 33428–33443. https://doi.org/10.52202/068431-2422.

280. Li, Y., Anumanchipalli, G.K., Mohamed, A., Chen, P., Carney, L.H., Lu, J., Wu, J., and Chang, E.F. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. Nat. Neurosci. *26*, 2213–2225. https://doi.org/10.1038/s41593-023-01468-4.

281. Tuckute, G., Feather, J., Boebinger, D., and McDermott, J.H. (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*, e3002366. https://doi.org/10.1371/journal.pbio.3002366.

282. Radford, A., Kim, A.W., Xu, T., Brockman, G., Mcleavy, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine

Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds. (PMLR), pp. 28492–28518.

283. Leonard, M.K., Gwilliams, L., Sellers, K.K., Chung, J.E., Xu, D., Mischler, G., Mesgarani, N., Welkenhuysen, M., Dutta, B., and Chang, E.F. (2024). Large-scale single-neuron speech sound encoding across the depth of human cortex. Nature *626*, 593–602. https://doi.org/10.1038/s41586-023-06839-2.

284. Khanna, A.R., Muñoz, W., Kim, Y.J., Kfir, Y., Paulk, A.C., Jamali, M., Cai, J., Mustroph, M.L., Caprara, I., Hardstone, R., et al. (2024). Single-neuronal elements of speech production in humans. Nature *626*, 603–610. https://doi.org/10.1038/s41586-023-06982-w.

285. Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996). Statistical learning by 8-month-old infants. Science *274*, 1926–1928. https://doi.org/10.1126/science.274.5294.1926.

286. Rovee-Collier, C., and Cuevas, K. (2008). The development of infant memory. In The Development of Memory in Infancy and Childhood, M.L. Courage and N. Cowan, eds. (Psychology Press), pp. 23–54. https://doi.org/10.4324/9780203934654-6.

287. Kutas, M., DeLong, K.A., and Smith, N.J. (2011). A look around at what lies ahead: prediction and predictability in language processing. In Predictions in the Brain, M. Bar, ed. (Oxford University Press), pp. 190–207. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065.

288. Kuperberg, G.R., and Jaeger, T.F. (2016). What do we mean by prediction in language comprehension? Lang. Cogn. Neurosci. *31*, 32–59. https://doi.org/10.1080/23273798.2015.1102299.

289. Ferreira, F., and Chantavarin, S. (2018). Integration and prediction in language processing: a synthesis of old and new. Curr. Dir. Psychol. Sci. *27*, 443–448. https://doi.org/10.1177/0963721418794491.

290. Kutas, M., and Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. Science *207*, 203–205. https://doi.org/10.1126/science.7350657.

291. Kutas, M., and Federmeier, K.D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.

292. Jain, S., and Huth, A. (2018). Incorporating context into language encoding models for fMRI. In Adv. Neural Inf. Process. Syst., *31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds. (Curran Associates, Inc.), pp. 6628–6637. https://proceedings.neurips.cc/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html.

293. Toneva, M., and 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, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds. (Curran Associates, Inc.) https://proceedings.neurips.cc/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html.

294. Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B., and Fedorenko, E. (2021). The neural architecture of language: integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. USA *118*, e2105646118. https://doi.org/10.1073/pnas.2105646118.

295. Caucheteux, C., and King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. Commun. Biol. *5*, 134. https://doi.org/10.1038/s42003-022-03036-1.

296. Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. *25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4.

297. Mischler, G., Li, Y.A., Bickel, S., Mehta, A.D., and Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. Nat. Mach. Intell. *6*, 1467–1477. https://doi.org/10.1038/s42256-024-00925-4.

298. Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., et al. (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. Neuron *112*, 3211–3222.e5. https://doi.org/10.1016/j.neuron.2024.06.025.

