DISSOCIATING LANGUAGE AND THOUGHT IN LARGE LANGUAGE MODELS

A PREPRINT

Kyle Mahowald*
The University of Texas at Austin
mahowald@utexas.edu

Anna A. Ivanova*
Massachusetts Institute of Technology
annaiv@mit.edu

Idan A. Blank
University of California Los Angeles
iblank@psych.ucla.edu

Nancy Kanwisher
Massachusetts Institute of Technology
ngk@mit.edu

Joshua B. Tenenbaum
Massachusetts Institute of Technology
jbt@mit.edu

Evelina Fedorenko
Massachusetts Institute of Technology
evelina9@mit.edu

November 7, 2023

ABSTRACT

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, showing 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. In short, LLMs are good models of language but incomplete models of human thought.

* *The two lead authors contributed equally to this work.*

1 Introduction

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 (Turing, 1950). The Turing test uses language as an interface between two agents, allowing human participants to probe the knowledge and reasoning capacities of two other agents to determine which of them is a human and which is a machine.[1] Although the utility of the Turing test has since been questioned, it has undoubtedly shaped the way society today thinks of machine intelligence (Boneh et al., 2019; French, 1990, 2000; Marcus et al., 2016; Moor, 1976; Pinar Saygin et al., 2000).

[1] In later versions of the test, the number of conversation partners has been reduced to one.

The popularity of the Turing test, combined with the language-thought coupling in everyday life, has led to several common fallacies related to the language-thought relationship. We focus on two of these.

The first 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 long 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; Bommasani et al., 2021; Chang and Bergen, 2023; Devlin et al., 2019; Vaswani et al., 2017), most notably OpenAI's GPT models (Brown et al., 2020; Radford et al., 2019) and their user interface known as ChatGPT (OpenAI, 2022). LLMs today can produce text that is difficult to distinguish from human output, outperform humans at some text comprehension tasks (Srivastava et al., 2022; Wang et al., 2018, 2019), and show superhuman performance on next-word prediction (Oh and Schuler, 2023). As a result, claims have emerged—both in the popular press and in the academic literature—that LLMs represent not only a major advance in language processing but are showing "sparks of artificial general intelligence" (Bubeck et al., 2023).

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 language models (e.g., the chatbot Eliza from the 1960s; Weizenbaum, 1966).[2].

The second 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 (e.g. Elazar et al., 2021), lack of commonsense reasoning abilities (e.g., the ability to predict the effects of gravity Marcus, 2020), and failure to understand what an utterance is really about (e.g., Bender and Koller, 2020; Bisk et al., 2020). While these efforts to probe model limitations are useful in identifying things that LLMs can't do, 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 reduce 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 the world). Solving the problem of formal linguistic competence (e.g., what counts as a valid string in a language) is far from trivial and indeed has been a major goal of modern linguistics. That said, language does not exist in a vacuum and is fundamentally embedded and social, so the formal capacity is of limited value without its situated context (e.g., Bucholtz and Hall, 2005; Clark, 1992, 1996; Grice, 1975; Hudley et al., 2020; Labov, 1978; Lakoff, 1972; Wittgenstein, 1953). Thus, both formal and functional linguistic competence are essential components of human language use.

Our motivation for the formal/functional distinction comes from the human brain. A wealth of evidence from cognitive science and neuroscience has established that language and thought in humans are robustly dissociable: the mechanisms dedicated to processing language are separate from the mechanisms responsible for reasoning, memory, and social skills (Section 2).

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 LLMs starting with GPT-3, their formal competence is essentially at ceiling, whereas the functional competence of contemporary LLMs remains patchy, with results depending drastically 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 available for a given language, functional linguistic competence improvements with scale are less impressive, so much so that, today, LLM developers have shifted away from simple scaling up techniques to specialized fine-tuning on tasks of interest.

The success of LLMs is a major development, with far-reaching implications. Their ability to learn the rules and patterns of language using a simple word prediction objective goes substantially beyond what many researchers would have predicted even 10 years ago (e.g., Everaert et al., 2015). But LLMs' success in mastering linguistic knowledge by predicting words using massive amounts of text does not guarantee that all aspects of thought and reasoning could

[2]Note that people also make a related fallacy, "bad at language -> bad at thought" (Mahowald & Ivanova, 2022). Individuals who are not native speakers of a language, who do not speak hegemonic dialects, or those suffering from disfluencies in their productions due to developmental or acquired speech and language disorders are often incorrectly perceived to be less smart and less educated (Hudley and Mallinson, 2015; Kinzler, 2021; Kinzler et al., 2009)

be learned that way. Understanding what prediction-based LLMs are trained to do and the specific problems they are trained, or not trained, to solve is crucial to understanding their abilities (McCoy et al., 2023).

Since 2023, "pure" LLMs trained on word-in-context prediction are commonly enhanced by additional fine-tuning and/or combined with specialized systems downstream. InstructGPT (Ouyang et al., 2022) and ChatGPT are examples of fine-tuning enhancements: there, a pre-trained LLM is fine-tuned with Reinforcement Learning from Human Feedback (RLHF; Christiano et al., 2017). GPT Plugins[3] and ToolFormers (Schick et al., 2023) are examples of the so-called "augmented language models" (Mialon et al., 2023), which aim to flexibly integrate API calls from an LLM with other specialized systems. There are also now multimodal models (like GPT-4 and earlier systems that pair images and captions like CLIP; Radford et al., 2021). These models are more than just LLMs and can learn based on more than just what is available in massive amounts of extant text.

Here, we posit that the next-word prediction objective allows a model to master formal linguistic competence but not functional linguistic competence. As a result, various attempts at LLM enhancements via fine-tuning or coupling with additional modules often aim to address specific gaps in the models' functional competence. These methods (e.g., the Reinforcement Learning from Human Feedback: RLHF) have led to meaningful gains in functional competence.

In the rest of the paper, we develop a framework for evaluating the competence of modern language models from a cognitive science perspective. In Section 2, we elaborate on the constructs of formal and functional linguistic competence and motivate this distinction based on the evidence from human cognitive science and neuroscience. In Section 3, 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. Then, in Section 4, we consider several domains required for functional linguistic competence—formal reasoning, world knowledge, situation modeling, and social-cognitive abilities—on which today's LLMs often fail, or at least perform much worse than humans. In Section 5, we discuss the implications of our framework for building and evaluating future models of language and Artificial General Intelligence (AGI). Finally, in Section 6, we summarize our key conclusions.

