Trends in Cognitive Sciences

CellPress

Feature Review

Dissociating language and thought in large language models

Kyle Mahowald,[1,5,*] Anna A. Ivanova,[2,5,*] Idan A. Blank,[3,*] Nancy Kanwisher,[4,*] Joshua B. Tenenbaum,[4,*] and Evelina Fedorenko[4,*]

Large language models (LLMs) have come closest among all models to date to mastering human language, yet opinions about their linguistic and cognitive capabilities remain split. Here, we evaluate LLMs using a distinction between formal linguistic competence (knowledge of linguistic rules and patterns) and functional linguistic competence (understanding and using language in the world). We ground this distinction in human neuroscience, which has shown that formal and functional competence rely on different neural mechanisms. Although LLMs are surprisingly good at formal competence, their performance on functional competence tasks remains spotty and often requires specialized fine-tuning and/or coupling with external modules. We posit that models that use language in human-like ways would need to master both of these competence types, which, in turn, could require the emergence of separate mechanisms specialized for formal versus functional linguistic competence.

Highlights

Formal linguistic competence (getting the form of language right) and functional linguistic competence (using language to accomplish goals in the world) are distinct cognitive skills.

The human brain contains a network of areas that selectively support language processing (formal linguistic competence), but not other domains like logical or social reasoning (functional linguistic competence).

In the late 2010s, large language models trained on word prediction tasks began achieving unprecedented success in formal linguistic competence, showing impressive performance on linguistic tasks that likely require hierarchy and abstraction.

Consistent performance on tasks requiring functional linguistic competence is harder to achieve for large language models and often involves augmentations beyond next word prediction.

Evidence from cognitive science and neuroscience can illuminate the capabilities and limitations of large language models and pave the way toward better, human-like models of both language and thought.

The language–thought conflation

When we hear a sentence, we typically assume that it was produced by a rational, thinking agent (another person). The sentences that people generate in day-to-day conversations are based on their world knowledge ('Not all birds can fly.'), their reasoning abilities ('You're 15, you can't go to a bar.'), and their goals ('Would you give me a ride, please?'). Thus, we often use other people's statements as a window into their minds.

In 1950, Alan Turing leveraged this tight relationship between language and thought to propose his famous test [1]. The Turing test uses language as an interface to cognition, allowing a human participant to probe the knowledge and reasoning capacities of two conversation partners to determine which of them is a human and which is a machine. Although the utility of the Turing test has since been questioned, it has undoubtedly shaped the way society today thinks of machine intelligence [2].

The popularity of the Turing test, combined with language–thought coupling in everyday life, has led to several common fallacies related to the language–thought relationship. One fallacy is that an entity (be it a human or a machine) that is good at language must also be good at thinking. If an entity generates coherent stretches of text, it must possess rich knowledge and reasoning capacities. Let's call this the 'good at language -> good at thought' fallacy. This fallacy has come to the forefront due to the recent rise of **large language models (LLMs)** (see Glossary), including OpenAI's GPT models, Anthropic's Claude, and more open alternatives [3] like Meta's LLaMa models and EleutherAI's GPT-J. LLMs today can produce text that is difficult to distinguish from human output, outperform humans at some text comprehension tasks [4,5], and show superhuman performance on next-word prediction [6]. As a result, claims have emerged, both in the popular press and in the academic literature, that LLMs are not only a major advance in language processing, but are also showing 'sparks of artificial general

[1]The University of Texas at Austin, Austin, TX, USA
[2]Georgia Institute of Technology, Atlanta, GA, USA
[3]University of California, Los Angeles, CA, USA
[4]Massachusetts Institute of Technology, Cambridge, MA, USA
[5]These authors contributed equally to this work

*Correspondence:
mahowald@utexas.edu (K. Mahowald),
a.ivanova@gatech.edu (A.A. Ivanova),
iblank@psych.ucla.edu (I.A. Blank),
ngk@mit.edu (N. Kanwisher),
jbt@mit.edu (J.B. Tenenbaum), and
evelina9@mit.edu (E. Fedorenko).

intelligence' [7]. However, when evaluating LLMs' capabilities, it is important to distinguish between their ability to think and their linguistic ability. The 'good at language -> good at thought' fallacy makes it easy to confuse the two, leading people to mistakenly attribute intelligence and intentionality to even the most basic dialog systems (e.g., the chatbot Eliza from the 1960s [8]).

The contrapositive of this fallacy is that a model that is bad at thinking must also be a bad model of language. Let's call this the 'bad at thought -> bad at language' fallacy. LLMs are commonly criticized for their lack of consistent, generalizable world knowledge [9], lack of commonsense reasoning abilities [10], and failure to understand what an utterance is really about [11]. Based on this evidence, some critics suggest that the models' failure to produce linguistic output that fully captures the richness and sophistication of human *thought* means that they are not good models of human *language*.

Both the 'good at language -> good at thought' and the 'bad at thought -> bad at language' fallacies stem from the conflation of language and thought. This conflation is unsurprising: it is still novel, and thus uncanny, to encounter an entity that generates fluent sentences despite lacking a human identity. Thus, our heuristics for understanding what a language model is doing, heuristics that emerged from our language experience with other humans, are broken.

To mitigate the language–thought conflation fallacies, we propose to systematically distinguish between two kinds of linguistic competence: **formal linguistic competence**, the knowledge of rules and statistical regularities of language, and **functional linguistic competence**, the ability to use language in real-world situations. Our motivation for the formal/functional distinction comes from the human brain, where these skills are robustly dissociable. Both formal and functional linguistic competence are essential components of human language use: an effective communicator needs to both generate grammatical, meaningful utterances and strategically use those utterances to achieve diverse, context-dependent goals [12,13].

Armed with this distinction, we evaluate the capabilities of contemporary LLMs and argue that LLMs exhibit a gap between formal and functional competence skills: for modern LLMs, formal competence in English is near human-level, whereas their functional competence remains patchy, with results depending on specific functional competence domains and on tasks within those domains. Moreover, whereas formal linguistic competence in LLMs improves drastically with the amount of training data, functional linguistic competence improvements with scale are less consistent, such that LLM developers have now shifted away from simple scaling-up of the language prediction task to more specialized methods targeting behaviors of interest [e.g., **reinforcement learning from human feedback (RLHF)**] [14] or coupling an LLM with external specialized modules (leading to so-called 'augmented language models' [15]).

Therefore, we posit that the next-word prediction objective allows a model to master formal but not necessarily functional linguistic competence. What is required for mastery of functional competence is harder to pin down, in part because so much of human cognition (commonsense reasoning, scientific knowledge, cultural knowledge) can be conveyed *through language* and thus learned *from language*, even if the underlying cognitive capacities are not themselves inherently linguistic. As a result, language models acquire a variety of non-linguistic capacities. But the ultimate ceiling for functional competence depends on important open questions about the information contained in the linguistic signal and the mechanisms used to leverage that information.

In the rest of the paper, we develop a framework for evaluating the competence of modern language models from a cognitive science perspective. In the first section, we elaborate on the

constructs of formal and functional linguistic competence and motivate this distinction based on evidence from human neuroscience. In the second section, we discuss the successes of LLMs in achieving formal linguistic competence, showing that models trained on word-in-context prediction capture numerous complex linguistic phenomena. In the third section, we consider several domains required for functional linguistic competence, formal reasoning, world knowledge, situation modeling, and social cognition, on which today's LLMs often fail, or at least perform worse than humans. In the fourth section, we discuss the implications of our framework for building and evaluating future models of language and thought before summarizing our key conclusions in the final section.

Formal versus functional linguistic competence

What does linguistic competence entail?

Formal linguistic competence. We define formal linguistic competence as a set of capacities required to produce and comprehend a given language. Broadly, being formally competent means getting the *form* of language right: knowing which strings could be valid words of a language (e.g., *bnick* cannot be a word in English but *blick* can [16]), how to productively combine morphemes to form novel words (e.g., Barack Obama-less-ness but not Barack Obama-ness-less [17]), learning enough about word meanings to know which words can go in which slots in a sentence [18], and knowing how to combine words into valid sentences.

Because of its centrality in the history of linguistics, it is the last of these (forming words into sentences) that we focus on in our discussion of formal competence. Most users of Standard Written English say: 'The dogs in my bedroom are asleep' rather than 'The dogs in my bedroom is asleep', because the verb 'to be' must match the number of the noun that is the subject of the sentence ('the dogs'), even though that verb is closer to an intervening, singular noun ('bedroom'). Linguistic competence also requires exquisite sensitivity to the regularities of idiosyncratic linguistic constructions. For instance, although English speakers know not to use the indefinite article 'a' with plural nouns, making a phrase like 'a days' ill-formed, they also know that it is allowed in a special construction where an adjective and a numeral intervene: 'a beautiful 5 days in New York' [19,20].

Human language users likely learn rules, plus thousands of idiosyncratic constructions [21], through some combination of sophisticated statistical learning [22–24] and innate conceptual and/or linguistic machinery [25–27]. The result is the human ability to understand and produce grammatical and coherent linguistic utterances.

Functional linguistic competence. In addition to being competent in the rules and statistical regularities of language, a competent language user uses language to accomplish goals in the world [12,28,29]: to talk about things that can be seen or felt or heard, to reason about diverse topics; to make requests; to cajole, prevaricate, and flatter. People use language in tandem with other perceptual and cognitive systems, such as our senses and our memory, and deploy words as part of a broader communication framework supported by our sophisticated social skills. A formal language system in isolation is useless unless it can interface with the rest of perception, cognition, and action.

The capacities required to use language to do things in the world are distinct from formal competence and depend crucially on non-linguistic cognition (Figure 1). Thus, we define functional linguistic competence as non-language-specific cognitive functions that are required when using language in tandem with non-language-specific capacities in real-world circumstances.

Glossary

Abstraction: for our purposes, a linguistic representation that allows for generalization. Part-of-speech is one such example: words like 'dog' and 'cat' belong to the abstract category of 'nouns'.
Architectural modularity: involves explicitly building distinct modules into a computational model, with each module responsible for achieving different goals.
Emergent modularity: induction of modularity through the model training process, without explicitly building it into the architecture.
Fine-tuning: a process by which, after a model is pretrained, it receives additional training on new data, often for a specific purpose.
Formal linguistic competence: the ability to get the *form* of language right. It includes knowledge of word formation (e.g., phonology and morphology), knowledge of word meaning, and knowledge of rules and statistical patterns for how words combine to create sentences. Note that our use of the term 'competence' differs from the classic competence/performance distinction in linguistics, given that in both models and humans separating competence and performance is often difficult.
Functional linguistic competence: the ability to use language to accomplish things in the world. It relies on a host of non-language-specific cognitive domains like formal reasoning, world knowledge, situation tracking, and social cognition.
Hierarchical structure: a crucial property of language that allows it to be more than just a linear sequence of words. Rather, how words go together in a sentence is better captured by a tree-like structure, where some words and phrases are nested inside larger phrases.
Language network: the interconnected set of brain regions that respond selectively to language but not to non-linguistic inputs and tasks.
Large language models (LLMs): models based on deep neural architectures (often but not always transformers) and trained on massive amounts of text using a word-in-context prediction task (sometimes with additional training objectives incorporated during or after the main training process). The term 'large' refers to the number of parameters in these models, which