299. Evanson, L., Bulteau, C., Chipaux, M., Dorfmüller, G., Ferrand-Sorbets, S., Raffo, E., Rosenberg, S., Bourdillon, P., and King, J.-R. (2025). Emergence of language in the developing brain. Preprint at arXiv. https://doi.org/10.48550/arXiv.2512.05718.

300. Michaelov, J.A., Bardolph, M.D., Van Petten, C.K., Bergen, B.K., and Coulson, S. (2024). Strong prediction: language model surprisal explains multiple N400 effects. Neurobiol. Lang. *5*, 107–135. https://doi.org/10.1162/nol_a_00105.

301. Hadidi, N., Feghhi, E., Song, B.H., Blank, I.A., and Kao, J.C. (2026). Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. Nat. Commun. *17*, 5769. https://doi.org/10.1038/s41467-026-72253-7.

302. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and de Lange, F.P. (2022). A hierarchy of linguistic predictions during natural language comprehension. Proc. Natl. Acad. Sci. USA *119*, e2201968119. https://doi.org/10.1073/pnas.2201968119.

303. Caucheteux, C., Gramfort, A., and King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat. Hum. Behav. *7*, 430–441. https://doi.org/10.1038/s41562-022-01516-2.

304. Azizpour, S., Westner, B.U., Szewczyk, J., Güçlü, U., and Geerligs, L. (2026). Reassessing prediction in the brain: pre-onset neural encoding during natural listening does not reflect pre-activation. eLife. https://doi.org/10.7554/elife.110085.1.

305. Schönmann, I., Szewczyk, J., de Lange, F.P., and Heilbron, M. (2026). Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding in naturalistic listening designs. eLife *14*, RP106543. https://doi.org/10.7554/eLife.106543.

306. Rabagliati, H., Gambi, C., and Pickering, M.J. (2016). Learning to predict or predicting to learn? Lang. Cogn. Neurosci. *31*, 94–105. https://doi.org/10.1080/23273798.2015.1077979.

307. Raviv, H., Tsyhanov, A., Gousios, K., Altenhof, A., Wang, H., Chen, B., Raviv, O., Rosenwein, T., Lew-Williams, C., Hasenfratz, L., et al. (2026). The First 1000 Days: an agent-based model of early language acquisition. Preprint at bioRxiv. https://doi.org/10.64898/2026.03.28.715023.

308. Evanson, L., Lakretz, Y., and King, J.R. (2023). Language acquisition: do children and language models follow similar learning stages? In Findings of the Association for Computational Linguistics: ACL (Association for Computational Linguistics), pp. 12205–12218. https://doi.org/10.18653/v1/2023.findings-acl.773.

309. Chang, F., Dell, G.S., and Bock, K. (2006). Becoming syntactic. Psychol. Rev. *113*, 234–272. https://doi.org/10.1037/0033-295X.113.2.234.

310. Borovsky, A., Elman, J.L., and Fernald, A. (2012). Knowing a lot for one's age: vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. J. Exp. Child Psychol. *112*, 417–436. https://doi.org/10.1016/j.jecp.2012.01.005.

311. Ramscar, M., Dye, M., and McCauley, S.M. (2013). Error and expectation in language learning: the curious absence of *mouses* in adult speech. Language *89*, 760–793. https://doi.org/10.1353/lan.2013.0068.

312. Zettersten, M. (2019). Learning by predicting: how predictive processing informs language development. In Patterns in Language and Linguistics, B. Busse and R. Moehlig-Falke, eds. (De Gruyter), pp. 255–288. https://doi.org/10.1515/9783110596656-010.

313. Reuter, T., Dalawella, K., and Lew-Williams, C. (2021). Adults and children predict in complex and variable referential contexts. Lang. Cogn. Neurosci. *36*, 474–490. https://doi.org/10.1080/23273798.2020.1839665.

314. Kray, J., Sommerfeld, L., Borovsky, A., and Häuser, K. (2024). The role of prediction error in the development of language learning and memory. Child Dev. Perspect. *18*, 190–203. https://doi.org/10.1111/cdep.12515.