2 Formal vs. functional linguistic competence

2.1 What does linguistic competence entail?

2.1.1 Formal linguistic competence

We define formal linguistic competence as a set of capacities required to produce and comprehend a given language. Specifically, it involves the knowledge of and flexible use of linguistic rules (e.g., Chomsky, 1957; Comrie, 1989; Pinker and Jackendoff, 2005), as well as of non-rule-like statistical regularities that govern that language (e.g., Bybee and Hopper, 2001; Goldberg, 2019; Jackendoff and Pinker, 2005). Well-recognized aspects of formal competence entail knowing a language's vocabulary and how it can be productively composed to form grammatical utterances. For example, 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 kinds of regularities that characterize 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 five days in New York" (Dalrymple and King, 2019; Keenan, 2013; Solt, 2007).

Human language users likely learn rules along with thousands of idiosyncratic constructions (Goldberg, 2019) through some combination of sophisticated statistical learning (Aslin, 2007; Aslin et al., 1998; Bresnan, 2007; Bybee and Hopper, 2001; Chater et al., 2006; Clark, 2014; Frank and Tenenbaum, 2011; Gerken, 2006; O'Donnell, 2011; Perfors et al., 2011; Saffran and Thiessen, 2003; Saffran et al., 1996; Spelke, 2004) and innate conceptual, and perhaps specifically linguistic, machinery (Berwick et al., 2011; Chomsky, 1957; Gleitman, 1993; Jackendoff and Jackendoff, 2002; Pinker, 2000; Pinker and Bloom, 1990; Pinker and Jackendoff, 2005). The result is the human ability to understand and produce grammatical and coherent linguistic utterances: "The customer ate." but not "The customer devoured.", "a beautiful five days in New York" and not "a beautiful five day in New York".

2.1.2 Functional linguistic competence

In addition to being competent in the rules and statistical regularities of language, a competent language user must be able to use language to do things in the world (Bloom, 2002; Bucholtz and Hall, 2004; Christiansen and Chater, 2016;

[3] https://openai.com/blog/chatgpt-plugins

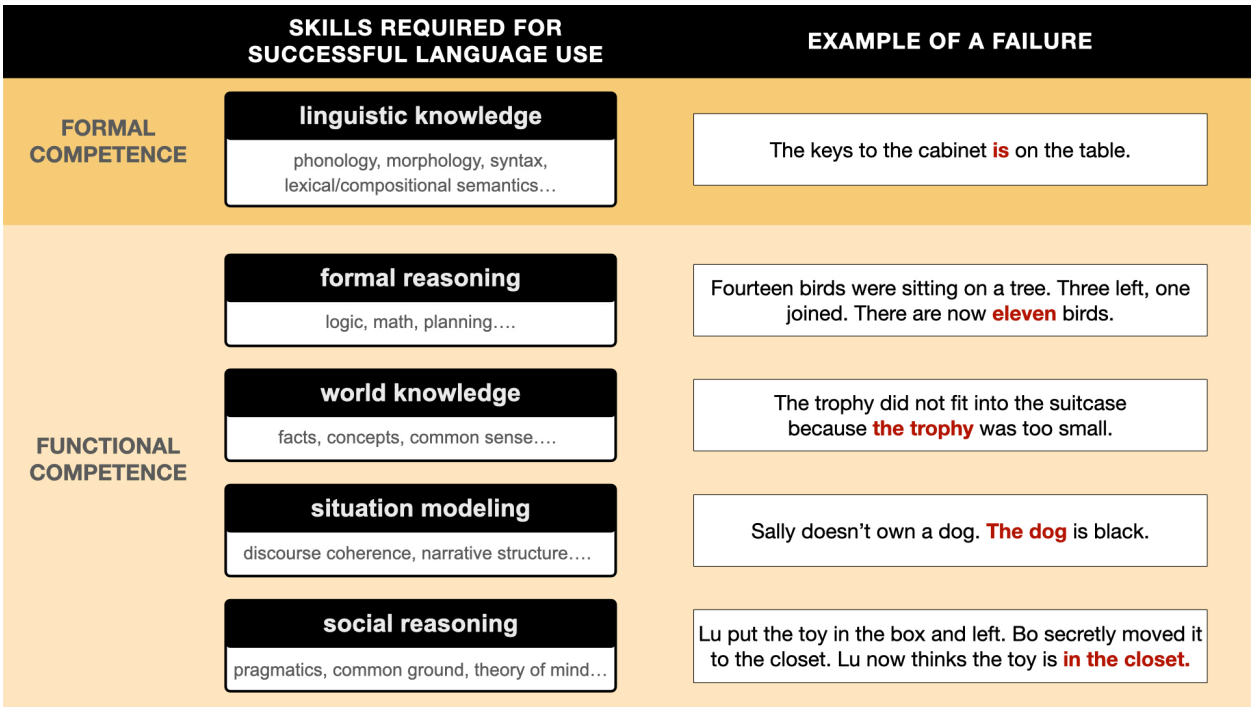

Figure 1: Successful use of language relies on multiple cognitive skills, some of which (required for formal competence) are language-specific and some (required for functional competence) are not. A failure to acquire a particular skill would result in a specific type of language use deficit. Determining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

Clark, 1996; Frank and Goodman, 2012; Grice, 1969, 1975; Slobin, 1996; Tomasello, 2010; Wilson and Sperber, 2002): to talk about things that can be seen or felt or heard, to explicitly reason about diverse topics, to make requests, to perform speech acts, to cajole, prevaricate, and flatter. People constantly use language to send and receive information 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 to a language user 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 aspects of 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.

2.2 Motivation for the distinction between formal vs. functional linguistic competence

As noted in Section 1, our motivation for the distinction between formal and functional linguistic competence comes from what we know about the functional 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. Below we briefly summarize a body of evidence from cognitive science and neuroscience that supports this dissociation.

2.2.1 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; e.g., Deniz et al., 2019; Fedorenko et al., 2010; MacSweeney et al., 2002; Regev et al., 2013; Scott et al., 2017) and production (Hu et al., 2022; Menenti et al., 2011); is sensitive to linguistic regularities at all levels: from phonological/sub-lexical, to word level, to phrase/sentence level (Bautista and Wilson, 2016; Blank and Fedorenko, 2020; Blank et al., 2016; Fedorenko et al., 2011, 2012, 2020; Regev et al., 2021); and supports linguistic operations that are related to both the processing of word meanings and those related to combinatorial semantic and syntactic processing (Fedorenko et al., 2020; Hu et al., 2022). Damage to the language network leads to linguistic deficits (e.g., Bates et al., 2003; Broca,

1865; Damasio, 1992; Mesulam, 2001; Mesulam et al., 2014; Saffran, 2000; Wernicke, 1874; Wilson et al., 2019). This tight link between the language network and language function indicates that these brain regions are responsible for language processing in humans.