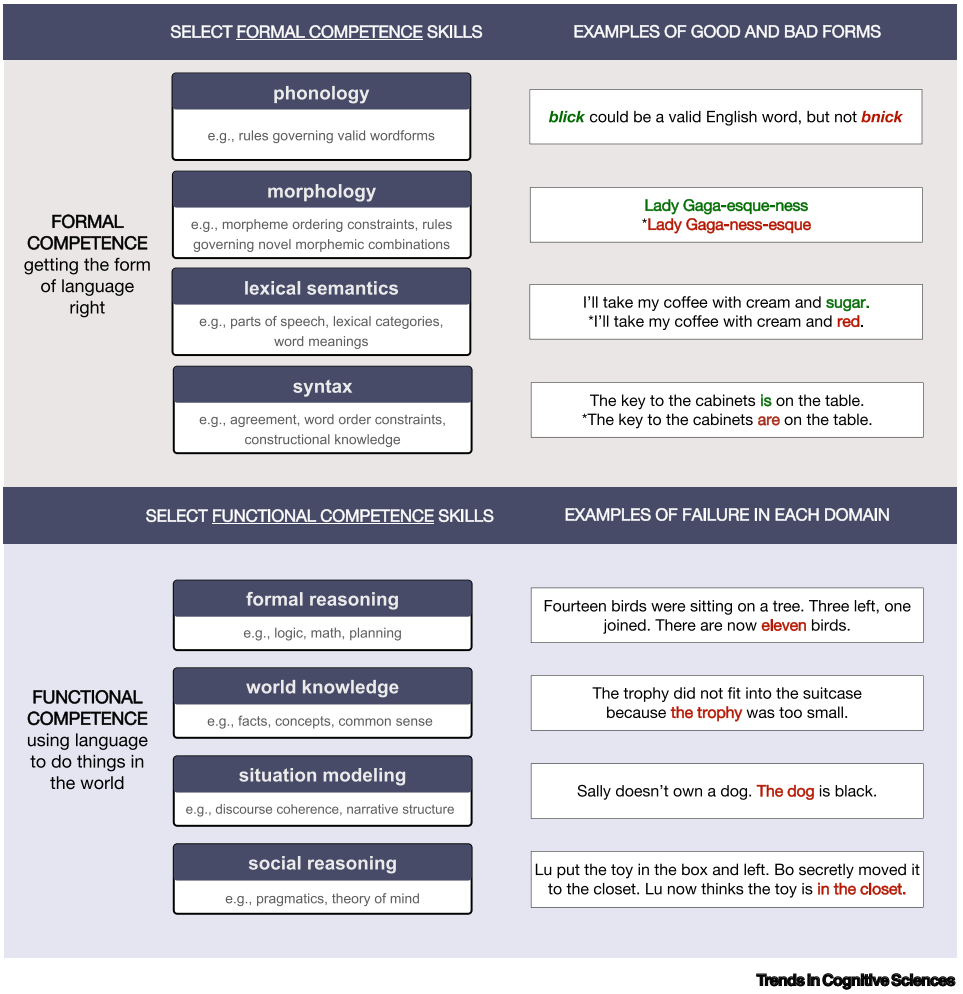

Figure 1. Separating formal and functional competence. Successful use of language relies on multiple cognitive skills, some of which (required for formal competence) are language-specific and some (required for functional competence) are not. Determining whether a particular failure stems from a gap in formal competence or functional competence is key to evaluating and improving language models.

ranges from millions to billions, as well as the size of the training data.
Pretraining: the process by which a model is first trained on a general task (for LLMs, typically a text prediction task) before being trained or used for a more specialized purpose.
Reinforcement learning from human feedback (RLHF): a process by which reinforcement learning techniques are used to impart human preferences (e.g., as to which of two model outputs is preferred) to a model. It seems to lead to significant improvements on functional tasks.
Theory of mind: a cognitive skill that enables thinking and reasoning about the minds of others (i.e., what others know, believe, want, etc.).
Tokens: the basic units in language models. In earlier language models, they were often words or morphemes. In today's LLMs, they are often inferred from large amounts of text using an algorithm like Byte Pair Encoding. They can resemble words and morphemes, but sometimes also linguistically unnatural units.

Motivation for the distinction between formal versus functional linguistic competence

Our motivation for the distinction between formal and functional linguistic competence comes from what we know about the architecture of the human mind. In humans, language is robustly dissociated from the rest of high-level cognition, as well as from perception and action. Next we briefly summarize a body of evidence from cognitive science and neuroscience that supports this dissociation.

The language network supports language processing in the human brain. Human language processing draws on a set of interconnected brain areas in the frontal and temporal lobes (typically in the left hemisphere). This **language network** supports both comprehension (spoken, written, and signed) [30–33] and production [34,35]; is sensitive to linguistic regularities at multiple levels: from phonological/sub-lexical [36] to phrase/sentence level [37,38]; and supports linguistic operations related both to the processing of word meanings and to combinatorial semantic and

syntactic processing [35,38]. Damage to the language network leads to linguistic deficits [39,40]. This tight link between the language network and language function indicates that these brain regions are responsible for language processing in humans.

The language network does not support non-linguistic cognition. The language network is remarkably selective for language. Evidence of a strong dissociation between language processing and non-linguistic abilities comes from two main sources: (i) functional brain imaging studies of neurotypical adults, and (ii) behavioral investigations of individuals with aphasia, a language impairment typically caused by a stroke or degeneration.

Brain imaging techniques like functional MRI (fMRI) are used to observe real-time activity in the language network in healthy individuals. Given its high spatial resolution, fMRI is well-suited to study whether any two cognitive abilities draw on the same brain structures. For example, to ask whether language and mathematical reasoning recruit the same brain areas, we can have participants perform a language task and a math task while in an MRI scanner and then test whether brain regions that are active during language processing are also active when participants solve a math problem. This approach reveals that the language network is extremely selective for language processing: it responds reliably when people listen to, read, or generate sentences, but not when they perform arithmetic tasks, engage in logical reasoning, understand computer programs, listen to music, categorize objects or events, reason about people's mental states, or process non-verbal communicative information like facial expressions or gestures [37,41–48].

Studies of individuals with aphasia provide a unique opportunity for testing which cognitive capacities rely on linguistic representations. Of particular interest are cases of 'global aphasia', which affects both production and comprehension. Individuals with global aphasia exhibit severe linguistic deficits that spare nothing but a small set of words. If some aspects of non-linguistic cognition draw on the same resources as language, then individuals with severe linguistic deficits should invariably exhibit impaired performance on the relevant non-linguistic tasks. However, despite the nearly complete loss of linguistic abilities, individuals with severe aphasia can have intact non-linguistic cognitive abilities: they can play chess, solve arithmetic problems, leverage their world knowledge to perform diverse tasks, reason about cause and effect, and navigate complex social situations [49].

In summary, evidence from brain imaging studies and from individuals with aphasia is remarkably consistent: the mechanisms that process language in the human brain do not support non-linguistic cognitive tasks. This sharp dissociation suggests that, in examining language models' functionality, we should separate their linguistic abilities from their abstract knowledge and reasoning abilities, which can be probed, and perhaps even learned, through a linguistic interface, but which require more than formal linguistic competence.

LLMs have largely mastered formal linguistic competence in English

In a 2019 interview, Chomsky remarked: 'We have to ask here a certain question: is [deep learning] engineering or is it science? [...] On engineering grounds, it's kind of worth having, like a bulldozer. Does it tell you anything about human language? Zero' (https://www.youtube.com/watch?v=cMscNuSUy0I). The view that deep learning models are not of scientific interest remains common in linguistics and, despite many arguments for integrating such models into research on human language processing and acquisition [50–52] and a chorus of arguments that they should be taken seriously as linguistic and cognitive models [53–55], their integration into language research still encounters resistance.

In this section, we evaluate the performance of LLMs *qua* language models by asking whether these models have made progress towards achieving formal linguistic competence, the kind of competence supported by the language-selective network in the human brain. We argue that LLMs have turned out to be surprisingly successful at mastering formal competence, qualitatively different in their formal linguistic capacities from models from before roughly 2018 in a way that was predicted by few practitioners in the field and which was unexpected given longstanding claims that grammatically competent systems would require strong language-specific priors. With surprise comes information: models' successes are informative for linguistic theorizing.

Statistical language models: some fundamentals

LLMs arose from several earlier approaches in computational linguistics, including statistical language modeling, word embeddings, and connectionism (an earlier term for the approach that morphed into today's deep learning). Similar to earlier statistical language models, LLMs are usually first trained on a word prediction task (the same task used for training n-gram models going back to Shannon's work in the mid-20th century; see [56] for historical overview). Similar to approaches in distributional semantics and word embeddings (for overviews, see [57,58]), LLMs represent linguistic information as vectors in a high-dimensional space. Similar to earlier connectionist approaches [59,60], LLMs are neural networks: a class of machine learning systems that was originally inspired by the human brain and learns its parameters from the input data. All of these approaches stand in contrast to models that use explicit, structured hierarchical representations of syntactic rules (see [61] for a discussion of these two divergent paradigms).

N-grams and word embedding models achieved some success in various domains in natural language processing (e.g., spelling correction, spam classification, sentiment analysis). However, they never approached human-level performance on general language tasks like text generation, leading to claims that purely statistical approaches would never be able to capture the richness of natural language, particularly in complex syntactic, morphological, and semantic domains (e.g., [62]). For instance, it has been claimed that statistical approaches, which use linear strings of words as input, are in principle unable to learn rare and complex syntactic features that require representing phrases and sentences hierarchically [63]. This pessimism is now challenged by LLMs.

LLMs are typically first trained on a training set constructed from a massive amount of text from the web. During **pretraining**, LLMs have a simple objective: predict a held-out **token** (the basic unit in LLMs, often but not always corresponding to words or morphemes [64]) based on a fixed number of previous tokens. The predicted token is then compared with the ground truth (which token actually occurred in that sentence), and the error signal is propagated back through the model to update its many parameters. The token prediction objective is often used as a pretraining step , and the model is then **fine-tuned** for a more specific task.

Although it is tempting [65] to move the goalposts and focus on what these models are still unable to do, we argue that the remarkable advances in LLMs' ability to capture various linguistic phenomena should not be overlooked. Significant formal linguistic abilities arise in models on the scale of GPT-2 or BERT and seem to plateau at a high level in contemporary LLMs (Box 1).

LLMs learn core aspects of human language processing

For LLMs to be useful as models of language processing in humans, we must be convinced that the models encode the abstract phonological, morphological, syntactic, and semantic rules that characterize human language (see Box 2 for a distinction between 'linguistic' and conceptual semantics). Although interesting differences exist between linguistic processing in LLMs and

Box 1. The path toward formal linguistic competence

When did LLMs achieve formal linguistic competence? Table I shows text generations from an n-gram model, an RNN (a recurrent neural network, which were state-of-the-art language models in the mid-2010s, before transformers), GPT-2, GPT-3, and GPT-4.

The n-gram model and RNN clearly fail to show formal linguistic competence: the sentences are not grammatical or sensible. By contrast, GPT-2 (a model produced just 8 years after the RNN in Row 2) generates grammatical output: the verbs agree with their subjects, parts of speech are used appropriately, relatively complex syntactic constructions like the relative clause 'that has nothing to do with who we are' is appropriately tied to a noun phrase. However, compared with the GPT-3 and GPT-4 output, it makes less sense: what does 'everything about being with people' mean? Finally, the outputs of GPT-3 and GPT-4 are both grammatical and semantically coherent, such that they can plausibly be taken for answers generated by a human.

Table I. Examples of texts generated by increasingly more advanced language models in response to a prompt 'The meaning of life is...' (following Sutskever *et al.* [194] and using their example verbatim in Row 2). Starting with GPT-2, outputs are grammatical and largely use the rules and patterns of English (formal linguistic competence).