315. Reddy, S., Chen, D., and Manning, C.D. (2019). CoQA: a conversational question answering challenge. Trans. Assoc. Comput. Linguist. *7*, 249–266. https://doi.org/10.1162/tacl_a_00266.

316. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E.M., Boureau, Y.-L., et al. (2021). Recipes for building an open-domain chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, P. Merlo, J. Tiedemann, and R. Tsarfaty, eds. (Association for Computational Linguistics), pp. 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24.

317. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). LaMDA: language models for dialog applications. Preprint at arXiv. https://doi.org/10.48550/arXiv.2201.08239.

318. Cai, J., Hadjinicolaou, A.E., Paulk, A.C., Soper, D.J., Xia, T., Wang, A.F., Rolston, J.D., Richardson, R.M., Williams, Z.M., and Cash, S.S. (2025). Natural language processing models reveal neural dynamics of human conversation. Nat. Commun. *16*, 3376. https://doi.org/10.1038/s41467-025-58620-w.

319. Yamashita, M., Kubo, R., and Nishimoto, S. (2025). Conversational content is organized across multiple timescales in the brain. Nat. Hum. Behav. *9*, 2066–2078. https://doi.org/10.1038/s41562-025-02231-4.

320. Zada, Z., Nastase, S.A., Speer, S., Mwilambwe-Tshilobo, L., Tsoi, L., Burns, S.M., Falk, E., Hasson, U., and Tamir, D.I. (2026). Linguistic coupling between neural systems for speech production and comprehension during real-time dyadic conversations. Neuron *114*, 774–787.e5. https://doi.org/10.1016/j.neuron.2025.11.004.

321. Spotorno, N., Koun, E., Prado, J., Van Der Henst, J.-B., and Noveck, I.A. (2012). Neural evidence that utterance-processing entails mentalizing: the case of irony. Neuroimage *63*, 25–39. https://doi.org/10.1016/j.neuroimage.2012.06.046.

322. Bašnáková, J., Weber, K., Petersson, K.M., van Berkum, J., and Hagoort, P. (2014). Beyond the language given: the neural correlates of inferring speaker meaning. Cereb. Cortex *24*, 2572–2578. https://doi.org/10.1093/cercor/bht112.

323. Vanlangendonck, F., Willems, R.M., and Hagoort, P. (2018). Taking common ground into account: specifying the role of the mentalizing network in communicative language production. PLOS One *13*, e0202943. https://doi.org/10.1371/journal.pone.0202943.

324. Paunov, A.M., Blank, I.A., and Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. J. Neurophysiol. *121*, 1244–1265. https://doi.org/10.1152/jn.00619.2018.

325. Haxby, J.V., Guntupalli, J.S., Nastase, S.A., and Feilong, M. (2020). Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. eLife *9*, e56601. https://doi.org/10.7554/eLife.56601.

326. Hasson, U., Ghazanfar, A.A., Galantucci, B., Garrod, S., and Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends Cogn. Sci. *16*, 114–121. https://doi.org/10.1016/j.tics.2011.12.007.

327. Zada, Z., Nastase, S.A., Li, J., and Hasson, U. (2025). Brains and language models converge on a shared conceptual space across different languages. Preprint at arXiv, arXiv:2506.20489v1. https://doi.org/10.48550/arXiv.2506.20489.

328. Perez-Nieves, N., Leung, V.C.H., Dragotti, P.L., and Goodman, D.F.M. (2021). Neural heterogeneity promotes robust learning. Nat. Commun. *12*, 5791. https://doi.org/10.1038/s41467-021-26022-3.

329. Lynn, C.W. (2026). Simple input–output dependencies explain neuronal activity. Nat. Phys. *22*, 1152–1159. https://doi.org/10.1038/s41567-026-03306-3.

