Evidence from Formal Logical Reasoning Reveals that the Language of Thought is not Natural Language

Hope Kean[1,2], Alexander Fung[*,1,2], Paris Jaggers[*,3], Jason Chen[1], Joshua S. Rule[1,5], Yael Benn[4], Joshua B. Tenenbaum[1], Steven T. Piantadosi[5], Rosemary A. Varley[3], Evelina Fedorenko[1,2]

[1]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology
[2]McGovern Institute for Brain Research, Massachusetts Institute of Technology
[3]Department of Psychology and Language Sciences, University College London
[4]Department of Psychology, Manchester Metropolitan University
[5]Department of Psychology, UC Berkeley

Acknowledgements

We would like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, including the technical team—Steve Shannon and Atsushi Takahashi. We would also like to thank Anya Ivanova, Colton Casto, Ben Lipkin, Chiebuka Ohams, and Ted Gibson for helpful comments on the manuscript. HK was supported by a Friends of McGovern graduate fellowship and a graduate fellowship from the K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center. JBT was supported by research funds from the Air Force Office of Scientific Research, the Office of Naval Research Science of AI program, and a Schmidt Sciences AI2050 Fellowship. JSR and STP were supported by the Division of Research on Learning in Formal and Informal Settings (EDU/DRL) grant 2201843. RV was supported by a Leverhume Research Fellowship (RF-2023-690\10). EF was supported by research funds from the McGovern Institute for Brain Research, the Department of Brain and Cognitive Sciences, MIT's Quest for Intelligence, and a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT. We would like to dedicate this work to S.A., an extraordinary individual whose time, effort, and remarkable mind and brain have made invaluable contributions to neuroscience over the past three decades, since becoming aphasic in his mid-forties. We are indebted to him and his family.

Keywords
Language of Thought, Language, Logical Reasoning, Aphasia, fMRI, Cognitive Neuroscience

Abstract

Humans are endowed with a powerful capacity for both inductive and deductive logical thought: we easily form generalizations based on a few examples and draw conclusions from known premises. Humans also arguably have the most sophisticated communication system in the animal kingdom: natural language allows us to express complex and structured meanings. Some have therefore argued for a tight relationship between complex thought and language, postulating that reasoning, including logical reasoning, relies on linguistic representations. We systematically investigated the relationship between logical reasoning and language using two complementary approaches. First, we used non-invasive brain imaging (fMRI) to examine neural activity as healthy adults engaged in inductive and deductive logical reasoning tasks. And second, we behaviorally evaluated logical abilities in individuals with extensive lesions to the language brain areas and consequent severe linguistic impairment. Our findings reveal that the language system is not engaged during logical reasoning, and patients with severe aphasia exhibit intact performance on logic tasks. Instead, inductive reasoning recruits the domain-general multiple demand system implicated broadly in goal-directed behaviors, whereas deductive reasoning draws on brain regions that are distinct from both the language and the multiple demand systems. Together, these results indicate that linguistic representations are neither utilized nor required for inductive or deductive logical reasoning.

Significance

Which brain areas allow humans to reason logically, to understand whether a conclusion follows from the premises? Are they the same areas that allow the assembly of words into structured representations? Scholars have debated for millennia whether logical reasoning is inextricably tied to natural language, or instead relies on a distinct "language of thought" (LOT). Using fMRI in healthy adults and evaluating logical ability in individuals with severe aphasia, we find that distinct neural systems support language processing vs. logical (inductive and deductive) reasoning. These results establish that language does not underpin logical inference and point to distinct representational systems for the logical LOT. This work contributes to our understanding of the division of cognitive labor in the human brain.

Introduction

In 350 BC, Aristotle wrote that "writing and speech are not the same for all people, but mental acts themselves, of which words signify, are the same for all people" (tr. Cooke & Tredennick, 1938). This reflection highlights a crucial distinction between linguistic symbols and the underlying cognitive structures they aim to express: although natural languages differ across cultures, the capacity for abstract thought and logical reasoning appears to be ubiquitously present across human societies. Still, because all natural languages share critical features some have argued that natural language is the underlying medium of complex thought (e.g., Chomsky, 1965; Carruthers, 2002).