Model	Text
3-gram model trained on *Moby Dick*	*The meaning of life is* gulped and gone, bleeds with keenest anguish at the same muteness of humanity over all creations. Just so with Ahab, 'says I;' look here you sabbee me, has every reason to be served.
RNN from [194]	*The meaning of life is* the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced.
GPT-2 [195]	*The meaning of life is* something simple that has nothing to do with who we are. Everything about being with people is tied to those characteristics, and people are just machines.
GPT-3 davinci-002 [89]	*The meaning of life is* a mystery to us all, and a question that will never be answered. Life is what we make of it, and it's up to each of us to find our own meaning in life.
GPT-4 (via ChatGPT interface)	*The meaning of life is* subjective and varies greatly depending on personal beliefs, values, and experiences. Some people find meaning in relationships, personal achievements, or spiritual beliefs, while others may see it as a journey of self-discovery, learning, or contributing to the greater good.

humans [66,67], there are also important similarities. Here, we review evidence that LLMs succeed as models of formal linguistic competence. We focus primarily on syntax, showing evidence of mastery on grammatical benchmarks as well as evidence for emergent syntactic structure in LLMs. But similar successes, both in performance and emergent structure, have been shown in other linguistic domains (e.g., emergent phonological structure [68], productive generation of morphologically complex neologisms [69], rich lexical semantic information [70], etc.).

Box 2. What about semantics?

Does semantics fall under formal or functional linguistic competence? The answer depends on what kind of semantics we are talking about.

One meaning of semantics, often associated with compositional and lexical semantics, concerns the way that meaning is derived from words and their combinations. We consider this aspect part of formal linguistic competence. Indeed, the language network in the brain responds both to lexical semantics(i.e., retrieving the meaning of individual words) and to compositional semantics (i.e., constructing the meaning of multi-word utterances) [38]. Insofar as LLMs are highly sensitive to lexical and combinatorial semantics, they parallel the language network.

The second meaning of semantics is something closer to 'general conceptual knowledge' (used in contexts such as 'non-verbal semantics', e.g., extracting the meaning of a picture). This definition is closely related to the notion of world knowledge, which we discuss in our section on world knowledge. Given the fact that conceptual knowledge and reasoning does not *have* to operate over linguistic inputs but is nonetheless essential for fluent language use, we classify it under functional linguistic competence.

LLMs perform well on benchmarks of diverse linguistic phenomena. By being trained for word prediction, transformer models learn a lot about the structure of language, including linguistic features that, even recently, were thought to be beyond the scope of statistical models. These models have succeeded not just on tests of general language understanding developed by the natural language processing (NLP) community (e.g., GLUE tasks [71]), but, critically for our purposes, on tests of linguistic competence in English and other languages with massive corpora available (see Box 3 for discussion of lower-resourced languages).

The benchmark BLiMP [72], for instance, contains minimal pairs of grammatical versus ungrammatical sentences across a diverse range of complex linguistic phenomena like filler-gap dependencies ('Bert knew what many writers find' versus '*Bert knew that many writers find') and negative polarity items ('The truck has clearly tipped over' versus '*The truck has ever tipped over'). Strikingly, a model [73] submitted to the BabyLM challenge [74] achieved 86% on BLiMP (cf. human baseline of 89%) despite being trained on an amount of data comparable with what a human child might be exposed to (Box 3). Models achieve similarly impressive results on other linguistic benchmarks like SyntaxGym [75] and there are now dozens of investigations of specific complex linguistic phenomena (some of which we discuss later).

Box 3. Limitations of LLMs as human-like language learners and processors

A preponderance of evidence suggests that LLMs acquire formal linguistic competence. Here, we address three common criticisms of LLMs as models of human language processing.

Excessive reliance on statistical regularities

LLMs succeed, in part, by learning statistical regularities and they can be 'right for the wrong reason' [196]. Many of LLMs' successes can be explained by patterns present in training data, such that a slight deviation from these patterns can drastically decrease performance [197]. In particular, adding noise or distracting information can degrade model performance [88,198] in ways that do not always affect humans. These results raise the related question of whether LLMs are just storing and regurgitating their inputs. But a study of GPT-2's novel behavior [69] showed that, although n-grams up to length 4 often appeared in the training set, GPT-2 generated mostly novel 5-grams and above. Thus, LLMs seem to be capable of some meaningful generalizable morphosyntactic knowledge beyond mere memorizing.

Unrealistic amounts of training data

Today's best LLMs are trained on vastly more data than a child is exposed to [199], and some evidence suggests that a model's training dataset would need to be unrealistically large to handle some constructions in a human-like way [200], without stronger priors [201,202]. This difference in the amount of input that models versus human language learners require is sometimes taken to imply that the models will necessarily be un-human-like. However, there is reason for optimism. The BabyLM competition [74] solicits submissions of language models trained on either a fixed corpus of 10M or 100M words, which are posited to be in the range of the number of words heard by a 10-year-old child. The best-performing model in 2023 [203] achieved impressive performance on the BLiMP benchmark for syntactic minimal pairs, suggesting the possibility that models can learn grammar from smaller amounts of data (see also [80], among others). It has also been shown that smaller models still provide good matches to human neural responses [204].

Although performance of smaller language models is not perfect, improvements in language models, including the use of more cognitively-inspired architectures and learning algorithms, could lead to strong performance with far less training data. As such, important questions are: what inductive biases are introduced by the model architectures and can those biases be made more human [201]?

Insufficient tests on languages other than English

Because LLMs are data-hungry, they work best on languages for which vast corpora are available. For most human languages, this is not the case. Moreover, the architectures themselves may be biased towards English and other European languages [205], and not all languages are equally easy to model given existing infrastructure [206]. That said, evidence is growing of strong performance in a variety of languages [207] and successful transfer of models to low-resource languages [208]. We should still, however, proceed with caution in assuming that the success of LLMs will extend to all languages. This is of particular concern for languages that are typologically distinct from English or have a different modality (e.g., signed languages).

LLMs learn hierarchical structure. In human languages, words are combined to make compositional meanings. In a multi-word sentence, the individual words' meanings do not simply get added linearly one by one. Instead, they can be combined hierarchically into tree-like structures.

The **hierarchical structure** in language manifests in many ways. One prominent example is non-local feature agreement. In English and many other languages, verbs agree with their subjects. For instance, a plural subject uses the verb 'are', whereas a singular subject uses 'is'. A bigram model, which simply stores frequencies of two-word strings, could learn that 'The keys are on the table' is more probable than 'The keys is on the table' by knowing that 'keys are' is more common than 'keys is'. But such a model would not be able to learn that the subject and verb agree even if arbitrarily far apart: for instance, 'The keys to the old, wooden kitchen cabinet are on the table' has six intervening words between the subject and verb, and yet 'are' still agrees with 'keys' and not with 'cabinet'. However, a model that learns the underlying hierarchical structure of English should be able to keep track of this long-distance subject–verb dependency [76].

Today's LLMs perform long-distance number agreement well above chance, preferring the grammatical over a non-grammatical sentence continuation even in the presence of intervening distractor words [77,78], although some models can be distracted by frequency effects (such as differences in the frequency between the singular and plural forms [79]). In a similar vein, LLMs can handle other constructions that require complex hierarchy, like filler-gap dependencies [80]. Finally, studies that examine the internal geometry of the models' sentence representations [81], studies that causally intervene on models' internal representations [82], and studies that turn on and off specific model 'neurons' [83,84] have provided mechanistic insights into how an LLM might represent hierarchical structure and establish non-local structural dependencies.

LLMs learn linguistic abstractions. Following [85], we define an **abstraction** as a generalized linguistic representation, such as a part-of-speech category (e.g., noun or verb) or grammatical role (e.g., subject or object), that goes beyond simple storage of input and allows for generalization. The very notion of subject–verb agreement, outlined in the previous section, relies on the abstract categories of subject and verb. As described in [77], in a sentence like 'dogs in the neighborhood often... (bark/barks)', a model might learn a shallow version of the agreement rule, namely, that the collocation of 'dogs' and 'bark' in the same sentence is more common than 'dogs' and 'barks'. However, a model that has an abstract representation of categories like grammatical subject, grammatical number, and verb should be able to handle long-distance number agreement even for novel combinations of words.

One way to test a model's knowledge of abstract rules is by using semantically nonsensical sentences, like 'The colorless green ideas I ate with the chair... (sleep/sleeps)'. Models have been shown to perform the agreement task well in several languages, even on these semantically anomalous sentences [77].

An even more stringent test for linguistic abstraction asks whether LLMs can apply morphosyntactic rules to novel words. A study of BERT's abstraction capabilities [86] showed that BERT has some ability to generalize grammatical categories. They give the model novel words, used in phrases, as input (e.g., 'the blick' where blick is likely a noun and 'they dax' where dax is likely a verb) and test whether, based on the input, the model can generalize the part-of-speech category (e.g., assign a higher score to 'I went to a blick' than to 'I went to a

dax'). They conclude that BERT succeeds partially at this task: it does learn to generalize, but only after repeated examples (but see [87,88] for ways in which the word itself affects compositional ability). Models also seem to (often) be able to use novel words appropriately [69,89].

A large body of work has tested linguistic abstraction in LLMs using a method called probing [90,91]. In this literature, a classifier is often trained to take as input internal model representations and then predict as output an abstract category, such as part-of-speech or dependency role. The logic of the probe is to test whether these abstract categories can be successfully recovered from the internal model states. Using this approach, it has been claimed that LLMs 'rediscover the classical NLP pipeline' [92], learning at various layers features like part-of-speech categories, parses, named entities, and semantic roles (although see [93]).

Importantly, a human-like language model is not expected to rely solely on abstract rules. Humans use diverse cues in their language learning and processing that sometimes override or conflict with strict hierarchical syntactic processing (e.g., [94,95]). Humans also rely, to varying extents, on memorizing previously seen input, as opposed to purely applying abstract rules [21,85]. Thus, when evaluating formal competence in LLMs, it is essential to directly compare their performance with that of humans [96]. For instance, a re-examination [97] of an earlier study [98] showed that apparent syntactic agreement deficits in GPT-2 occurred on instances that were also challenging for humans. Overall, LLMs clearly learn some linguistic abstraction, even if the degree of that abstraction remains a matter of debate (as it does for humans).

LLMs learn constructions. Recent evidence suggests that LLMs learn syntactic constructions [99–101]. These constructions can be idiosyncratic, lexically sensitive, and relatively rare, such as 'a beautiful 5 days in Austin' [102]. LLMs also show some amount of sensitivity to the preposing in prepositional phrase construction ('Surprising though it may be...'), even when the gap crosses a finite clause boundary ('Surprising though *I know* it may be') [103]. They achieve this sensitivity even though such examples crossing the finite clause boundary are vanishingly rare: only 58 examples out of seven billion sentences in a corpus. The fact that models can learn that some vanishingly rare constructions are grammatical, whereas other equally rare constructions are not, suggests that LLMs meaningfully learn something about syntax.

Models are also sensitive to the form of the comparative correlative 'the better the syntax, the better the semantics' [104]. However, this sensitivity does not mean that they are sensitive to the semantic implications of the construction. Indeed, it appears that inferences based on these sentences can be a challenge (e.g., knowing that if I say 'the better the syntax, the better the semantics' and then tell you that the syntax is better, this means the semantics is better). This asymmetry nicely illustrates the formal/functional distinction: the model clearly knows how to use the construction and get the form right without necessarily being able to access the intended meaning. We discuss these issues in more detail in later sections.