330. Stanojevic, A., Woźniak, S., Bellec, G., Cherubini, G., Pantazi, A., and Gerstner, W. (2024). High-performance deep spiking neural networks with 0.3 spikes per neuron. Nat. Commun. *15*, 6793. https://doi.org/10.1038/s41467-024-51110-5.

331. Whittington, J.C.R., and Bogacz, R. (2019). Theories of error back-propagation in the brain. Trends Cogn. Sci. *23*, 235–250. https://doi.org/10.1016/j.tics.2018.12.005.

332. Kozachkov, L., Kastanenka, K.V., and Krotov, D. (2023). Building transformers from neurons and astrocytes. Proc. Natl. Acad. Sci. USA *120*, e2219150120. https://doi.org/10.1073/pnas.2219150120.

333. Warstadt, A., and Bowman, S.R. (2022). What artificial neural networks can tell us about human language acquisition. In Algebraic Structures in Natural Language (CRC Press), pp. 17–60. https://doi.org/10.1201/9781003205388-2.

334. Raviv, H., Hasenfratz, L., Gousios, K., Faryna, M., Beaty, R., Johnson, D., Chen, B., Altenhof, A., Ryan, B., Greenberg, C.A., et al. (2026). The First 1,000 days (1kD) project - collecting and analyzing an ultra-dense naturalistic dataset of Human baby development. Preprint at bioRxiv. https://doi.org/10.64898/2026.03.19.712982.

335. Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., et al. (2023). Findings of the BabyLM Challenge: sample-efficient pretraining on developmentally plausible corpora. In Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning, A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, and T. Linzen, et al., eds. (Association for Computational Linguistics), pp. 1–6. https://doi.org/10.18653/v1/2023.conll-babylm.1.

336. Hu, M.Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., and Wilcox, E.G. (2024). Findings of the second BabyLM Challenge: sample-efficient pretraining on developmentally plausible corpora. In The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, M.Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt, and E.G. Wilcox, eds. (Association for Computational Linguistics), pp. 1–21. https://aclanthology.org/2024.conll-babylm.1/.

337. Charpentier, L., Choshen, L., Cotterell, R., Gul, M.O., Hu, M.Y., Liu, J., Jumelet, J., Linzen, T., Mueller, A., Ross, C., et al. (2025). Findings of the third BabyLM Challenge: accelerating language modeling research with cognitively plausible data. In Proceedings of the First BabyLM Workshop, Charpentier., L. Choshen, R. Cotterell, M.O. Gul, M.Y. Hu, J. Liu, J. Jumelet, T. Linzen, A. Mueller, and C. Ross, et al., eds. (Association for Computational Linguistics), pp. 399–420. https://doi.org/10.18653/v1/2025.babylm-main.28.

338. Hosseini, E.A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., and Fedorenko, E. (2024). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. Neurobiol. Lang. (Camb) *5*, 43–63. https://doi.org/10.1162/nol_a_00137.

339. Christiansen, M.H., and Chater, N. (2016). The now-or-never bottleneck: a fundamental constraint on language. Behav. Brain Sci. *39*, e62. https://doi.org/10.1017/S0140525X1500031X.

340. Bastiaansen, M., and Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. Prog. Brain Res. *159*, 179–196. https://doi.org/10.1016/S0079-6123(06)59012-0.

341. Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S.J., and Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron *56*, 1127–1134. https://doi.org/10.1016/j.neuron.2007.09.038.

342. Martin, A.E., and Doumas, L.A.A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. PLOS Biol. *15*, e2000663. https://doi.org/10.1371/journal.pbio.2000663.

343. Meyer, L., Sun, Y., and Martin, A.E. (2020). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. Lang. Cogn. Neurosci. *35*, 1089–1099. https://doi.org/10.1080/23273798.2019.1693050.

344. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. Long Papers. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, *1*, M. Walker, H. Ji, and A. Stent, eds. (Association for Computational Linguistics), pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202.