2.2.2 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: a) functional brain imaging studies of neurotypical adults, and b) behavioral investigations of individuals with aphasia—a language impairment caused by damage to the language network, typically as a result of a stroke or degeneration.

Brain imaging techniques like functional MRI (fMRI) can be used to observe activity in the language network in healthy individuals in real time. Given its high spatial resolution, fMRI is especially 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 robustly and reliably when people listen to, read, or generate sentences (Section 2.2.1), but not when they perform arithmetic tasks, engage in logical reasoning, understand computer programs, listen to music, categorize objects or events, watch others' actions, reason about people's mental states, or process non-verbal communicative information like facial expressions or gestures (e.g., Amalric and Dehaene, 2019; Benn et al., 2021; Blank et al., 2014; Chen et al., 2021; Deen et al., 2015; Fedorenko et al., 2011; Ivanova et al., 2020; Jouravlev et al., 2019; Liu et al., 2020; Monti et al., 2007, 2009, 2012; Paunov et al., 2019, 2022; Pritchett et al., 2018; Shain et al., 2022).

Studies of individuals with aphasia provide a unique opportunity for testing which cognitive capacities rely on linguistic representations. Of particular interest are cases of so-called 'global aphasia', which affects both production and comprehension. Individuals with global aphasia exhibit severe linguistic deficits that, at best, 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, some individuals with severe aphasia have intact non-linguistic cognitive abilities: they can play chess, compose music, solve arithmetic problems and logic puzzles, leverage their world knowledge to perform diverse tasks, reason about cause and effect, and navigate complex social situations (Basso and Capitani, 1985; Bek et al., 2010; Klessinger et al., 2007; Lecours and Joanette, 1980; Luria et al., 1965; Varley, 1998; Varley and Siegal, 2000; Varley et al., 2001, 2005; Willems et al., 2011, see Fedorenko and Varley, 2016, for a review).

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, it is important to separate their linguistic abilities from their abstract knowledge and reasoning abilities, which can be probed—and perhaps even learned—through a linguistic interface but require much more than formal linguistic competence.

3 LLMs have mastered formal linguistic competence

Chomsky said in a 2019 interview (Lex Fridman, 2019): "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." The view that deep learning models are not of scientific interest remains common in linguistics and psycholinguistics, and, despite many arguments for integrating such models into research on human language processing and acquisition (Blank, 2023; Jain et al., 2023; Lappin, 2021; Linzen, 2019; Pater, 2019; Warstadt and Bowman, 2022) and an increasing chorus of claims that they should be taken seriously as linguistic and cognitive models (Baroni, 2021; Cao and Yamins, 2021; Frank, 2023; Piantadosi, 2023), 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 that is supported by the language-selective network in the human brain. We argue that these models are surprisingly and impressively successful at mastering this specific domain—dramatically more successful than the best systems from 10-15 years ago.

3.1 Statistical language models: some history and fundamentals

LLMs have arisen from a number of 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 trained on a word prediction task (the same task used for training n-gram models going back at least to Shannon's work in the mid-20th century; see Jurafsky and Martin (2009) and Lee (2003) for a historical overview). Similar to approaches in distributional semantics and word embeddings (for overviews, see Baroni and Lenci, 2010; Erk, 2012; Lenci, 2008)), LLMs represent linguistic information as vectors in a high-dimensional space. And, similar to earlier connectionist approaches (e.g., Elman, 1990, 1993; Rumelhart and McClelland, 1986, 1987), they use neural networks that are modeled on the human brain, whereby a series of model weights are learned and passed through a network in order to generate a response. All of these approaches stand in contrast to models that use explicit, structured hierarchical representations of syntactic rules (see Norvig, 2012, 2017 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; Jurafsky and Martin (2009); Lee (2003)). 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., Pinker and Prince, 1988). Everaert et al. (2015) specifically claim that statistical approaches, which use linear strings of words as input, are unable to learn complex syntactic features that require representing phrases and sentences hierarchically rather than linearly. This pessimism is now challenged by LLMs.

Here, we focus on a class of LLMs known as transformers—and their augmented descendants like ChatGPT. How do these models work? First, a training set is constructed from a massive amount of text from the web. During training, LLMs have a simple objective: predict the next token based on a fixed number of previous tokens (typically between 500 and a few thousand).[4] The predicted word piece is then compared with the ground truth (which word piece actually occurred in that training sentence), and the feedback signal is propagated back through the model to update its many (>100 billion) parameters.

Although it is tempting to move the goalposts and focus on what these models are still unable to do (see Bowman, 2022 for a discussion of the dangers of focusing on failures in NLP), we argue that the remarkable advances in LLMs' ability to capture various linguistic phenomena should not be overlooked. We argue that a large portion of formal linguistic competence arises as early as models of the scale of GPT-2 or BERT and plateaus in contemporary LLMs (Box 1). For the rest of this section, we focus on the grammaticality and coherence of LLM productions.

Box 1. The path toward formal linguistic competence

When did LLMs achieve formal linguistic competence? Table 1 shows text generations from an n-gram model, a state-of-the-art 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.

The output of GPT-2 (a model produced just 8 years after the RNN in Row 2) is grammatical: 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 to 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.

3.2 Large language models 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

[4] A token is the basic unit in LLMs. For most models of languages with alphabetic scripts, they end up corresponding to a mixture of words and morphemes.

Table 1: Examples of texts generated by increasingly more advanced language models in response to a prompt "The meaning of life is..." (following Sutskever et al. (2011) and using their example verbatim). Starting with GPT-2, outputs are grammatical and demonstrate proper use of 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 (Sutskever et al., 2011)	***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 (Radford et al., 2019)	***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 (Brown et al., 2020)	***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.

2 for a distinction between "linguistic" and conceptual semantics). Although important and interesting differences exist between linguistic processing in LLMs and humans (Lenci, 2023; Van Schijndel and Linzen, 2021), there are also important similarities. Here, we review evidence that LLMs succeed as models of formal linguistic competence (see Ettinger, 2020; Hupkes et al., 2020; Linzen and Baroni, 2021; Lovering and Pavlick, 2022; Manning et al., 2020; Press et al., 2022, for further discussions of the importance of emergent linguistic competence).

3.2.1 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 NLP community (e.g., GLUE tasks Wang et al., 2018, 2019), but, critically for our purposes, on tests of linguistic competence.