One prominent proposal—the "language of thought" (LOT) hypothesis (Fodor, 1975)—has emphasized the compositional and hierarchical nature of thoughts. According to this hypothesis, thoughts are composed of smaller atomic pieces in a structured format with hierarchical relations among the component elements, similar to how computer programs are built out of a small collection of primitive operations, or sentences are built out of words. Building on this similarity with language, some have then explicitly argued that the language of thought is natural language (e.g., Davidson, 1967, 1975; Dennett, 1991, 1996, 2017; Chomsky, 1993, 1995; Carruthers, 2002; for earlier claims, see Wittgenstein, 1921; for counter-arguments, see Pinker, 1994; Fodor, 1998). Many thoughts can certainly be cast into natural language: after all, an effective communication system must allow for the transmission of internal mental representations. However, people also express many ideas using non-linguistic symbols, mathematical expressions, visual schematics or diagrams, and so on. Therefore, a degree of isomorphism between certain thoughts and linguistic expressions need not entail that *specifically linguistic* representations are used for thinking, any more than the fact that language shares hierarchical structures with music entails that language *is* the substrate for music.

Indeed, empirical data have been accumulating that suggest that linguistic representations are neither utilized nor necessary for thinking. Patients with acquired linguistic impairments (aphasia) appear to engage in diverse forms of thought, as evidenced by their intact performance on tasks requiring mathematical reasoning (Varley et al., 2005), causal reasoning (Varley & Siegal, 2000), and Theory of Mind (Apperley et al., 2004, 2006). Also, the brain areas that support language comprehension and production (see Fedorenko et al., 2024 for a review) are not engaged during many cognitively demanding tasks. These tasks include solving arithmetic and algebraic problems (Fedorenko et al., 2011; Monti et al., 2012; Amalric et al., 2019), understanding computer code (Ivanova et al., 2020; Liu et al., 2020), engaging in social reasoning (Paunov et al., 2019, 2022; Shain, Paunov, Chen et al., 2023), or performing executive function tasks (Fedorenko et al., 2011; Malik-Moraleda, Ayyash et al., 2022; Hiersche et al., 2024), which require maintenance of information in working memory and inhibiting irrelevant distractors—skills intimately linked to fluid intelligence (Engle et al., 1999; Duncan, 2010).

However, the relationship between language and *logical reasoning* specifically has only received limited attention in the prior neuroscience literature, in spite of the fact that logic is a hallmark domain where the role of language has been emphasized (Kroger et al., 2008). One prior fMRI study has examined the relationship between language and deductive reasoning by comparing neural responses during a logic inference task (judging the validity of a conclusion given a

premise) versus a linguistic inference task (judging the validity of thematic roles across two sentences) relative to lower-level control conditions (judging expression well-formedness) (Monti et al., 2009). A partial dissociation was observed: although both the logic and the language contrast elicited left-lateralized responses in the frontal, temporal, and parietal cortex, some brain areas—in the vicinity of areas traditionally implicated in language processing (Geschwind, 1970; Goodglass, 1993; Friederici et al., 2002; Hagoort, 2019; Fedorenko et al., 2024)—showed a stronger response to linguistic inference, whereas other areas responded more strongly during logic inference (see Coetzee et al., 2022 for concordant evidence from TMS). However, recent work calls these findings into question. Research over the last decade has a) developed more robust ways of identifying language brain areas (Fedorenko et al., 2010; Lipkin et al., 2022), and b) established that the kind of linguistic paradigm employed by Monti et al. (2009) plausibly engages both language areas but also brain areas sensitive to general cognitive effort (e.g., Diachek, Blank, Siegelman et al., 2020), which makes previously observed dissociations more difficult to interpret. Another past study examined logical, as well as mathematical, reasoning and argued that neither type of reasoning engaged the language areas (Kroger et al., 2008), but that study did not include a language task, which is necessary to make such a claim. Thus, these prior findings are intriguing, but have not conclusively answered the question of whether logical reasoning engages linguistic processing mechanisms.