345. Dentella, V., Günther, F., and Leivada, E. (2023). Systematic testing of three Language Models reveals low language accuracy, absence of

response stability, and a yes-response bias. Proc. Natl. Acad. Sci. USA *120*, e2309583120. https://doi.org/10.1073/pnas.2309583120.

346. Dentella, V., Günther, F., Murphy, E., Marcus, G., and Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. Sci. Rep. *14*, 28083. https://doi.org/10.1038/s41598-024-79531-8.

347. Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., and Levy, R. (2024). Language models align with human judgments on key grammatical constructions. Proc. Natl. Acad. Sci. USA *121*, e2400917121. https://doi.org/10.1073/pnas.2400917121.

348. Juzek, T.S. (2024). The Syntactic Acceptability Dataset (Preview): a resource for machine learning and linguistic analysis of English. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds. (ELRA and ICCL), pp. 16113–16120. Proceedings of the Language Resources and Evaluation Conference. https://doi.org/10.63317/3f2yamzokm9i. https://aclanthology.org/2024.lrec-main.1401/.

349. Rakshit, S., Hu, J., Mahowald, K., and Goldberg, A.E. (2026). A suite of LMs comprehend puzzle statements as well or better than humans. Open Mind (Camb) *10*, 431–440. https://doi.org/10.1162/OPMI.a.344.

350. Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. (2024). Understanding the effects of RLHF on LLM generalisation and diversity. In International Conference on Learning Representations, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, eds., pp. 20620–20653. https://proceedings.iclr.cc/paper_files/paper/2024/hash/5a68d05006d5b05dd9463dd9c0219db0-Abstract-Conference.html.

351. McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. *102*, 419–457. https://doi.org/10.1037/0033-295X.102.3.419.

352. Norman, K.A., and O'Reilly, R.C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychol. Rev. *110*, 611–646. https://doi.org/10.1037/0033-295X.110.4.611.

353. Giallanza, T., Campbell, D., Cohen, J.D., and Rogers, T.T. (2024). An integrated model of semantics and control. Psychol. Rev. *132*, 1128–1177. https://doi.org/10.1037/rev0000485.

354. Webb, T.W., Frankland, S.M., Altabaa, A., Segert, S., Krishnamurthy, K., Campbell, D., Russin, J., Giallanza, T., O'Reilly, R., Lafferty, J., et al. (2024). The relational bottleneck as an inductive bias for efficient abstraction. Trends Cogn. Sci. *28*, 829–843. https://doi.org/10.1016/j.tics.2024.04.001.

355. Zador, A.M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. Nat. Commun. *10*, 3770. https://doi.org/10.1038/s41467-019-11786-6.

356. Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., and Nanda, N. (2024). Gemma Scope: open sparse autoencoders everywhere all at once on Gemma 2. In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen, eds. (Association for Computational Linguistics), pp. 278–300. https://doi.org/10.18653/v1/2024.blackboxnlp-1.19.

357. Panichello, M.F., and Buschman, T.J. (2021). Shared mechanisms underlie the control of working memory and attention. Nature *592*, 601–605. https://doi.org/10.1038/s41586-021-03390-w.

358. Doerig, A., Sommers, R.P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G.W., Kording, K.P., Konkle, T., van Gerven, M.A.J., Kriegeskorte, N., et al. (2023). The neuroconnectionist research programme. Nat. Rev. Neurosci. *24*, 431–450. https://doi.org/10.1038/s41583-023-00705-w.

359. McCloskey, M. (1991). Networks and theories: the place of connectionism in cognitive science. Psychol. Sci. *2*, 387–395. https://doi.org/10.1111/j.1467-9280.1991.tb00173.x.

360. Wood, J.N., Pandey, L., and Wood, S.M.W. (2024). Digital twin studies for reverse engineering the origins of visual intelligence. Annu. Rev. Vis. Sci. *10*, 145–170. https://doi.org/10.1146/annurev-vision-101322-103628.