LLMs are predictive of activity in the human language network

As discussed earlier, language processing in humans relies on a dedicated brain network. This network exhibits all the hallmarks of formal linguistic competence: it is sensitive to abstract hierarchical rules in isolated phrases and sentences [31,105–107], in naturalistic narratives [108–111], and in syntactically well-formed but semantically empty ('jabberwocky') stimuli [31,49,106]. The language network is also sensitive to specific word co-occurrences (e.g., as evidenced by sensitivity to n-gram surprisal [108]), indicating that it learns not only the rules, but also the patterns of language. The language network's selectivity for linguistic versus non-linguistic

inputs, along with its sensitivity to linguistic rules and patterns, allows us to operationalize formal linguistic competence as a set of computations that in humans take place within the language network.

If LLMs and the human language network perform similar computations to achieve formal linguistic competence, we expect to observe similarities in their internal organization (see [54] for similar arguments in the domain of vision). And indeed, LLMs and the human language network exhibit non-trivial similarities.

First, the internal architecture of LLMs resembles that of the language network. Both operate at the level of abstract linguistic units (words/tokens) rather than modality-specific representations, such as pixels or acoustic waveforms, and combine these unit-level representations into composite representations of phrases and sentences. Neither system shows clear spatial segregation for syntactic and semantic processing (LLMs: [92,112]; brain: [38,111]), indicating that these processes are tightly functionally coupled in both.

Second, one can establish a direct mapping between internal LLM representations and neural activity patterns within the language network. This mapping can be successfully used to predict brain responses to novel sentences and words in previously unseen contexts [113–115]. This similarity between sentence activation patterns in LLMs and the brain is suggestive of similar representational mechanisms that support computations in these systems.

We do not claim that the correspondence between LLMs and the language network is one-to-one. For instance, LLMs learn patterns outside traditional human linguistic competency, such as predicting newline characters [116]. Nevertheless, the fact that internal representations learned by contemporary LLMs contain sufficient information to predict the language network's responses to diverse linguistic strings indicates at least some correspondence between LLMs' representations and those in the language network.

Using LLMs as models of formal linguistic competence in humans

LLMs today generate highly coherent, grammatical texts that can be indistinguishable from human output. In doing so, they exhibit knowledge of hierarchical structure and linguistic abstractions, while resembling human brain responses during language processing. These models are not perfect learners of abstract linguistic rules, but neither are humans. We therefore conclude that LLMs possess substantial formal linguistic competence, at least in English.

LLMs have already overturned claims about the fundamental impossibility of acquiring certain linguistic knowledge, including hierarchical structure and abstract categories, from the statistics of linguistic input alone [117]. If language modeling continues to improve (including learning from more realistic kinds and amounts of data; Box 3), this would allow testing more general versions of this 'poverty of the stimulus' argument [118], including specific tests of what inductive biases might be necessary to successfully learn the rules and statistical regularities of human language. As such, LLMs have substantial value in the scientific study of language learning and processing.

Non-augmented LLMs fall short on functional linguistic competence

Real-life language use is impossible without non-linguistic cognitive skills. Understanding a sentence, reasoning about its implications, and deciding how to respond all rely on cognitive capacities that go beyond formal competence. In this section, we ask: how good are contemporary LLMs at functional linguistic competence?

We focus on four key capacities that are not language-specific but are nevertheless crucial for language use in real-life settings: (i) formal reasoning: a host of abilities including logical and mathematical reasoning, computational thinking, and novel problem solving; (ii) world knowledge: factual and commonsense knowledge about agents, objects, properties, actions, events, and ideas; (iii) situation modeling: the dynamic tracking of objects, agents, and events as a narrative/conversation unfolds over time; and (iv) social reasoning: understanding the social context of linguistic exchanges. An average conversation requires the use of all these capacities, yet none of them are specific to language use.

For each domain, we first describe its neural mechanisms in humans and then discuss how well contemporary LLMs have mastered the domain. We conclude that, unlike formal competence, functional competence of LLMs is uneven, often requiring specialized fine-tuning and/or lacking human-like robustness and generality. In Box 4, we highlight the importance of properly evaluating LLMs; evaluation issues can occur in studies of either formal or functional competence, but we believe they have led to a particularly large amount of overclaiming about models' functional competence.

Box 4. Important considerations for evaluating functional competence

In discussing different domains of functional competence, it is important to highlight two key considerations.

(i). Fine-tuning on the task and the challenge of closed models

When discussing whether an LLM excels in a particular domain, it is essential to note whether the model has been fine-tuned on the task of interest. Formal competence skills can be observed in many LLMs trained on word-in-context prediction, without the need to specifically fine-tune them on syntactic trees or other grammatical abstractions. Functional competence skills, however, are often boosted through additional fine-tuning on some relevant corpus, the task of interest, or using more general fine-tuning techniques like reinforcement learning based on human feedback (RLHF). Claims such as 'LLMs succeed at theory of mind' often apply to fine-tuned models, not the generic word-in-context prediction machines.

An extreme case of task-specialized fine-tuning is when the task materials are, in fact, included in the model's training set. LLMs can memorize large amounts of information (and this trend becomes more pronounced in larger models [209]), such that prompting them with the exact sentences (or similar sentences) they have encountered during training will give highly inflated performance metrics.

Identifying whether performance boost on a given task comes from generic changes to the model (such as increasing the model size or the amount of training data) versus task-specific changes (such as fine-tuning on a similar task) is essential for further model understanding and improvement.

The closed nature of many LLMs (like the latest GPT models) make evaluating them in these ways difficult or impossible. Thus, we believe that useful scientific knowledge will increasingly come from studying models whose training data, architecture, and training procedure are transparent and able to be studied [3,53].

(ii). Generalizable, robust performance

When probing a particular ability, it is important to evaluate the models' performance across a variety of tasks, prompts, and scenarios. A failure to generalize beyond a particular surface-level form of the input may indicate that a model is using a non-human-like computational mechanism. For instance, it is sometimes the case that a model might perform well on a particular benchmark by leveraging low-level co-occurrence patterns (e.g., learning to predict that sentence A entails sentence B just because sentences A and B have overlapping lexical items), but as soon as these obvious patterns are removed, the model's performance might drop to chance levels [196].

A particular danger when evaluating model abilities is excessive reliance on single examples. As in any scientific endeavor, assessing a phenomenon requires multiple observations to ensure generalizability and replicability. Thus, we here emphasize systematic benchmark-based evaluations rather than single examples (although those can be informative for illustrating a phenomenon or for initiating a more in-depth exploration).

Formal reasoning

Language allows people to discuss highly abstract ideas, turn ideas into scientific and philosophical theories, construct logical syllogisms, and engage in formal debates. Unsurprisingly, language is often considered a cornerstone of complex reasoning [119,120]. However, neuroscience provides evidence that language and formal reasoning dissociate in cognitive systems and so a model that has mastered formal linguistic competence will not necessarily exhibit logical reasoning abilities.

Humans. Despite their close interplay, language and reasoning rely on distinct cognitive and neural systems. Unlike language, formal reasoning engages brain regions known as the multiple demand network [121], named so because these regions are engaged in many cognitively demanding tasks: logic [47], mathematical reasoning [41], physical reasoning [122], and computer code comprehension [46,123]. Human patient studies have provided causal evidence for the role of the multiple demand network in logical reasoning by showing that the amount of damage to these regions correlates negatively with performance on standard tests of fluid intelligence [124,125]. Importantly, the multiple demand network supports reasoning even when the task is presented linguistically [41,47,123], similar to how LLMs receive their prompts.

LLMs. Multiple studies have pointed out LLMs' limitations on tasks requiring formal reasoning, such as math problems. GPT-3 performs well on two-digit addition and subtraction but not on more complex tasks, such as three-digit addition or two-digit multiplication [89]. GPT-4 similarly shows good performance on small-digit but not large-digit mathematical operations [126]. Reasoning tests that break common co-occurrence patterns in the input or require multi-step operations also lead to model failure [127,128].

The most commonly cited reason for these failures is the failure of artificial neural nets to generalize to patterns outside their training distribution [128,129]. This generalization gap can be partially bridged by 'chain of thought' approaches, whereby a model is prompted to generate intermediate computation steps before arriving at an answer [130]. However, even these approaches do not lead to foolproof results [126]. Thus, more and more researchers pair LLMs with external modules that can carry out structural logical and mathematical computations, such as the Mathematica plugin (https://www.wolfram.com/wolfram-plugin-chatgpt/) or a probabilistic reasoning engine [131]. The shift toward augmenting LLMs with reasoning-specific modules is consistent with evidence from neuroscience: language and formal reasoning are distinct cognitive capacities that work best when they are supported by separate processing mechanisms.

World models 1: factual and commonsense knowledge

A commonly debated capacity in LLMs is their ability to leverage internal world models [131,132]. We break down the notion of world models into two components: world knowledge (factual and commonsense, this section) and situation tracking (the ability to maintain and update information about objects, agents, etc.; next section).

Humans. Evidence from neuroscience shows a dissociation between linguistic and semantic (world) knowledge. Individuals with language deficits may struggle to produce grammatical utterances and retrieve contextually appropriate words, but their ability to reason about objects and events presented non-linguistically often remains intact [42,133]. However, individuals who suffer from semantic dementia (a neurodegenerative disorder) retain the ability to speak but struggle with tasks that rely on world knowledge (e.g., knowing that pumpkins are typically orange)

even when the stimuli are presented non-verbally as pictures [134]. Thus, linguistic and semantic knowledge can be disentangled.

LLMs. LLMs have access to a wealth of knowledge about the world: word co-occurrence patterns in texts on the web contain both factual information (e.g., who was the first man on the moon) and commonsense information (e.g., the taste of lemon) [135]. If this information can be effectively extracted, LLMs would be able to serve as off-the-shelf knowledge bases [136]. However, world knowledge contained in LLM representations suffers from several major shortcomings.

First, LLMs routinely generate false statements, informally known as 'hallucinations'. This observation is unsurprising: their training objective is to generate plausible sentence continuations, with no reference to the underlying factual correctness of the resulting claims. Some developers have fine-tuned LLMs to provide links to sources that back up their claims; however, those citations can also be inaccurate [137].

Second, LLM outputs are often inconsistent: the same prompt phrased in different ways can elicit different responses [138]. They can also get 'distracted' by intervening information (e.g., an irrelevant claim inserted between a premise and a conclusion) [88].

Third, commonsense knowledge is often under-represented in language corpora: people are much more likely to communicate new or unusual information rather than commonly known facts [139]. As a result, LLMs can struggle on commonsense knowledge benchmarks [140], especially once low-level statistical cues are controlled for [9].

And fourth, explicitly stated factual knowledge is easy to access but hard to maintain, requiring constant updates; for instance, the answer to 'Who is the current president of the USA?' will change every 4 or 8 years. Whereas humans can update their knowledge representations via a single sentence, updating world knowledge in LLMs requires locating and editing this particular bit of knowledge in their internal parameters, a non-trivial task [141], especially because these edits should affect some other bits of knowledge (e.g., that the previously current president is now the past president) but leave many other facts unaffected [142].

A more human-like approach to world knowledge representation might require dissociating linguistic representation/processing and world knowledge storage/updates. Such approaches exist (e.g.,[143]) but have not yet reached dominance in the field, typically because of relatively low coverage of existing knowledge bases. Although we cannot rely on LLMs alone for accurate world knowledge claims, we might use them as a starting point for constructing detailed knowledge bases [144] and commonsense schemata [145].