Two other bodies of work bear relevance to the logic-language relationship. The first comes from developmental psychology and suggests that the timelines of linguistic and logic development diverge. Inductive reasoning capacities appear to come online very early (e.g., Gopnik, 1982; Goddu & Gopnik, 2025). The representations are argued to be sensorimotor, not linguistic, in nature, and the main debate concerns their domain specificity (e.g., Carey, 2009; Gelman, 2003) vs. generality (e.g., Gopnik et al., 2004; Gopnik & Wellman, 2012) (e.g., do infants have distinct representations for reasoning about the physical world vs. about social agents?). For deductive reasoning, the picture is less clear: some have argued that one-year-old preverbal infants can already reason disjunctively (Cesana-Arlotti et al., 2018), which implies a separation of logical and linguistic ability. However, others have reported failures in simple disjunctive inferences as late as age 2.5 years (Mody & Carey, 2016). These late failures may suggest that a certain level of linguistic competence is required for the development of these capacities, but they are also consistent with the slow developmental trajectory of executive abilities (Tervo-Clemmens et al., 2023; Best & Miller, 2010; Diamond, 2013; Zelazo & Carlson, 2020), which may be needed to deal with the task demands in more complex paradigms.

The second body of work comes from artificial intelligence, where recent advances have provided additional motivation for investigating the relationship between language and logic. In particular, neural network language models (Devlin et al., 2019; Radford et al., 2018; Brown et al., 2020) not only achieve human-level performance on diverse language tasks (Hendrycks et al., 2021; Achiam et al., 2023; Rein et al., 2024), but also exhibit impressive successes on certain reasoning tasks (e.g., ARC: Chollet et al., 2024; BIG-Bench: Srivastava et al., 2022; LogiQA: Liu et al., 2020; MATH: Hendrycks et al., 2021; WildBench: Lin et al., 2024; Templates: Boix-Adserà et al., 2024). These findings beg the question of whether linguistic competence inevitably leads to logical ability, although many have pointed out that reasoning in large language models lacks robustness and generalizability (e.g., Mahowald, Ivanova, et al., 2024; Qiu et al., 2024; Shojaee, Mirzadeh et al., 2025; McCoy et al., 2023; Nezhurina et al., 2025).

To systematically examine the role of language in human logical reasoning, we adopted a two-pronged approach. We first used functional MRI in healthy adults to test whether the language network (Fedorenko et al., 2024) would be engaged during logical reasoning tasks (Study 1). Next, we used behavioral experiments in individuals with severe aphasia resulting from extensive damage to left peri-Sylvian cortex, to test whether linguistic ability is necessary for logical reasoning (Study 2). In each study, we examined two paradigmatic forms of logical reasoning: inductive and deductive reasoning. The ***inductive reasoning*** paradigm (**Figure 1**, left panel) was adapted from Rule et al. (2024; for earlier, related paradigms, see Bruner et al., 1956; Bongard, 1967; Gentner, 1983; Hofstadter & Mitchell, 1994; Tenenbaum, 1999). Participants were presented with an input number list and an output list (e.g., `[5, 7] → [7, 5, 7]`) and asked to infer the rule that governs the input-to-output transformation. They could then test their hypothesis on a new input list, and so on, until they guess the correct rule. The rules involve a combination of mathematical operations (e.g., $f(l) = [\, x + 2 \mid x \in l \,]$), "add 2 to every number"), list operations (e.g., $f(l) = [\, x_i \mid x_i \in l, \forall j < i, x_i \neq x_j \,]$, "leave only unique, non-repeated elements"), and structural operations (e.g., $f(l) = l[3:] \,^{\wedge}\, l[0:3]$, "rotate the list by three elements"), each of which can be written as short computer programs. The fMRI version also included a control condition where participants were told the rule and asked to apply it to new input lists.