The benchmark BLiMP (Warstadt et al., 2020), for instance, contains minimal pairs of grammatical vs. ungrammatical sentences across a diverse range of complex linguistic phenomena like filler-gap dependencies (Bert knew what many writers find vs. *Bert knew that many writers find) and negative polarity licensing (The truck has clearly tipped over. vs. *The truck has ever tipped over.) These examples are designed to be challenging. RoBERTa base, a relatively small transformer LLM, systematically assigns higher probability scores to grammatical than to ungrammatical sentences, achieving human-level performance on 6 out of 12 item types (Warstadt and Bowman, 2022). As of mid-2023, the best performing model evaluated on BLiMP achieved the performance of 84%, with human performance being 89% (Liang et al., 2022). Similarly impressive results are seen on other linguistic benchmarks like SyntaxGym (Gauthier et al., 2020).

3.2.2 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 (e.g., Adger, 2003; Bresnan, 1982; Chomsky, 1957, 1965; Frazier, 1979; Jackendoff and Jackendoff, 2002). For instance, in the phrase "the keys to the cabinet", words should be combined as follows: the first "the" combines with "keys", the second "the" combines with "cabinet", "the cabinet" combines with "to", and the resulting phrase "to the cabinet" combines with "the keys".

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 non-hierarchical 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 must 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 the nearby "cabinet". However, a model that learns the underlying hierarchical structure of English will have no trouble keeping track of the subject-verb dependency (Bernardy and Lappin, 2017; Finlayson et al., 2021; Gulordava et al., 2018; Kuncoro et al., 2018; Lakretz et al., 2019; Lasri et al., 2022; Linzen et al., 2016; Lu et al., 2020; Mueller et al., 2022).

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 words (Gulordava et al., 2018; Linzen and Baroni, 2021), although some of the earlier models can be distracted by frequency effects (such as differences in the frequency between the singular and plural forms; Wei et al., 2021; Yu et al., 2020). In a similar vein, LLMs can handle other constructions, like filler-gap dependencies (Wilcox et al., 2018) or negative polarity (Marvin and Linzen, 2018; Warstadt et al., 2019). Finally, studies that examine the internal geometry of the models' sentence representations (e.g., Hewitt and Manning, 2019), studies that causally intervene on models and test agreement (Ravfogel et al., 2021), and studies that turn on and off specific model "neurons" (e.g., Lakretz et al., 2019; Mueller et al., 2022) have provided mechanistic insights into how an LLM might represent hierarchical structure and establish non-local structural dependencies.

3.2.3 LLMs learn abstractions

Following Ambridge (2020), 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. Gulordava et al. (2018) gives the example that, 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 rule knowledge is by using semantically nonsensical sentences, like "The colorless green ideas I ate with the chair... (sleep/sleeps)". Testing on Italian, English, Hebrew, and Russian, Gulordava et al. (2018) found that models performed well even on these semantically anomalous sentences.

An even more stringent test for linguistic abstraction asks whether LLMs can apply morphosyntactic rules to novel words. For instance, Kim and Smolensky (2021) show 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. That is, given the example in the preceding sentence as input, the model should be able to know that "I went to a blick" is more probable than "I went to a dax" since blick was used as a noun. They conclude that BERT succeeds partially at this task: it does learn to generalize, but only after repeated examples (but see Kim et al., 2022; Misra et al., 2023, for ways in which the word itself affects compositional ability). More recent models, such as GPT-3, seem to be able to use a novel word appropriately right away, at least if prompted correctly (Brown et al., 2020; McCoy et al., 2023).

Further, a large body of work has tested for linguistic abstraction in LLMs using a method called probing (Belinkov, 2022; Conneau et al., 2017; Ettinger et al., 2016; Giulianelli et al., 2018; Lasri et al., 2022; Tenney et al., 2019). In this literature, a classifier is typically trained on top of internal model representations to ask whether an abstract category, such as part-of-speech or dependency role, can be recovered from the model. Tenney et al. (2019) argued that LLMs "rediscover the classical NLP pipeline," learning at various layers features like part-of-speech categories, parses, named entities, and semantic roles (although see de Vries et al., 2020; Niu et al., 2022). However, one important limitation of such probing studies is that, even if abstract categories can be decoded from model representations, a model might not necessarily be using this knowledge (a problem that can be at least somewhat ameloriated with models that causally intervene on representations, e.g., Belinkov, 2022; Elazar et al., 2021; Finlayson et al., 2021; Ravfogel et al., 2021; Tucker et al., 2022; Wu et al., 2022).

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., Bates and MacWhinney, 1989; Gibson and Pearlmutter, 1998; MacDonald et al., 1994; MacWhinney and MacWhinney, 1987; Rayner et al., 2006; Tanenhaus et al., 1995). Humans also rely, to varying extents, on memorizing previously seen input, as opposed to purely learning abstract rules (Ambridge, 2020; Bod, 2009; Bybee and Hopper, 2001; Goldberg, 2009, 2019; Langacker, 1988, 2010; O'Donnell, 2015). Thus, when evaluating formal competence in

LLMs, it is essential to directly compare their performance with that of humans. For instance, Lampinen et al., 2022 reexamined an earlier study by Lakretz et al., 2021, which showed syntactic agreement deficits in GPT-2, and showed that the cases where the model failed were, in fact, also challenging for humans.

Overall, LLMs clearly achieve at least some degree of abstraction. The degree of that abstraction remains a matter of debate, as it does for humans, but the fact that LLMs show evidence of representing hierarchical structure and abstract linguistic patterns suggests that powerful language models can learn linguistic rules and regularities from textual input.

3.2.4 LLMs learn constructions

Recent evidence suggests that LLMs learn syntactic *constructions* (Tayyar Madabushi et al., 2020; Tseng et al., 2022; Weissweiler et al., 2023). These constructions can be idiosyncratic, lexically sensitive, and relatively rare, such as "a beautiful five days in Austin" (Mahowald, 2023).

Potts (2023) shows that early versions of GPT-3 are sensitive 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"). Remarkably, models learn that this crossing of the finite clause boundary is grammatical even though such examples are vanishingly rare: Potts painstakingly finds only 58 examples out of *7 billion* sentences in a corpus. So even today's most massive models are unlikely to see very many of these constructions—and they certainly won't see them with a wide range of lexical items and contexts. Yet these models know that variants like this, which almost never appear in the corpus, are grammatical, whereas other kinds of strings that almost never appear are not. We agree with Potts that these results suggest that LLMs meaningfully learn something about the language's syntax.

Weissweiler et al. (2022) show that models are also sensitive to the form of the comparative correlative "the better the syntax, the better the semantics". However, this sensitivity does not mean that they are sensitive to the semantic implications of the construction. Indeed, Weissweiler et al. (2022) show 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 *functional competence*. In Section 4, we discuss functional competence in greater detail.

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 aspect of semantics, often associated with compositional and lexical semantics, concerns the way that meaning is derived from words and their combinations. We consider this 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 (e.g., Fedorenko et al., 2012, 2020). The behavior of LLMs with respect to lexical and combinatorial semantics resembles that of the language network (Section 3.3).