World models 2: situation tracking

People can follow the plot of a story that spans multiple chapters or even multiple books. We can also remember many details weeks or months after a conversation. We accomplish these feats by leveraging language inputs to create a 'situation model', a mental model of entities, relations between them, and a sequence of states they had been in or events they had participated in [146]. Does the language network in humans construct a situation model based on its inputs? And how good are LLMs at building and updating situation models over time?

Humans. The language network in humans does not appear to track structure above the clause level [147,148]. Instead, integration of meaning over longer periods of time likely takes place within the so-called default network [149]. Crucially, the default network tracks both linguistic and non-linguistic narratives [150], indicating that situation modeling is not a language-specific skill.

LLMs. Situation modeling in LLMs faces two main challenges: (i) extracting information from many sentences in a row, and (ii) integrating incoming inputs to appropriately update information about entities and their states.

The first problem is currently being tackled by continuously increasing the models' context window (i.e., the number of words they can process in one go). This approach will inevitably run into computational challenges: when summarizing a book, having a model that simultaneously attends to each word in that book is vastly inefficient (although see some attempts to overcome this issue, e.g., [151]). A human-like solution to this problem might include hierarchical processing (e.g., generating a summary for each chapter and then for the whole book; for related approaches, see [152,153]).

Even when LLMs operate over shorter spans of text that easily fit inside their context windows, the question is: can they update their internal representations to track changes in the world? Some evidence suggests that they can [154], although LLMs make characteristically non-human-like mistakes when it comes to situation modeling: for instance, their outputs can refer to non-existent discourse entities ('Arthur doesn't own a dog. The dog is brown.' [155]). Thus, whether robust situation model building over shorter span of text is feasible using an LLM-only architecture remains a matter of debate.

Social reasoning

'Water!'

Wittgenstein famously used single-word utterances like this to show that linguistic meaning radically depends on context. Although this word's literal meaning is straightforward, the intended meanings are more varied. Is the word being gasped by a thirsty person in the desert? By a hiker warning his friend of a hidden stream? An impatient diner talking to a waiter? Work in cognitive science and linguistics has come to recognize that these context-dependent aspects of language are not just peripheral but a central part of human language production and understanding [12,28].

The set of skills required to infer the intended meaning of an utterance beyond its literal content is known as pragmatics. Pragmatics likely engages a variety of neural mechanisms [156–158], including both the language network and other brain regions. Thus, different types of pragmatic reasoning can be classified either as formal or as functional competence. Here, we focus on one core functional competence capacity required for pragmatics: social reasoning.

Humans. A wealth of neuroscientific evidence shows that the human brain has dedicated machinery for processing social information [44,159]. The most relevant to our current discussion is the **theory of mind** network [160], a set of brain regions that are engaged when their owner is attempting to infer somebody's mental state (with our without the use of language [161,162]). The specific contributions of the theory of mind network to language understanding can be divided into two broad categories. First, just like other functionally specialized brain

modules, it is engaged when processing semantic content that is specifically related to its domain: narratives that require inferring the mental state of the characters engage the theory of mind network [162] and texts that require inferring the characters' intentions evoke greater activity than those that do not [163,164]. Second, the theory of mind network is engaged more strongly during nonliteral language comprehension, including phenomena like jokes, sarcasm, indirect speech, and conversational implicature [158,165], in other words, in situations where understanding the meaning of an utterance requires inferring the intentions of the speaker. Thus, successful language understanding relies on our broader, non-language-specific social inference skills.

LLMs. Recent models, trained with RLHF, have shown strong performance in interpreting nonliteral utterances, such as metaphors and polite deceits, suggesting they can reach human or near-human performance on at least some pragmatic tasks [166]. That said, LLMs exhibit uneven performance across pragmatic domains: their ability to interpret sarcasm or to complete jokes was limited even as their metaphor comprehension abilities soared [166]. Overall, at least some forms of pragmatic inference might be acquired via pre-training coupled with targeted fine-tuning. It remains an open question whether aspects of pragmatics that are easiest for LLMs are those that are supported by the language network in humans.

LLMs' ability to solve theory of mind tasks has been subject to particular controversy. These tasks require both social knowledge and the ability to maintain a situation model. A typical example would feature character X moving an object from location A to location B while character Y is not around and so, does not see the move. The goal is to predict the true location of the object (location B) and the location where character Y believes the object is (location A). A bold claim that instruction-tuned LLMs have mastered theory-of-mind tasks [167] was quickly countered by a demonstration that including basic controls (such as character Y being told about the true object location) brought LLM performance to below-chance levels [168]. Several other studies have identified limitations in LLM performance on theory of mind tasks [169–171] (cf. [172]). One solution to overcome these limitations has been to augment an LLM with a symbolic tracker of entity states and character beliefs [173], an approach that mirrors the separation between language and theory of mind processing in humans.

Language input can bootstrap functional competence capabilities

Many non-linguistic cognitive capabilities can be substantially enhanced by language input. In humans, this relationship is particularly salient during development: babies learn new conceptual categories more easily when they are accompanied by linguistic labels [174] and children with delayed language access have delayed social reasoning abilities [175]. Even in adulthood, knowledge of specific number words predicts the ability to conceptually represent exact numbers [176]. Coupled with the fact that language inputs contain vast quantities of information about the world, and that language is both a crucial data source and representational substrate for much of people's world knowledge, this evidence suggests that, in principle, a model trained exclusively on language input could acquire much of functional linguistic competence.

Thus, we do not argue that functional linguistic competence is out of reach for language-based models; our main goals are: (i) to highlight the conceptual distinction between formal and functional linguistic competence, which in the human brain draw on separate neural circuits; and (ii) to demonstrate the gulf between LLMs' formal and functional linguistic abilities. These facts lead to a speculation that, like the human brain, models that can master language use would also require or benefit from separate mechanisms for formal and functional competence. We discuss this idea next.

Toward models that use language like humans

In this paper, we have advanced the thesis that formal and functional linguistic competence are distinct capabilities, with formal competence relying on distinct language machinery and function competence requiring the integration of diverse brain networks. We have shown that formal competence emerges in contemporary LLMs as a result of the word-in-context prediction objective; however, this objective alone appears insufficient for equipping LLMs with functional linguistic competence skills. Based on the neuroscience evidence, we suggest that models that succeed at real-life language use will need to be *modular*, mimicking the division of labor between formal and functional competence in the human brain.

We see at least two ways to separate LLM circuits responsible for formal and functional competence: explicitly building modularity into the architecture of the system (we call this **architectural modularity**) or naturally inducing modularity through the training process, both through the training data and the objective function (we call this **emergent modularity**).

Architectural modularity has a long history; it involves stitching together separate components, perhaps with quite specialized architectures [177,178]. Modern-day examples include a transformer language model paired with a separate memory module (e.g., [143,179]) or a model for visual question answering, which includes a language module, a vision module, and a reasoning module [180,181]. Such modular models achieve high task performance, are more efficient (i.e., can be trained on smaller datasets and have lower compute demands during inference), and show better generalizability (i.e., perform well on datasets with previously unseen properties). The modules of such models can be trained separately or together, similarly to how humans can flexibly combine different cognitive skills when learning to perform novel complex tasks.

Recently, the desire for this kind of modularity has expanded to include attempts to augment language models with the ability to call separate programs, as in including API calls [182], mathematical calculators [183], planners [184], and other kinds of modules that do specific structured operations.

Another approach in this vein uses LLMs as modules to translate a natural language query into code, which can then be passed to a symbolic module, which then generates an answer. Wong, Grand, *et al.* [131] outline a research program for this approach, showing that a version of GPT-3 fine-tuned to generate both natural language and code (Codex) can translate text input into meaningful structured probabilistic programs; inference in these programs can be used to reason over relational domains (like kinship systems), grounded domains (like visual scenes), and situations that require planning and understanding the plans of others. Their approach demonstrates a promising avenue for integrating what LLMs succeed at (namely, formal linguistic competence) with other cognitive modules that benefit from symbolic structure and abstraction.

The emergent modularity approach involves training models end-to-end (similarly to contemporary LLMs) while creating the conditions that facilitate the emergence of specialized model subcomponents over the course of training. Modular structure has been shown to spontaneously emerge in some end-to-end neural network systems in domains other than language (e.g., [185,186]), which suggests that emergent modularity may constitute an optimal solution to many complex tasks. One strategy for this approach to be successful is for the model architecture to incentivize the development of individual, specialized modules within the model. Transformers, the most popular architecture today, satisfy this condition to some extent by allowing different attention heads to attend to different input features (e.g., [187–189]); certain approaches promote modularization even more explicitly (e.g., by endowing transformers with a mixture-of-

experts architecture that incentivizes separate 'experts' to carry out different computations [190–192]).

A modular model architecture is much better aligned with the brain's functional architecture for language, which includes separate components for formal and functional competence. Is it possible to build formally and functionally competent systems that do not mimic the modular structure of the human brain? In theory, yes: systems with different underlying architectures (e.g., modular versus non-modular) can, in principle, exhibit similar behaviors. However, explicitly disentangling formal and functional competence skills at the architectural level is perhaps the most fail-safe path toward ensuring that an artificial intelligence (AI) model uses language in a human-like way.

Concluding remarks

Over the past few years, the discourse around language models has consisted of a curious mix of overclaiming and underclaiming [65]. While some claim models are on the verge of intelligence, others have pointed out the many failures of LLMs on a broad range of tasks, from number multiplication to generating factually true statements. Here, we have put these seemingly inconsistent reactions in dialog with prior and ongoing work in computational linguistics, cognitive science, and neuroscience. In particular, we argue that LLMs are remarkably successful on tasks that require a particular type of structural and statistical linguistic competence: formal linguistic competence. Although their performance is not yet fully human-like, these models achieve an impressive degree of success in representing and using hierarchical relationships among words and building representations that are sufficiently abstract to generalize to new words and constructions. As such, these LLMs are underused in linguistics as candidate models of human language processing.

We also review some of the LLMs' failures on tasks that target real-life language use, such as reasoning, while highlighting that the capabilities these tasks require are fundamentally distinct from formal language competence and rely on machinery in the human brain distinct from language processing machinery.

The failures of LLMs on non-linguistic tasks do not undermine their utility as models of language processing. After all, the brain areas that support language processing in humans also cannot do math, solve logical problems, or even track the meaning of a story across sentences or paragraphs. If we take the human mind and brain – a good example of generalized intelligence, as a guide, we might expect that future advances in developing intelligent systems will require combining language models with models that represent abstract knowledge and support complex reasoning, rather than expecting a single model (trained with a single word prediction objective) to do it all. Finally, to detect and monitor such advances, we need benchmarks that cleanly separate formal and functional linguistic competence (Box 5). Formally and rigorously evaluating functional competence in LLMs will be informative for both science and engineering (see Outstanding questions).

To those who have argued that the most interesting aspects of human language cannot be learned from data alone, we say that LLMs compellingly demonstrate the possibility of learning complex syntactic features from linguistic input (even if, as of now, much more input is required than a typical child gets exposed to). To those who criticize LLMs for their inability to do complex arithmetic or to reason about the world, we say, give language models a break: given a strict separation of language and non-linguistic capabilities in the human mind, we should evaluate these capabilities separately, recognizing successes in formal linguistic competence even when non-

Box 5. The need for better benchmarks

To assess progress on the road toward building models that use language in human-like ways, it is important to develop benchmarks that evaluate both formal and functional linguistic competence. This distinction can reduce the confusion that arises when discussing these models by combating the 'good at language -> good at thought' and the 'bad at thought -> bad at language' fallacies.