For ***deductive reasoning***, we used three paradigms. The first was introduced in Coetzee & Monti (2018) and was used for the fMRI component of the study (**Figure 1**, center panel). Participants were presented with a classic syllogism consisting of two premises and a conclusion (e.g., i. If A is B, then A is also C. ii. A is not C. iii. Therefore, A is not B) and were asked to judge the validity of the conclusion. Half of the syllogisms used real words (e.g., "If the block is large..."), and the other half used nonwords (e.g., If the tep is ag…"). Although deductive reasoning was required for all problems, the problems varied in the difficulty of the necessary deduction. The easier problems, known as Modus Ponens, had the following forms: "If the block is large, then it is not yellow. The block is large. So, the block is not yellow." (the conclusion follows from the premises), or "If the block is large, then is it not yellow. The block is large. So, the block is yellow." (the conclusion does not follow from the premises). The more complex problems, known as Modus Tollens, had the following forms: "If the block is large, then it is not yellow. The block is yellow. So, the block is not large." (the conclusion follows from the premises), or "If the block is large then it is not yellow. The block is not yellow. So, the ball is large." (the conclusion does not follow from the premises). This is a classic paradigm in the study of logical reasoning (Wason, 1966; Johnson-Laird & Wason, 1970; Wason & Johnson-Laird, 1972; Johnson-Laird, 1975; Rips, 1983) and the contrast between the Modus Tollens versus Modus Ponens problems elegantly isolates deductive reasoning demands. However, it is not suitable for patients with aphasia because it uses verbal stimuli. As a result, the second, complementary paradigm relied on non-verbal matrix reasoning (Cattell, 1949; for prior use in fMRI, see Woolgar et al., 2013). Participants were presented with a matrix of four abstract geometric patterns and were asked to decide which of the four was an outlier (**Figure 1**, right panel). Patients with aphasia performed a similar version (Wechsler, 2011), except that a matrix of abstract geometric patterns was missing a pattern, and participants were asked to decide which option from a set of possible answers completes the matrix (**Figure 1**, bottom right panel). This type of paradigm, commonly used in assessing the fluid reasoning capacity across diverse populations (Cattell, 1940; Raven, 2000; Wechsler, 2008; Primi et al., 2010; Flanagan & Harrison, 2012; Weiss et al., 2016; Roth et al., 2015; Bediou et al., 2018,

2023; Zorowtiz et al., 2024), primarily engages deductive reasoning by requiring participants to evaluate logical constraints and eliminate incorrect options. However, participants may additionally employ inductive strategies to generate hypotheses about the rules that govern the given patterns, so this paradigm may tax both deductive and inductive reasoning.

If linguistic representations underlie logical inference, then a) the language brain areas should be engaged during the induction task, the deduction tasks, or both, as measured with fMRI; and b) patients with profound aphasia should be unable to perform these tasks. To foreshadow our findings, we do not find support for the idea that language is the medium of logical reasoning.

Results

The language network is not engaged during inductive nor deductive reasoning.

As expected based on a large body of prior work (e.g., see Fedorenko et al., 2024 for a review), the areas of the language network show a robust response to language processing, evidenced by a stronger response during the sentence condition compared to the nonword-list condition estimated in independent data ($p< 0.001$ at the network level; **Figure 2B**, **Table 1**). These areas have been established in past work to support computations related to retrieving word meanings from memory, syntactic-structure building, and semantic composition during both language comprehension and production (Fedorenko et al., 2010; Menenti et al., 2011; Pallier et al., 2011; Giglio et al., 2022; Shain, Kean et al., 2024). Especially important for the hypothesis that the language system mediates logical reasoning is the sensitivity of these brain areas to hierarchical linguistic structure (Just et al., 1996; Ben-Shachar et al., 2003; Bornkessel et al., 2005; Caplan et al., 2008; Pallier et al., 2011; Tyler et al., 2011; Blank et al., 2016; Ding et al., 2016; Heilbron et al., 2022; Shain et al., 2022).

Critically, the language-processing brain areas show little or no response to the inductive and deductive reasoning contrasts (**Figure 2A-B**, **Table 1**). The two deductive contrasts elicit no significant response (ps>0.1 at the network level; see **SI Figure 1a** and **SI Table 1** for evidence that the results are similar for the five language areas separately), although as expected, the responses are strong to both conditions of the verbal deductive paradigm (**SI Figure 1b**). The inductive contrast elicits a reliable response at the network level ($p<0.05$), but the effect is small, with the language contrast being over four times stronger, and the critical induction condition eliciting a response that is at or below the level of the control condition of the language task (reading nonword lists) (**SI Figure 1b**).

Patients with profound aphasia can nevertheless reason logically.

Two individuals with profound aphasia, S.A. and G.S., showed preserved logical reasoning abilities. Both individuals have sustained extensive damage to their left peri-Sylvian cortex (**Figure 2C**) and, consequently, exhibit severe linguistic impairments (**Figure 2D**, **SI Table 3**). In particular, critical to the hypothesis that linguistic syntactic structures mediate logical reasoning, these patients show severe grammatical impairments in both comprehension and production, with at or near-chance performance on multiple syntactic assessments. For example, both exhibited

marked impairment in assigning thematic roles in understanding reversible sentences, such as *The diver splashed the dolphin*. Nevertheless, despite their profound aphasia, both patients show typical-like performance on the inductive and the deductive reasoning tasks. In the inductive reasoning task (**Figure 1**, left panel), they solved the vast majority of the rules presented to them (19/25 and 39/40, respectively). This performance level is comparable to a normative sample of control participants; neither patient significantly differs from the control sample (Crawford-Howell single-case test *p*s>0.499). Similarly, in the deductive matrix reasoning task, both patients solved the vast majority of the problems (25/30 and 26/30), which puts them +2.3 and +1.8 standard deviations above the mean based on the age-matched normative data available for this task (WASI-II; Wechsler, 2011) (these raw scores correspond to *T*-scores of 73 and 68; normative $T = 50$, SD = 10).