The second sense 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 Section 4.2). 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.

3.3 LLMs are predictive of activity in the human language network

As discussed in Section 2.2.1, language processing in humans relies on a dedicated brain network. As one might expect, the human language network exhibits the hallmarks of formal linguistic competence: it is sensitive to abstract hierarchical rules in isolated phrases and sentences (e.g., Ding et al., 2016; Fedorenko et al., 2010, 2016; Law and Pylkkänen, 2021; Nelson et al., 2017; Pallier et al., 2011; Shain et al., 2023; Snijders et al., 2009), in naturalistic narratives (e.g., Brennan et al., 2020; Heilbron et al., 2022; Reddy and Wehbe, 2021; Shain et al., 2020, 2022), and in syntactically well-formed but semantically empty ("jabberwocky") stimuli (e.g., Fedorenko and Varley, 2016;

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

Although a preponderance of evidence suggests that LLMs acquire formal linguistic competence, their behavior is not fully human-like. Below we consider three common criticisms of LLMs as models of human language processing.

Excessive reliance on statistical regularities

Part of what makes LLMs succeed on diverse tasks is the fact that they can pick up on statistical regularities to achieve good performance. As a result, the models can be "right for the wrong reason" (McCoy et al., 2019) and leverage certain features in the input that aren't the ones being tested (Chaves, 2020). For instance, adding noise or distracting information can degrade model performance on a variety of tasks (e.g. Belinkov and Bisk, 2017; Kassner and Schütze, 2020; Khayrallah and Koehn, 2018; Wallace et al., 2019). In some (but not all) of these cases, it is shown that such noise does not similarly affect humans. Other evidence suggests that LLMs can be misled by simple frequency effects, such as, in a task where it has to choose between a singular and plural form of a particular verb, always choosing the plural form of verbs for which the plural form is much more frequent (Chaves and Richter, 2021).
These findings lead to the question of whether LLMs simply store and regurgitate output. This does not appear to be the case: McCoy et al. (2023) explicitly investigated the extent to which GPT-2 output is retrieved from the training set and found that, although n-grams up to length 4 often appeared in the training set, GPT-2 generated mostly novel 5-grams and above. They also showed that the model routinely generates plausible novel words that do not appear in the training set. Thus, LLMs generate output based on a combination of word co-occurrence knowledge and abstract morphosyntactic rules.

Unrealistic amounts of training data

Most LLMs that achieve near-human performance are trained on vastly more data than a child is exposed to (Warstadt and Bowman, 2022). van Schijndel et al. (2019) have found that a model's training dataset would need to be unrealistically large in order to handle some constructions in a human-like way. Therefore, even if linguistic information is, in principle, learnable from statistical regularities in the input, in practice, human language learners likely rely on pre-existing biases in order to learn quickly from sparse and noisy input—biases that today's state-of-the-art models lack (McCoy et al., 2018, 2020; Yedetore et al., 2023). This difference in the amount of input that models vs. human language learners require is sometimes taken to imply that the resulting model representations will necessarily be fundamentally unlike human linguistic representations.
However, there is reason for optimism. Ongoing work is actively exploring the extent to which models that are trained on more realistic amounts and kinds of input and/or in otherwise more realistic ways can still learn critical aspects of language (Warstadt and Bowman, 2022). An ongoing competition, BabyLM, 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. Several studies have found that some syntactic generalizations can occur in BERT-like architectures that are trained with millions (as opposed to billions) of words (Hu et al., 2020; Kobzeva et al., 2023; Wilcox et al., 2022; Zhang et al., 2021). For example, the ELECTRA model uses a binary word prediction objective, instead of an unconstrained word prediction task, and achieves comparable performance to models trained on far more data (Clark et al., 2020). BabyBERTa, a model trained on 5 million words of child-directed speech (Huebner et al., 2021)—similar in scale to what a human child would encounter—learned some syntactic generalizations comparable to RoBERTa, a high-performing model that is trained on 30 billion words. And Hosseini et al. (2022) found that models trained on just millions of tokens already provide a good match to human neural responses during language processing. Performance of smaller language models is far from perfect; for example, BabyBERTa could not represent several aspects of grammar that human children find intuitive. Critically, however, improvements in language models—including the use of more cognitively-inspired architectures and learning algorithms—could lead to strong performance with orders of magnitude less training data than today's state-of-the-art LLMs (see Zhuang et al., 2021, for evidence for the success of a similar approach in vision models). In that vein, important questions are: what inductive biases are introduced by the model architectures, whether those biases resemble the ones that enable humans to learn language (McCoy et al., 2018, 2020; Ravfogel et al., 2019), and whether better architectures could better capture these biases and enable faster learning on less data.

Insufficient tests on languages other than English

Because LLMs are data-hungry, they only work well on languages for which vast corpora are available. For most human languages, this is not the case. More worryingly, the architectures themselves may be biased towards English and other European languages (Blasi et al., 2022): not all languages are equally easy to model given the existing infrastructure (Cotterell et al., 2018; Mielke et al., 2019; Ravfogel et al., 2018). That said, evidence is growing of strong performance in a variety of languages (Kobzeva et al., 2023; Martin et al., 2020; Pires et al., 2019; Wang et al., 2019), and successful transfer of models to low-resource languages (Wang et al., 2020). Nevertheless, we should 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 language or have an entirely different modality (e.g., signed languages; Yin et al. 2021).

Fedorenko et al., 2010; Humphries et al., 2006; Matchin and Wood, 2020; Matchin et al., 2017, 2019; Pallier et al., 2011; Shain et al., 2021). The language network is also sensitive to specific word co-occurrences (e.g., as evidenced by sensitivity to n-gram surprisal; Shain et al., 2020), indicating that it learns not only the rules, but also the patterns of language. The language network's selectivity for linguistic vs. 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 Cao and Yamins, 2021; Schrimpf et al., 2020; Yamins et al., 2014 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 (Section 2.2.1). Both systems then compose these unit-level representations into composite representations of phrases and sentences. Interestingly, neither system shows clear spatial segregation for syntactic and semantic processing (LLMs: e.g., Durrani et al., 2020; Huang et al., 2021; Tenney et al., 2019; brain: e.g., Dick et al., 2001; Fedorenko et al., 2020; Reddy and Wehbe, 2021, 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 (e.g., Caucheteux and King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021). 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 1:1. Indeed, LLMs learn large amounts of patterns from natural texts that would hardly qualify as formal competence, such as predicting newline characters (Michaud et al., 2023). Nevertheless, the fact that internal representations learned by contemporary LLMs contain sufficient information to predict responses to diverse pieces of text in the human language network indicates that at least some of LLMs representations are driven by the same factors as the language network.