Several existing benchmarks already evaluate formal linguistic competence in LLMs [72,75] and can be complemented by additional tests of core linguistic features: hierarchy and abstraction. Benchmarks for evaluating different domains of functional linguistic competence, like commonsense world knowledge (e.g., WinoGrande [210]), can often be 'hacked' by LLMs by leveraging flawed heuristics [211]. This issue is likely exacerbated in large-scale heterogeneous datasets like BIG-bench [5]. Moreover, functional competence benchmarks often rely on a certain, often underspecified level of required formal linguistic competence skills and/or mix together different functional competence abilities. Designing benchmarks that carefully disentangle different components of language knowledge and use would therefore constitute an important step toward a more informative assessment of LLMs.

linguistic capabilities lag behind. Finally, to those who are looking to improve the state of machine learning systems, we suggest that, instead of, or in addition to, continuously scaling up the models [193], more promising solutions will come in the form of modular architectures, built-in or emergent, that, like the human brain, integrate language processing with additional systems that carry out perception, reasoning, and action.

Acknowledgments

For helpful conversations, we thank Jacob Andreas, Alex Warstadt, Dan Roberts, Kanishka Misra, students in the 2023 UT Austin Linguistics 393 seminar, the attendees of the Harvard LangCog journal club, the attendees of the UT Austin Department of Linguistics SynSem seminar, Gary Lupyan, John Krakauer, members of the Intel Deep Learning group, Yejin Choi and her group members, Allyson Ettinger, Nathan Schneider and his group members, the UT NLL Group, attendees of the KUIS AI Talk Series at Koç University in Istanbul, Tom McCoy, attendees of the NYU Philosophy of Deep Learning conference, Sydney Levine, organizers and attendees of the ILFC seminar, and others who have engaged with our ideas. We also thank Aalok Sathe for help with document formatting and references.

K.M. acknowledges funding from National Science Foundation (NSF) Grant 2104995. A.I. was supported by funds from the Quest Initiative for Intelligence. E.F. was supported by National Institutes of Health (NIH) awards R01-DC016607, R01-DC016950, and U01-NS121471 and by research funds from the Brain and Cognitive Sciences Department, McGovern Institute for Brain Research, and the Simons Foundation through the Simons Center for the Social Brain.

Declaration of interests

The authors declare no conflicts of interest.

References

1. Turing, A.M. (1950) Computing machinery and intelligence. *Mind* 59, 433–460
2. Chang, Y. *et al.* (2023) A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Tech.*, Published online January 23, 2024. https://doi.org/10.1145/3641289
3. Bommasani, R. *et al.* (2023) The foundation model transparency index. *arXiv*, Published online October 19, 2023. https://doi.org/10.48550/arXiv.2310.12941
4. Wang, A. *et al.* (2019) SuperGLUE: a stickier benchmark for general-purpose language understanding systems. 33rd Conference on Neural Information Processing Systems
5. Srivastava, A. *et al.* (2023) Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *arXiv*, Published online June 12, 2023. https://doi.org/10.48550/arXiv.2206.04615
6. Oh, B.-D. and Schuler, W. (2023) Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Trans. Assoc. Comput. Linguist.* 11, 336–350
7. Bubeck, S. *et al.* (2023) Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv*, Published online April 13, 2023. https://doi.org/10.48550/arXiv.2303.12712
8. Weizenbaum, J. (1966) Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 36–45
9. Elazar, Y. *et al.* (2021) Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguist.* 1012–1031
10. Marcus, G. (2020) The next decade in AI: four steps towards robust artificial intelligence. *arXiv*, Published online February 19, 2020. https://doi.org/10.48550/arXiv.2002.06177
11. Bender, E.M. and Koller, A. (2020) Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D. *et al.*, eds), pp. 5185–5198, Association for Computational Linguistics
12. Grice, H. (1975) Logic and conversation. In *Syntax and Semantics 3, Speech Acts* (Cole, P. and Morgan, J.L., eds), pp. 41–58, Academic Press
13. Clark, H.H. (1992) *Arenas of Language Use*, University of Chicago Press
14. Ouyang, L. *et al.* (2022) Training language models to follow instructions with human feedback. *Adv. Neural Inf. Processing. Syst.* 35, 27730–27744

Outstanding questions

How much functional competence can be acquired from the linguistic signal? Humans use language as a substrate for knowledge and so LLMs acquire non-linguistic information from the linguistic signal. How much information in this signal can be used to bootstrap functional competence? Are there aspects of functional competence that cannot be learned from language at all?

How can we train competent language models on smaller amounts of data? LLMs have achieved remarkable linguistic competence but they are trained on data very unlike what human children encounter. Although LLMs receive vastly more words (several orders of magnitude), they lack the richly structured and interactive input thought to be essential to child language acquisition. Would benefits emerge from training models in more interactive and human-like ways?

Will the formal competence successes of LLMs transfer to other world languages? Most LLM evaluations have taken place in English and a handful of other world languages. Building models for lower-resourced languages and evaluating them on both formal and functional dimensions is an important ongoing project.

How long will the LLM growth continue? Ten years ago, most researchers in the field would not have predicted that LLMs would be as advanced as they are today. Will current AI approaches lead to further revolutionary achievements in language and thought or would the mastery of functional linguistic competence require radically new approaches?

Which is more promising: architectural modularity or emergent modularity? If we want to build human-like modular systems, will it require explicitly building in functionally distinct components or can they be induced through end-to-end fine-tuning [e.g., with reinforcement learning from human feedback (RLHF)?]

How modular are today's LLMs? We argue that an LLM that achieves both formal and functional competence may need to rely on separate mechanisms

15. Mialon, G. *et al.* (2023) Augmented language models: a survey. *arXiv*, Published online February 15, 2023. https://doi.org/10.48550/arXiv.2302.07842
16. Halle, M. (1962) Phonology in generative grammar. *Word* 18, 54–72
17. Aronoff, M. and Fudeman, K. (2022) *What Is Morphology?*, John Wiley & Sons
18. Cruse, D.A. (1986) *Lexical Semantics*, Cambridge University Press
19. Dalrymple, M. and King, T.H. (2019) An amazing four doctoral dissertations. *Argumentum* 15, 2019
20. Keenan, C. (2013) A pleasant three days in Philadelphia: arguments for a pseudopartitive analysis. *PWPL* 19, 11
21. Goldberg, A.E. (2019) *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*, Princeton University Press
22. Bresnan, J. (2007) Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots Linguist. Search Evidential Base* 96, 77–96
23. A. Clark. Distributional learning as a theory of language acquisition. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, page 29, Gothenburg, Sweden, April 2014. Association for Computational Linguistics
24. Saffran, J. *et al.* (1996) Statistical learning by 8-month-old infants. *Science* 274, 1926
25. Chomsky, N. (1957) *Syntactic Structures*, Mouton
26. Gleitman, L.R. (1993) A human universal: the capacity to learn a language. *Mod. Philol.* 90, S13–S33
27. Jackendoff, R. (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution*, Oxford University Press
28. Clark, H.H. (1996) *Using Language*, Cambridge University Press
29. Bucholtz, M. and Hall, K. (2004) Language and identity. *Companion Linguist. Anthropol.* 1, 369–394
30. Deniz, F. *et al.* (2019) The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *J. Neurosci.* 39, 7722–7736
31. Fedorenko, E. *et al.* (2010) New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* 104, 1177–1194
32. MacSweeney, M. *et al.* (2002) Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain* 125, 1583–1593
33. Scott, T.L. *et al.* (2017) A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn. Neurosci.* 8, 167–176
34. Menenti, L. *et al.* (2011) Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychol. Sci.* 22, 1173–1182
35. Hu, J. *et al.* (2023) Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cereb. Cortex* 33, 4384–4404
36. T. I. *et al.* (2021) High-level language brain regions are sensitive to sub-lexical regularities. *bioRxiv*, Published online June 11, 2021. https://doi.org/10.1101/2021.06.11.447786
37. Fedorenko, E. *et al.* (2011) Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci.* 108, 16428–16433
38. Fedorenko, E. *et al.* (2020) Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition* 203, 104348
39. Bates, E. *et al.* (2003) Voxel-based lesion-symptom mapping. *Nat. Neurosci.* 6, 448–450
40. Wilson, S.M. *et al.* (2019) Language mapping in aphasia. *J. Speech Lang. Hear. Res.* 62, 3937–3946
41. Amalric, M. and Dehaene, S. (2016) Origins of the brain networks for advanced mathematics in expert mathematicians. *Proc. Natl. Acad. Sci. USA* 113, 4909–4917
42. Benn, Y. *et al.* (2023) The language network is not engaged in object categorization. *Cereb. Cortex* 33, 10380–10400
43. Chen, X. *et al.* (2023) The human language system, including its inferior frontal component in "Broca's area," does not support music perception. *Cereb. Cortex* 33, 7904–7929
44. Deen, B. *et al.* (2015) Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb. Cortex* 25, 4596–4609
45. Jouravlev, O. *et al.* (2019) Speech-accompanying gestures are not processed by the language-processing mechanisms. *Neuropsychologia* 132, 107132
46. Liu, Y.-F. *et al.* (2020) Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *eLife* 9, e59340
47. Monti, M.M. *et al.* (2012) Thought beyond language: neural dissociation of algebra and natural language. *Psychol. Sci.* 23, 914–922
48. Paunov, A.M. *et al.* (2022) Differential tracking of linguistic vs. mental state content in naturalistic stimuli by language and theory of mind (ToM) brain networks. *Neurobiology of. Language* 3, 419–440
49. Fedorenko, E. and Varley, R.A. (2016) Language and thought are not the same thing: evidence from neuroimaging and neurological patients: language versus thought. *Ann. N. Y. Acad. Sci.* 1369, 132–153
50. Linzen, T. (2019) What can linguistics and deep learning contribute to each other? Response to Pater. *Language* 95, e99–e108
51. Blank, I.A. (2023) What are large language models supposed to model? *Trends Cogn. Sci.* 27, 987–989
52. Jain, S. *et al.* (2023) Computational language modeling and the promise of in silico experimentation. *Neurobiol. Lang.*, Published online March 9, 2023. https://doi.org/10.1162/nol-a-00101
53. Frank, M.C. (2023) Openly accessible LLMs can help us to understand human cognition. *Nat. Hum. Behav.* 7, 1825–1827
54. Cao, R. and Yamins, D. (2021) Explanatory models in neuroscience: part 1–taking mechanistic abstraction seriously. *arXiv*, Published online April 10, 2021. https://doi.org/10.48550/arXiv.2104.01490
55. Baroni, M. (2022) On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. In *Algebraic Structures in Natural Language* (Lappin, S. and Bernardy, J.-P., eds), pp. 1–16, CRC Press
56. Jurafsky, D. and Martin, J.H. (2009) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd edn), Pearson Prentice Hall
57. Baroni, M. and Lenci, A. (2010) Distributional memory: a general framework for corpus-based semantics. *Computat. Linguist.* 36, 673–721
58. Erk, K. (2012) Vector space models of word meaning and phrase meaning: a survey. *Lang. Linguist. Compass* 6, 635–653
59. Rumelhart, D.E. and McClelland, J.L. (1986) *Parallel Distributed Processing*, MIT Press
60. Elman, J. (1993) Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99
61. Norvig, P. (2012) Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance* 9, 30–33
62. Pinker, S. and Prince, A. (1988) On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* 28, 73–193
63. Everaert, M.B. *et al.* (2015) Structures, not strings: linguistics as part of the cognitive sciences. *Trends Cogn. Sci.* 19, 729–743
64. Sennrich, R. *et al.* (2016) Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Erk, K. and Smith, N.A., eds), pp. 1715–1725, Association for Computational Linguistics
65. Bowman, S. (2022) The dangers of underclaiming: reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Muresan, S. *et al.*, eds), pp. 7484–7499, Association for Computational Linguistics
66. Lenci, A. (2023) Understanding natural language understanding systems. A critical analysis. *Sistemi Intelligenti* 35, 277–302
67. Van Schijndel, M. and Linzen, T. (2021) Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cogn. Sci.* 45, e12988

for different competence types. Mechanistic interpretability studies can shed light on the extent to which different cognitive tasks might segregate even within today's LLMs.