Inductive and deductive reasoning elicit strong responses outside the language network.

Given that our logic tasks did not engage the language system, we wanted to ensure that they elicit a response somewhere in the brain. We first examined neural responses in one plausible candidate system—the Multiple Demand (MD) network (Duncan & Owen, 2000; Duncan, 2010; Duncan et al., 2020). This network comprises a set of bilateral frontal and parietal brain areas that show strong activity during diverse cognitively demanding tasks, such as standard executive function tasks (Duncan & Owen, 2001; Niendam et al., 2012; Hugdahl et al., 2015; Shashidhara et al., 2019; Assem et al., 2020a). Damage to the MD network is associated with decreases in fluid intelligence (Woolgar et al., 2010, 2018). In addition, and most relevantly, certain forms of reasoning appear to rely on the MD network, including mathematical reasoning (Monti et al., 2012; Fedorenko et al., 2013; Amalric et al., 2019) and the kind of reasoning necessary to understand computer code (Ivanova et al., 2020; Liu et al., 2020). Replicating much prior work (e.g., Fedorenko et al., 2013; Assem et al., 2020b), the areas of the MD network show a robust response to a spatial working memory task, including a stronger response during the more demanding condition ($p<0.001$ at the network level; **SI Figure 2** and **SI Table 2**). The inductive reasoning task elicited a robust response across the MD network, evidenced by a stronger response during the rule induction condition compared to the control, rule application condition ($p<0.001$ at the network level; **SI Figure 2**, **SI Table 2**). However, the deductive reasoning task did not engage the MD network ($p>.48$ at the network level). This finding aligns with Coetzee & Monti (2018), who reported a dissociation between brain areas sensitive to deductive reasoning and those sensitive to general task difficulty—a key signature of the MD network (see also Kroger et al., 2008 for concordant data). A whole-brain search revealed several frontal and parietal areas that showed a stronger response to the Modus Tollens condition compared to the Modus Ponens condition estimated in independent data ($p<0.001$ across the set of areas; **SI Figure 2**).

Discussion

To shed light on the long-standing debate on the role of linguistic syntactic representations in formal logical reasoning, we examined the responses of the language brain areas—robustly sensitive to syntactic structure (Pallier et al., 2011; Giglio et al., 2022; Shain et al., 2022; Shain, Kean et al., 2024)—to inductive and deductive reasoning. In a complementary approach, we examined logical reasoning abilities in individuals with severe damage to the language areas. The

results converged on a clear answer: linguistic representations are not engaged nor necessary for logical reasoning. Below we contextualize these findings with respect to the broader literature.