3.4 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 at least some knowledge of hierarchical structure and abstract linguistic categories while successfully capturing 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 (although see Box 3 on their limitations).

LLMs have already overturned the claims about the fundamental impossibility of acquiring certain linguistic knowledge—including hierarchical structure and abstract categories—from the statistics of linguistic input alone (Piantadosi, 2023). If language modeling continues to improve (including learning from more realistic kinds and amounts of data; Box 2), this would allow testing more general versions of this "poverty of the stimulus" argument (Berwick et al., 2011; Chomsky, 1991; McCoy et al., 2018; Pullum and Scholz, 2002), including specific tests of which, if any, inductive biases are required to learn the rules and statistical regularities of human language. As such, LLMs have substantial value in the scientific study of language learning and processing.

4 Non-augmented LLMs fall short on functional linguistic competence

So far, we have argued that LLMs have largely mastered formal linguistic competence—effectively producing fluent language. We have also argued that, for most of human history, fluent language has implied fluent thought. Thus, when interacting with LLMs today, people naturally attribute substantial thinking abilities to these models simply by the virtue of the "good at language -> good at thought" fallacy. In this section, we ask: how good are contemporary LLMs at functional linguistic competence?

Real-life language use is impossible without non-linguistic cognitive skills. Understanding a sentence, reasoning about its implications, and deciding what to say all rely on cognitive capacities that go way beyond formal 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 reasoning, mathematical reasoning, relational reasoning, computational thinking, and novel problem solving; **ii) world knowledge**—knowledge of objects and their properties, actions, events, social agents, facts, and ideas; **iii) situation modeling**—the dynamic tracking of protagonists,

locations, and events as a narrative/conversation unfolds over time; and **iv) social reasoning**—understanding the social context of linguistic exchanges, including what knowledge is shared, or in 'common ground', what the mental states of conversation participants are, and pragmatic reasoning ability. A simple conversation typically requires the use of all four of these capacities, yet none of them are specific to language use.

For each domain, we first discuss its neural mechanisms in humans and then briefly discuss how well contemporary LLMs have mastered this domain. We conclude that, unlike formal competence, functional competence of LLMs is uneven across domains, often requiring specialized fine-tuning and/or lacking the robustness and generality of functional competence skills in humans. In Box 4, we highlight the importance of properly conducting LLM evaluations; 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 in case of functional competence (e.g., Bubeck et al., 2023).

Box 4. Important considerations for evaluating functional competence

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

A. Fine-tuning on the task

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. As discussed in Section 3, 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; Tirumala et al., 2022), so prompting them with the exact sentence they have encountered during training will give highly inflated performance metrics.

B. 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 often the case that a model might perform well on a particular benchmark by leveraging low-level co-occurrence patterns (e.g., "knowing" that the sky is blue because of a frequent co-occurrence of the words "sky" and "blue"), but as soon as these obvious patterns are removed, the model's performance might drop to chance levels.

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).

4.1 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 (e.g., Baldo et al., 2005, 2010; Carruthers, 2002; Dennett, 1994; Grigoroglou and Ganea, 2022; Hinzen, 2013). 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* (Duncan, 2010, 2013; Duncan et al., 2020), named so because they are active in response to many cognitively demanding tasks: logic (Coetzee and Monti, 2018; Monti et al., 2007, 2009), mathematical reasoning (Amalric and Dehaene, 2016, 2019; Fedorenko et al., 2013; Monti et al., 2012; Pinel and Dehaene, 2009), physical reasoning (Fischer et al., 2016; Pramod et al., 2022;

Schwettmann et al., 2019), and computer code comprehension (Ivanova et al., 2020; Liu et al., 2020). 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 (Gläscher et al., 2010; Woolgar et al., 2010, 2018). Importantly, the multiple demand network supports reasoning even when the task is presented linguistically (Amalric and Dehaene, 2016, 2019; Ivanova et al., 2020; Monti et al., 2012) — similar to how LLMs receive their prompts.

LLMs. Several studies have criticized the extent to which language models can engage in tasks that require formal reasoning, such as math problems expressed in words. Patel et al. (2021) showed that, although models can appear to solve math problems (e.g., the dataset in Miao et al., 2020), they actually rely on heuristics and fail on more complicated problems. Similarly, the creators of GPT-3 show that it performs well on two-digit addition and subtraction but not on more complex tasks, such as three-digit addition or two-digit multiplication (Brown et al., 2020). GPT-4 similarly shows good performance on small-digit but not large-digit mathematical operations (Dziri et al., 2023). Reasoning tests that break common co-occurrence patterns in the input or require multi-step operations also lead to model failure (Chowdhery et al., 2022; Talmor et al., 2020). The most commonly cited reason for these failures is the failure of artificial neural nets to generalize to patterns outside their training distribution (Lake and Baroni, 2018; Loula et al., 2018; Zhang et al., 2022). This generalization gap can be partially bridged by the so-called "chain of thought" approaches, whereby a model is prompted to generate intermediate computation steps before arriving at an answer (Wei et al., 2022). However, even these approaches do not lead to foolproof results (Dziri et al., 2023) and might apply only to a subset of logic problems (Prystawski and Goodman, 2023). Thus, more and more researchers pair LLMs with external modules that can carry out structural logical and mathematical computations, such as the Mathematica plugin [5] or a probabilistic reasoning engine (Wong et al., 2023).

Overall, evidence from LLMs 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.

4.2 World knowledge and commonsense reasoning

Language provides a wealth of knowledge about the world (i.e., semantic knowledge). Basic 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). Might language processing mechanisms then subserve the more general capacity for storing semantic content? Do LLMs that master formal linguistic competence concurrently master world knowledge?

Humans. Evidence from neuroscience shows a dissociation between linguistic and semantic 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 as pictures often remains intact (Benn et al., 2021; Ivanova et al., 2021; Varley and Siegal, 2000). On the other hand, individuals who suffer from semantic dementia (a neurodegenerative disorder affecting primarily anterior temporal lobes) struggle with tasks that rely on world knowledge (e.g., knowing that a zebra has stripes) regardless of whether the stimuli are presented verbally or as pictures (Patterson et al., 2007). Thus, linguistic and semantic knowledge and processing rely on distinct neural circuits.

LLMs. Language models acquire a wealth of world knowledge contained in word co-occurrence patterns (Grand et al., 2022), so much so that attempts have been made to use them as off-the-shelf knowledge bases (Petroni et al., 2019). And there is some evidence that, even without any visual input, they can learn representations of space (Patel and Pavlick, 2021) and color (Abdou et al., 2021). However, world knowledge contained in LLM representations suffers from several major shortcomings.