Should LLMs be described as individual language users or as distributions over potential user outputs? There are different ways to think about LLMs in the context of language use, for example, as individual language users ('agent-based view') or as tools augmenting human activities, like a calculator ('tool-based view') [212–214]. Which of these views will be the most fruitful way to think about LLMs as they gain additional competencies and become more widely used?

How much grounding do LLMs need to continue improving? How limited are text-only approaches, compared with approaches across different modalities [215,216]?

How much can LLMs ultimately tell us about human language and cognition? What are the cases where LLMs fall short as models of human language use? Are these discrepancies solvable or will they require language researchers to develop different paradigms?

68. Beguš, G. (2021) CiwGAN and fiwGAN: encoding information in acoustic data to model lexical learning with generative adversarial networks. *Neural Netw.* 139, 305–325
69. McCoy, R.T. *et al.* (2023) How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Trans. Assoc. Comput. Linguist.* 11, 652–670
70. Chronis, G. and Erk, K. (2020) When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (Fernández, R. and Linzen, T., eds), pp. 227–244, Association for Computational Linguistics
71. Wang, A. *et al.* (2018) GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Association for Computational Linguistics
72. Warstadt, A. *et al.* (2020) BLiMP: the benchmark of linguistic minimal pairs for English. *Trans. Assoc. Comput. Linguist.* 8, 377–392
73. Samuel, D. (2023) Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (Warstadt, A. *et al.*, eds), pp. 221–237, Association for Computational Linguistics
74. Warstadt, A. *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* (Warstadt, A. *et al.*, eds), pp. 1–34, Association for Computational Linguistics
75. Gauthier, J. *et al.* (2020) SyntaxGym: an online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 70–76, Association for Computational Linguistics
76. Linzen, T. *et al.* (2016) Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguist.* 4, 521–535
77. Gulordava, K. *et al.* (2018) Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers)* (Walker, M. *et al.*, eds), pp. 1195–1205, Association for Computational Linguistics
78. Linzen, T. and Baroni, M. (2021) Syntactic structure from deep learning. *Annu. Rev. Linguist.* 7, 195–212
79. Yu, C. *et al.* (2020) Word frequency does not predict grammatical knowledge in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Webber, B. *et al.*, eds), pp. 4040–4054, Association for Computational Linguistics
80. Wilcox, E.G. *et al.* (2022) Using computational models to test syntactic learnability. *Linguist. Inq.*, Published online April 20, 2023. https://doi.org/10.1162/ling-a-00491
81. 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(Vol. 1: Long and Short Papers)* (Burstei n, J. *et al.*, eds), pp. 4129–4138, Association for Computational Linguistics
82. Ravfogel, S. *et al.* (2021) Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (Bisazza, A. and Abend, O., eds), pp. 194–209, Association for Computational Linguistics
83. Mueller, A. *et al.* (2022) Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* (Fokkens, A. and Srikumar, V., eds), pp. 95–109, Association for Computational Linguistics
84. Lakretz, Y. *et al.* (2019) The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long and Short Papers)* (Burstein, J., ed.), pp. 11–20, Association for Computational Linguistics
85. Ambridge, B. (2020) Against stored abstractions: A a radical exemplar model of language acquisition. *First Lang.* 40, 509–559
86. Kim, N. and Smolensky, P. (2021) Testing for grammatical category abstraction in neural language models. In *Proceedings of the Society for Computation in Linguistics 2021* (Ettinger, A. *et al.*, eds), pp. 467–470, Association for Computational Linguistics
87. Kim, N. *et al.* (2022) Uncontrolled lexical exposure leads to overestimation of compositional generalization in pretrained models. *arXiv*, Published online December 21, 2022. https://doi.org/10.48550/arXiv.2212.10769
88. Misra, K. *et al.* (2023) COMPS: conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pretrained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (Vlachos, A. and Augenstein, I., eds), pp. 2928–2949, Association for Computational Linguistics
89. Brown, T.B. *et al.* (2020) Language models are few - shot learners. In *Advances in Neural Information Processing Systems* (159), pp. 1877–1901
90. Ettinger, A. *et al.* (2016) Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 134–139, Association for Computational Linguistics
91. Belinkov, Y. (2022) Probing classifiers: promises, shortcomings, and advances. *Comput. Linguist.* 48, 207–219
92. Tenney, I. *et al.* (2019) BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Korhonen, A. *et al.*, eds), pp. 4593–4601, Association for Computational Linguistics
93. Niu, J. *et al.* (2022) Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics* (Calzolari, N. *et al.*, eds), pp. 3143–3153, International Committee on Computational Linguistics
94. MacDonald, M.C. *et al.* (1994) The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676
95. Bates, E. and MacWhinney, B. (1989) Functionalism and the competition model. In *The Crosslinguistic Study of Sentence Processing* (MacWhinney, B. and Bates, E., eds), pp. 3–73, Cambridge University Press
96. Dasgupta, I. *et al.* (2022) Language models show human-like content effects on reasoning. *arXiv*, Published online October 30, 2023. https://doi.org/10.48550/arXiv.2207.07051
97. Lampinen, A.K. (2022) Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *arXiv*, Published online February 16, 2023. https://doi.org/10.48550/arXiv.2210.15303
98. Lakretz, Y. *et al.* (2021) Causal transformers perform below chance on recursive nested constructions, unlike humans. *arXiv*, Published online October 14, 2021. https://doi.org/10.48550/arXiv.2110.07240
99. Weissweiler, L. *et al.* (2023) Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)* (Bonial, C. and Tayyar Madabushi, H., eds), pp. 85–95, Association for Computational Linguistics
100. Tseng, Y.-H. *et al.* (2022) CxLM: a construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Calzolari, N. *et al.*, eds), pp. 6361–6369, European Language Resources Association
101. Tayyar Madabushi, H. *et al.* (2020) CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics* (Scott, D. *et al.*, eds), pp. 4020–4032, International Committee on Computational Linguistics

102. Mahowald, K. (2023) A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (Vlachos, A. and Augenstein, I., eds), pp. 265–273, Association for Computational Linguistics
103. Potts, C. (2023) *Characterizing English Preposing in PP constructions*, LingBuzz lingbuzz/007495
104. Weissweiler, L. *et al.* (2022) The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y. *et al.*, eds), pp. 10859–10882, Association for Computational Linguistics
105. Fedorenko, E. *et al.* (2016) Neural correlate of the construction of sentence meaning. *Proc. Natl. Acad. Sci.* 113, E6256–E6262
106. Pallier, C. *et al.* (2011) Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci.* 108, 2522–2527
107. Law, R. and Pylkkänen, L. (2021) Lists with and without syntax: a new approach to measuring the neural processing of syntax. *J. Neurosci.* 41, 2186–2196
108. Shain, C. *et al.* (2020) fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* 138, 107307
109. Brennan, J.R. *et al.* (2020) Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia* 146, 107479
110. Heilbron, M. *et al.* (2022) A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl. Acad. Sci.* 119, e2201968119
111. Reddy, A.J. and Wehbe, L. (2021) Can fMRI reveal the representation of syntactic structure in the brain? *Adv. Neural Inf. Proces. Syst.* 34, 9843–9856
112. Huang, J.Y. *et al.* (2021) Disentangling semantics and syntax in sentence embeddings with pre-traied language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Toutanova, K. *et al.*, eds), pp. 1372–1379, Association for Computational Linguistics
113. Caucheteux, C. and King, J.-R. (2022) Brains and algorithms partially converge in natural language processing. *Commun. Biol.* 5, 1–10
114. Goldstein, A. *et al.* (2022) Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* 25, 369–380
115. Schrimpf, M. *et al.* (2021) The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* 118,
116. Michaud, E.J. *et al.* (2023) The quantization model of neural scaling. *arXiv*, Published online January 13, 2024. https://doi.org/10.48550/arXiv.2303.13506
117. Piantadosi, S.T. (2023) *Modern language models refute Chomsky's approach to language*, Lingbuzz lingbuzz/007180
118. Chomsky, N. (1991) Linguistics and cognitive science: problems and mysteries. In *The Chomskyan Turn*, Blackwell
119. Dennett, D.C. (1994) The role of language in intelligence. In *What Is Intelligence? The Darwin College Lectures* (Khalfa, J., ed.), Cambridge University Press
120. Carruthers, P. (2002) The cognitive functions of language. *Behav. Brain Sci.* 25, 657–674 discussion 674–725
121. Duncan, J. (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci.* 14, 172–179
122. Fischer, J. *et al.* (2016) Functional neuroanatomy of intuitive physical inference. *Proc. Natl. Acad. Sci.* 113, E5072–E5081
123. Ivanova, A.A. *et al.* (2020) Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife* 9, e58906
124. Woolgar, A. *et al.* (2010) Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proc. Natl. Acad. Sci.* 107, 14899–14902
125. Woolgar, A. *et al.* (2018) Fluid intelligence is supported by the multiple-demand system not the language system. *Nat. Hum. Behav.* 2, 200–204
126. Dziri, N. *et al.* (2023) Faith and fate: limits of transformers on compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems*
127. Valmeekam, K. *et al.* (2022) Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*
128. Wu, Z. *et al.* (2023) Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv*, Published online August 1, 2023. https://doi.org/10.48550/arXiv.2307.02477
129. Zhang, H. *et al.* (2023) On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23* (Elkind, E., ed.), pp. 3365–3373, International Joint Conferences on Artificial Intelligence Organization
130. Wei, J. *et al.* (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Proces. Syst.* 35, 24824–24837
131. Wong, L. *et al.* (2023) From word models to world models: translating from natural language to the probabilistic language of thought. *arXiv*, Published online June 23, 2023. https://doi.org/10.48550/arXiv.2306.12672
132. Yildirim, I. and Paul, L. (2023) From task structures to world models: what do LLMs know? *arXiv*, Published online October 6, 2023. https://doi.org/10.48550/arXiv.2310.04276
133. Ivanova, A.A. *et al.* (2021) The language network is recruited but not required for nonverbal event semantics. *Neurobiol. Lang.* 2, 176–201
134. Patterson, K. *et al.* (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* 8, 976–987
135. Grand, G. *et al.* (2022) Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* 6, 975–987
136. Petroni, F. *et al.* (2019) Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Association for Computational Linguistics
137. Liu, N. *et al.* (2023) Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (Bouamor, H. *et al.*, eds), pp. 7001–7025, Association for Computational Linguistics
138. Sclar, M. *et al.* (2023) Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. *arXiv*, Published online October 17, 2023. https://doi.org/10.48550/arXiv.2310.11324
139. Gordon, J. and Van Durme, B. (2013) Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, pp. 25–30, Association for Computing Machinery
140. Liu, X. *et al.* (2022) Things not written in text: exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Muresan, S. *et al.*, eds), pp. 2365–2376, Association for Computational Linguistics
141. Kim, Y. *et al.* (2023) Carpe diem: on the evaluation of world knowledge in lifelong language models. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*
142. Meng, K. *et al.* (2022) Locating and editing factual associations in GPT. *Adv. Neural Inf. Proces. Syst.* 35, 17359–17372
143. Borgeaud, S. *et al.* (2022) Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pp. 2206–2240, PMLR
144. Cohen, R. *et al.* (2023) Crawling the internal knowledge-base of language models. In *Findings of the Association for Computational Linguistics: EACL 2023* (Vlachos, A. and Augenstein, I., eds), pp. 1856–1869, Association for Computational Linguistics
145. Chersoni, E. *et al.* (2019) A structured distributional model of sentence meaning and processing. *Nat. Lang. Eng.* 25, 483–502
146. Van Dijk, T.A. and Kintsch, W. (1983) *Strategies of Discourse Comprehension*, Academic Press
147. Lerner, Y. *et al.* (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31, 2906–2915