First, LLMs routinely generate factually false claims, 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, an empirical study by Liu et al. (2023) has shown that those citations are often inaccurate.

Second, LLM outputs are often inconsistent: the same prompt phrased in different ways can elicit different responses (Elazar et al., 2021; Ravichander et al., 2020). They can also get "distracted" by intervening information, e.g., an irrelevant claim inserted between a premise and a conclusion Misra et al. (2023).

The two issues above apply to both factual and commonsense knowledge, but some problems with knowledge assessments pertain specifically to one or the other type. Factual knowledge often requires updates; for instance, the answer to "Who is the current president of the US?" will change every few 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, especially because these edits should

[5] https://www.wolfram.com/wolfram-plugin-chatgpt/

affect some other bits of knowledge (e.g., that the previously current president is now the past president) but leave many other facts unaffected (Meng et al., 2022). And commonsense knowledge is often underrepresented in language corpora: people are much more likely to communicate new or unusual information rather than commonly known facts Gordon and Van Durme (2013). As a result, LLMs struggle when comparing sizes of objects (e.g., "a table is smaller than an airplane" Liu et al., 2022; Talmor et al., 2020), distinguishing likely and unlikely events (Kauf et al., 2022), or performing on general commonsense reasoning benchmarks once the statistical cues are controlled for Elazar et al. (2021).

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., Borgeaud et al., 2022; Guu et al., 2020) but have not yet reached dominance in the field, typically because of relatively low coverage of existing knowledge bases. But 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 and commonsense schemata (Chersoni et al., 2019; Cohen et al., 2023).

4.3 Situation modeling

People can easily follow the plot of a story that spans multiple chapters or, sometimes, multiple book volumes. We can also have a three-hour conversation with a friend, and the next day remember most of what was discussed. We accomplish these impressive feats not by having a dedicated memory slot for every word that we read or heard, but by creating, based on the linguistic information, a "situation model" — a relatively abstract mental model of entities, relations between them, and a sequence of states they had been in or events they had participated in (Van Dijk et al., 1983). 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 (e.g., Blank and Fedorenko, 2020; Jacoby and Fedorenko, 2020; Lerner et al., 2011 see also Yeshurun et al., 2017). Instead, integration of meaning over longer periods of time likely takes place within the so-called default network (e.g., Blank and Fedorenko, 2020; Ferstl and von Cramon, 2002; Kuperberg et al., 2000; Lerner et al., 2011; Simony et al., 2016, ; for a general review of the default network, see Buckner and DiNicola, 2019). Crucially, the default network tracks both linguistic and non-linguistic narratives (Baldassano et al., 2017, 2018), indicating that situation modeling is not a language-specific skill.[6]

LLMs. Situation modeling in LLMs faces two main challenges: (1) extracting information from many sentences in a row; (2) making use of incoming inputs to track entities and events.

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 can simultaneously attend to each word in that book is vastly inefficient (although see some attempts to overcome this issue, e.g. (Su et al., 2021)). 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 Moirangthem and Lee, 2020; Ruan et al., 2022).

However, 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 (Li et al., 2021, cf. Kim and Schuster, 2023), 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."; Schuster and Linzen, 2022). Thus, whether robust situation model building over shorter span of text is feasible using an LLM-only architecture remains a matter of debate (see also theory of mind studies in Section 4.4).

4.4 Social reasoning

"Water!"

[6]Recent work has suggested that the default network may consist of two distinct interdigitated sub-networks (Braga et al., 2019; Deen and Freiwald, 2022; DiNicola et al., 2020). One of these sub-networks appears to correspond to the Theory of Mind network (Saxe and Kanwisher, 2003) discussed in 4.4 below. The exact contributions of the other sub-network remain debated, with different proposals linking its functions to episodic projection (placing oneself into the past, when remembering things, or into the future, when imagining things; Buckner et al., 2008), scene construction and situation modeling (Hassabis and Maguire, 2009) or spatial cognition in general (Deen and Freiwald, 2022).

Wittgenstein famously used single-word utterances like this to show that linguistic meaning radically depends on context. Although this word's literal interpretation is simply a reference to a physical entity, 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 kind of grounded, context-dependent aspects of language are not just peripheral but a central part of human language production and understanding (Bloom, 2002; Christiansen and Chater, 2016; Clark, 1996; Frank and Goodman, 2012; Grice, 1969, 1975; Slobin, 1996; Tomasello, 2010; Wilson and Sperber, 2002).

The process of inferring the intended meaning of the utterance beyond its literal content is known as pragmatics (Levinson et al., 1983). This processes likely engage a variety of cognitive mechanisms (Andrés-Roqueta and Katsos, 2017; Floyd et al., 2023; Levinson, 2000; Paunov et al., 2022). Here, we focus on one core capacity required for pragmatics: social reasoning.

Humans. A wealth of neuroscientific evidence shows that the human brain has dedicated machinery for processing social information (Adolphs, 1999, 2009; Deen et al., 2015; Isik et al., 2017; Kanwisher et al., 1997; Lee Masson and Isik, 2021; Saxe, 2006; Tarhan and Konkle, 2020; Walbrin et al., 2018, e.g.,). Perhaps the most relevant to our current discussion is the theory of mind network, a set of brain regions that are engaged when their owner is attempting to infer somebody's mental state (Fletcher et al., 1995; Gallagher et al., 2000; Jacoby et al., 2016; Saxe and Kanwisher, 2003; Saxe and Powell, 2006; Saxe et al., 2006). 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 regardless of whether the actual stimuli are texts or movies (Jacoby et al., 2016; Paunov et al., 2022), and texts that require inferring the characters' intentions evoke greater activity than those that do not (Ferstl and von Cramon, 2002; Fletcher et al., 1995; Saxe and Powell, 2006). Second, the theory of mind network is engaged more strongly during nonliteral language comprehension, including phenomena like jokes, sarcasm, indirect speech, and conversational implicature (Feng et al., 2017, 2021; Hauptman et al., 2022; Jang et al., 2013; Spotorno et al., 2012; van Ackeren et al., 2012, see Hagoort and Levinson, 2014, for a review) — in other words, in situations where understanding the meaning of an utterance requires inferring the intentions of the speaker. Thus, successful language understanding and use relies on our broader, non-language-specific social inference skills.