148. Jacoby, N. and Fedorenko, E. (2020) Discourse-level comprehension engages medial frontal theory of mind brain regions even for expository texts. *Lang. Cogn. Neurosci.* 35, 780–796
149. Buckner, R.L. and DiNicola, L.M. (2019) The brain's default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* 20, 593–608
150. Baldassano, C. *et al.* (2017) Discovering event structure in continuous narrative perception and memory. *Neuron* 95, 709–721
151. Su, J. *et al.* (2024) Roformer: enhanced transformer with rotary position embedding. *Neurocomputing* 568, 127063
152. Moirangthem, D.S. and Lee, M. (2020) Abstractive summarization of long texts by representing multiple compositionalities with temporal hierarchical pointer generator network. *Neural Netw.* 124, 1–11
153. Ruan, Q. *et al.* (2022) HiStruct+: improving extractive text summarization with hierarchical structure information. In *Findings of the Association for Computational Linguistics: ACL 2022* (Muresan, S. *et al.*, eds), pp. 1292–1308, Association for Computational Linguistics
154. Kim, N. and Schuster, S. (2023) Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Rogers, A. *et al.*, eds), pp. 3835–3855, Association for Computational Linguistics
155. Schuster, S. and Linzen, T. (2022) When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Carpuat, M. *et al.*, eds), pp. 969–982, Association for Computational Linguistics
156. Andrés-Roqueta, C. and Katsos, N. (2017) The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders. *Front. Psychol.* 8, 996
157. Levinson, S. (2000) *Presumptive Meanings: The Theory of Generalized Conversational Implicature*, MIT Press
158. Hauptman, M. *et al.* (2023) Non-literal language processing is jointly supported by the language and theory of mind networks: evidence from a novel meta-analytic fMRI approach. *Cortex* 162, 96–114
159. Saxe, R. (2006) Uniquely human social cognition. *Curr. Opin. Neurobiol.* 16, 235–239
160. Gopnik, A. and Wellman, H.M. (1992) Why the child's theory of mind really is a theory. *Mind Lang.* 7, 145–171
161. Saxe, R. and Kanwisher, N. (2003) People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage* 19, 1835–1842
162. Jacoby, N. *et al.* (2016) Localizing pain matrix and theory of mind networks with both verbal and non-verbal stimuli. *NeuroImage* 126, 39–48
163. Ferstl, E.C. and von Cramon, D.Y. (2002) What does the frontomedian cortex contribute to language processing: coherence or theory of mind? *NeuroImage* 17, 1599–1612
164. Saxe, R. and Powell, L.J. (2006) It's the thought that counts: specific brain regions for one component of theory of mind. *Psychol. Sci.* 17, 692–699
165. Hagoort, P. and Levinson, S.C. (2014) Neuropragmatics. In *The Cognitive Neurosciences* (5th edn), pp. 667–674, MIT Press
166. Hu, J. *et al.* (2023) A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Rogers, A. *et al.*, eds), pp. 4194–4213, Association for Computational Linguistics
167. Kosinski, M. (2023) Theory of mind may have spontaneously emerged in large language models. *arXiv*, Published online November 11, 2023. https://doi.org/10.48550/arXiv.2302.02083
168. Ullman, T. (2023) Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv*, Published online March 14, 2023. https://doi.org/10.48550/arXiv.2302.08399
169. Shapira, N. *et al.* (2023) Clever hans or neural theory of mind? Stress testing social reasoning in large language models. *arXiv*, Published online May 24, 2023. https://doi.org/10.48550/arXiv.2305.14763
170. Sap, M. *et al.* (2022) Neural theory-of-mind? On the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y. *et al.*, eds), pp. 3762–3780, Association for Computational Linguistics
171. Trott, S. *et al.* (2023) Do large language models know what humans know? *Cogn. Sci.* 47, e13309
172. Gandhi, K. *et al.* (2023) Understanding social reasoning in language models with language models. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*
173. Sclar, M. *et al.* (2023) Minding language models' (lack of) theory of mind: a plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Rogers, A. *et al.*, eds), pp. 13960–13980, Association for Computational Linguistics
174. Waxman, S.R. (2002) Early word-learning and conceptual development: everything had a name, and each name gave birth to a new thought. In *Blackwell Handbook of Childhood Cognitive Development*, pp. 102–126, Blackwell Publishing
175. Pyers, J.E. and Senghas, A. (2009) Language promotes false-belief understanding: evidence from learners of a new sign language. *Psychol. Sci.* 20, 805–812
176. Pitt, B. *et al.* (2022) Exact number concepts are limited to the verbal count range. *Psychol. Sci.* 33, 371–381
177. Bottou, L. and Gallinari, P. (1990) A framework for the cooperation of learning algorithms. *Adv. Neural Inf. Proces. Syst.* 3, 781–788
178. Ronco, E. and Gawthrop, P.J. (1997) Neural networks for modelling and control. *Rapp. Tech.* 97008,
179. Liu, Q. *et al.* (2022) Relational memory-augmented language models. *Trans. Assoc. Comput. Linguist.* 10, 555–572
180. Mao, J. *et al.* (2019) The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, ICLR
181. Hudson, D. and Manning, C.D. (2019) Learning by abstraction: the neural state machine. In *Advances in Neural Information Processing Systems*, pp. 5901–5914
182. Schick, T. *et al.* (2023) Toolformer: language models can teach themselves to use tools. *arXiv*, Published online February 9, 2023. https://doi.org/10.48550/arXiv.2302.04761
183. Cobbe, K. *et al.* (2021) Training verifiers to solve math word problems. *arXiv*, Published online November 18, 2021. https://doi.org/10.48550/arXiv.2110.14168
184. Liu, B. *et al.* (2023) LLM+P: empowering large language models with optimal planning proficiency. *arXiv*, Published online September 27, 2023. https://doi.org/10.48550/arXiv.2304.11477
185. Yang, G.R. *et al.* (2019) Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* 22, 297–306
186. Dobs, K. *et al.* (2022) Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* 8, eabl8913
187. Manning, C.D. *et al.* (2020) Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci.* 117, 30046–30054
188. Vaswani, A. *et al.* (2017) Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30, 5998–6008
189. Vig, J. and Belinkov, Y. (2019) Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Linzen, T. *et al.*, eds), pp. 63–76, Association for Computational Linguistics
190. Goyal, A. *et al.* (2022) *Coordination among neural modules through a shared global workspace*. Proceedings of ICLR
191. Kudugunta, S. *et al.* (2021) Beyond distillation: task-level mixture-of-experts for efficient inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (Moens, M.-F. *et al.*, eds), pp. 3577–3599, Association for Computational Linguistics
192. Zhou, Y. *et al.* (2022) Mixture-of-experts with expert choice routing. *Adv. Neural Inf. Proces. Syst.* 35, 7103–7114
193. Kaplan, J. *et al.* (2020) Scaling laws for neural language models. *arXiv*, Published online January 23, 2020. https://doi.org/10.48550/arXiv.2001.08361

194. Sutskever, I. *et al.* (2011) Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1017–1024
195. Radford, A. *et al.* (2019) Language models are unsupervised multitask learners. *OpenAI Blog* 1, 2019
196. McCoy, T. *et al.* (2019) Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Korhonen, A. *et al.*, eds), pp. 3428–3448, Association for Computational Linguistics
197. McCoy, R.T. *et al.* (2023) Embers of autoregression: understanding large language models through the problem they are trained to solve. *arXiv*, Published online September 24, 2023. https://doi.org/10.48550/arXiv.2309.13638
198. Kassner, N. and Schütze, H. (2020) Negated and misprimed probes for pretrained language models: birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D. *et al.*, eds), pp. 7811–7818, Association for Computational Linguistics
199. Warstadt, A. and Bowman, S.R. (2022) What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language* (Lappin, S. and Bernardy, J.-P., eds), pp. 17–60, CRC Press
200. van Schijndel, M. *et al.* (2019) Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Inui, K. *et al.*, eds), pp. 5831–5837, Association for Computational Linguistics
201. McCoy, R.T. *et al.* (2020) Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Trans. Assoc. Comput. Linguist.* 8, 125–140
202. Yedetore, A. *et al.* (2023) How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Rogers, A. *et al.*, eds), pp. 9370–9393, Association for Computational Linguistics
203. Georges Gabriel Charpentier, L. and Samuel, D. (2023) Not all layers are equally as important: every layer counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (Warstadt, A. *et al.*, eds), pp. 238–252, Association for Computational Linguistics
204. Hosseini, E.A. *et al.* (2024) Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiol. Lang.*, Published online January 19, 2024. https://doi.org/10.1162/nol-a-00137
205. Blasi, D. *et al.* (2022) Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (Muresan, S. *et al.*, eds), pp. 5486–5505, Association for Computational Linguistics
206. Mielke, S.J. *et al.* (2019) What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Korhonen, A. *et al.*, eds), pp. 4975–4989, Association for Computational Linguistics
207. Martin, L. *et al.* (2020) CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D. *et al.*, eds), pp. 7203–7219, Association for Computational Linguistics
208. Wang, Z. *et al.* (2020) Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Cohn, T. *et al.*, eds), pp. 2649–2656, Association for Computational Linguistics
209. Tirumala, K. *et al.* (2022) Memorization without overfitting: analyzing the training dynamics of large language models. *Adv. Neural Inf. Proces. Syst.* 35, 38274–38290
210. Sakaguchi, K. *et al.* (2020) Winogrande: an adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence* (34), pp. 8732–8740
211. Elazar, Y. *et al.* (2021) Back to square one: artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Moens, M.-F. *et al.*, eds), pp. 10486–10500, Association for Computational Linguistics
212. Yiu, E. *et al.* (2023) Transmission versus truth, imitation versus innovation: What what children can do that large language and language-and-vision models cannot (yet). *Perspect. Psychol. Sci.*, Published online October 26, 2023. https://doi.org/10.1177/17456916231201401
213. Lederman, H. and Mahowald, K. (2024) Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of LLMs. *arXiv*, Published online January 10, 2024. https://doi.org/10.48550/arXiv,2401.04854
214. Mitchell, M. and Krakauer, D.C. (2023) The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci.* 120, e2215907120
215. Pavlick, E. (2023) Symbols and grounding in large language models. *Phil. Trans. R. Soc. A* 381, 20220041
216. Mollo, D.C. and Millière, R. (2023) The vector grounding problem. *arXiv*, Published online April 4, 2023. https://doi.org/10.48550/arXiv.2304.01481