LLMs. Recent versions of OpenAI's GPT models show a marked improvement in interpreting non-literal utterances, such as metaphors and polite deceits, suggesting that it can reach human or near-human performance on at least some pragmatic tasks (Hu et al., 2022). This improvement likely comes from the fact that these models are fine-tuned on such pragmatics-requiring scenarios, although the fine-tuning details for OpenAI's 2022-2023 models remain unknown as of late 2023. That said, LLMs exhibit uneven performance across pragmatic domains: their ability to interpret sarcasm or complete jokes was limited even as their metaphor comprehension abilities soared (Hu et al., 2022). Overall, at least some forms of pragmatic inference might be acquired via targeted fine-tuning.

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 (Section 4.3). 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 fine-tuned LLMs (GPT-3.5 and up) have mastered theory of mind tasks (Kosinski, 2023) 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 (Ullman, 2023). Several other studies have identified limitations in LLM performance on theory of mind tasks (Sap et al., 2022; Shapira et al., 2023; Trott et al., 2023, cf. Gandhi et al., 2023). One solution to overcome these limitations has been to augment an LLM with a symbolic tracker of entity states and character beliefs (Sclar et al., 2023), an approach that mirrors the separation between language and theory of mind processing in humans.

5 Toward models that use language like humans

In this paper, we have advanced the thesis that formal and functional linguistic competence are distinct capabilities, recruiting different machinery in the human brain. 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.

As discussed throughout Section 4, to overcome LLM limitations in specific functional competence domains, researchers have employed targeted fine-tuning techniques, as well as augmenting base LLMs with additional modules that support extra-linguistic skills. Both of these approaches hold promise, and equipping LLMs with functional linguistic

competence results in *modular* models that mimic 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 (e.g., Bottou and Gallinari, 1990; Ronco and Gawthrop, 1997). More recent examples include a transformer language model paired with a separate memory module (e.g., Borgeaud et al., 2022; d'Autume et al., 2019; Liu et al., 2022) or a model for visual question answering, which includes a language module, a vision module, and a reasoning module (e.g., Andreas et al., 2016; Hudson and Manning, 2019; Johnson et al., 2017; Mao et al., 2019; Yi et al., 2018). 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 (Schick et al., 2023), mathematical calculators (Cobbe et al., 2021), planners (Liu et al., 2023), and other kinds of modules that do specific structured operations.

Another approach in this vein uses LLMs as module that translates a natural language query into code, which can then be passed to a symbolic module, which then generates an answer. Wong et al. (2023) outline a program for this approach, showing that GPT-3 can translate text input into meaningful structured probabilistic programs, which can be used to reason over relational domains (like kinship systems), grounded domains (like visual scenes), and situations that require planning. Their approach demonstrates a promising avenue for integrating what LLMs succeed at (namely, the kind of formal linguistic competence we outline in this paper) 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 sub-components 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., Dobs et al., 2022; Yang et al., 2019), 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. Manning et al., 2020; Vaswani et al., 2017; Vig and Belinkov, 2019); certain approaches promote modularization even more explicitly, e.g., by endowing transformers with a mixture-of-experts architecture (Goyal et al., 2022; Kudugunta et al., 2021; Zhou et al., 2022).

A modular language model architecture is much better aligned with the fact that real-life language use is a complex capability that requires both language-specific knowledge (formal competence) and various non-language-specific cognitive abilities (functional competence). Whether built-in or emerging naturally, modularity can lead the models to mirror the functional organization of the human brain and, consequently, make their behavior more human-like.

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 (e.g., Gauthier et al., 2020; Warstadt et al., 2020) and can be complemented by additional tests of core linguistic features: hierarchy and abstraction (Section 3.2). Benchmarks for evaluating different domains of functional linguistic competence, like commonsense reasoning (e.g., WinoGrande from Sakaguchi et al., 2019; HellaSwag from Zellers et al., 2019), can often be "hacked" by LLMs by leveraging flawed heuristics (Elazar et al., 2021). This issue is likely exacerbated in large-scale heterogeneous datasets like BIG-bench (Srivastava et al., 2022). 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.

6 Conclusion

Over the last few years, the discourse around language models has consisted of a curious mix of overclaiming and underclaiming (Bowman, 2022). On the one hand, some have claimed that GPT models show "sparks" of general intelligence (Bubeck et al., 2023). On the other hand, a steady stream of articles within the academic literature 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 dialogue 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 specialized machinery in the human brain.

The many 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 artificial general intelligence 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).

To those who have argued that 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-linguistic capabilities lag behind. Finally, to those who are looking to language models as a route to AGI, we suggest that, instead of, or in addition to, continuously scaling up the models (Kaplan et al., 2020), 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 planning.

Acknowledgements

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 and his group members, 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.

Funding Sources

KM acknowledges funding from NSF Grant 2104995. AI was supported by funds from the Quest Initiative for Intelligence. EF was supported by 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.

Conflicts of Interest

The authors declare no Conflicts of Interest.

References

[1] Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.9. URL https://aclanthology.org/2021.conll-1.9.

[2] David Adger. *Core Syntax: A Minimalist Approach*. Oxford University Press Oxford, 2003.

[3] Ralph Adolphs. The Human Amygdala and Emotion. *The Neuroscientist*, 5(2):125–137, March 1999. ISSN 1073-8584. doi: 10.1177/107385849900500216. URL https://doi.org/10.1177/107385849900500216. Publisher: SAGE Publications Inc STM.

[4] Ralph Adolphs. The Social Brain: Neural Basis of Social Knowledge. *Annual review of psychology*, 60:693–716, 2009. ISSN 0066-4308. doi: 10.1146/annurev.psych.60.110707.163514. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2588649/.

[5] Marie Amalric and Stanislas Dehaene. Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences of the United States of America*, 113(18): 4909–4917, May 2016. ISSN 1091-6490. doi: 10.1073/pnas.1603205113.

[6] Marie Amalric and Stanislas Dehaene. A distinct cortical network for mathematical knowledge in the human brain. *NeuroImage*, 189:19–31, April 2019. ISSN 10538119. doi: 10.1016/j.neuroimage.2019.01.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811919300011.

[7] Ben Ambridge. Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6):509–559, 2020.

[8] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016.

[9] Clara Andrés-Roqueta and Napoleon Katsos. The Contribution of Grammar, Vocabulary and Theory of Mind in Pragmatic Language Competence in Children with Autistic Spectrum Disorders. *Frontiers in Psychology*, 8, 2017. ISSN 1664-1078. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00996.

[10] R.N. Aslin. What's in a look? *Developmental Science*, 10(1):48–53, 2007.

[11] R.N. Aslin, J.R. Saffran, and E.L. Newport. Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science*, 9(4):321–324, 1998.

[12] Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W. Pillow, Uri Hasson, and Kenneth A. Norman. Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, 95(3):709–721.e5, August 2017. ISSN 1097-4199. doi: 10.1016/j.neuron.2017.06.041.