Accumulating evidence for language selectivity. Our results showing that the left-lateralized fronto-temporal language network is not engaged in logical induction nor deduction adds to a growing body of evidence for the selectivity of the language network for linguistic computations. In particular, prior studies have evaluated diverse hypotheses about overlap between linguistic processing and different perceptual and cognitive tasks. Some have focused on the social-communicative function of language and argued for overlap with other social functions (Grice, 1968, 1975; Sperber & Wilson, 1986), but studies have shown that non-verbal communicative signals, such as facial expressions and gestures, are processed in brain areas distinct from the language network (e.g., Deen et al., 2015; Pritchett et al., 2018; Jouravlev et al., 2019). Others emphasized the hierarchical structure of language and argued for overlap with the processing of other structured stimuli, such as music (Patel, 2003; Fitch & Martins, 2014), but studies have shown that music perception tasks do not engage the language areas (Fedorenko et al., 2011; Rogalsky et al., 2011; Chen et al., 2023). Yet others have argued that language processing requires reliance on domain-general executive resources (Thompson-Schill, 2005; Kaan & Swaab, 2002; Novick et al., 2005, 2014; January et al., 2009), but executive tasks do not engage the language system (Fedorenko et al., 2011; Malik-Moraleda, Ayyash et al., 2022; Hiersche et al., 2024), and demanding linguistic computations are processed within the language network (Blank et al., 2016; Shain et al., 2020, 2022; Quillen et al., 2021; Wehbe et al., 2021; see Fedorenko & Shain, 2021 for a review). Finally, and of greatest relevance to the current investigation, some hypotheses have highlighted the similarity between linguistic structure and structure in some domains of abstract reasoning, such as mathematical or logical thinking (Chomsky, 1957; Marcus, 2001; Carruthers, 2002), or even reasoning in particular domains, such as intuitive physics (McCarthy & Hayes, 1969; Kowalski & Sergot, 1986; Pinto & Reiter, 1993) or social reasoning (de Villiers & de Villiers, 2000). A number of studies have examined the relationship between language processing and these types of reasoning and found no overlap: e.g., mathematical reasoning (Fedorenko et al., 2011; Amalric & Dehaene, 2016; Amalric et al., 2019; for evidence from aphasia, see Varley et al., 2005), computer code comprehension (Ivanova et al., 2020; Liu et al., 2020), physical reasoning (Kean et al., 2025), or social reasoning (Paunov et al., 2019, 2022; Shain, Paunov, Chen et al., 2022; Du et al., 2024; for evidence from aphasia, see Varley & Siegal, 2000; Apperley et al., 2006; Willems et al., 2011). Our study adds to this body of work, showing that formal logical reasoning—including both induction and deduction—does not recruit nor require linguistic representations.

Why do language processing and logical reasoning dissociate? Despite the fact that language can be used to express complex ideas, the representations and computations that support linguistic processing (i.e., retrieving words from memory and combining them into structured representations) appear to be distinct from those that mediate formal reasoning abilities, such as mathematical and logical reasoning. This dissociation presumably stems from the distinct demands associated with linguistic communication vs. formal reasoning. One key difference may have to do with the kinds of meanings that language vs. these other systems typically express: namely, natural languages tend to express meanings related to the external and internal world, but mathematics, logic, and computer code express mostly abstract, relational meanings that do not

bear a direct or necessary connection to the external world (see Malik-Moraleda et al., 2025 for further discussion).

Another reason may be that the linguistic format is actually not well-suited for formal reasoning, in spite of superficial similarities in the structured nature of both linguistic expressions and formal logical expressions. In particular, linguistic representations are noisy and ambiguous in ways that make them unreliable for supporting formal inference. Natural language is riddled with referential vagueness, scope ambiguities, and under-specification, requiring pragmatic enrichment, all of which can obscure the logical structure of a sentence. Evidence from AI research shows boosts in LLM performance on reasoning tasks when linguistic inputs are converted into first-order logic, which can then be fed into an external theorem solver (e.g., the LINC system; Olausson, Gu, Lipkin, Zhang, et al., 2024; the SatLM system; Ye et al., 2023; the LogicGuide system; Poesia et al., 2023; the Logic-LM system; Pan et al., 2023; and many others, e.g. Nye et al., 2021; Borazjanizadeh and Piantadosi, 2024). This work suggests that natural language is not a reliable medium for inference, which requires context-independent, structurally explicit representations.

Alternatives to linguistic representations. A central outstanding question concerns the form of non-linguistic representations, such as those supporting logical induction and deduction, in the human mind and brain. Multiple competing theoretical accounts exist, which make distinct predictions about the format and neural implementation of the representations involved in reasoning. For example, according to the *mental model* (or mental simulation) framework, reasoners construct analog simulations of the problem, exploiting visuo-spatial representations and working memory resources to model the possible consequences of premises (Johnson-Laird et al., 1983; Knauff, 2013). These accounts predict the engagement of brain areas that support visual imagery and spatial working memory (e.g., Knauff et al., 2002; Alfred et al., 2020). Alternatively, *script- or schema-based* approaches propose that reasoning draws on cached relational templates (structured representations of events or situations abstracted from prior experience) (Piaget, 1923; Bartlett, 1932; Schank & Abelson, 1977; Bower et al., 1979). These accounts suggest that reasoning is scaffolded by context- and domain-specific knowledge structures, drawing on multiple semantic networks shaped by prior experience (e.g., Gilboa & Marlatte, 2017; Masís-Obando et al., 2022). And the *mental logic* proposals treat abstract reasoning as the manipulation of propositional (sometimes, language-like) symbols according to syntactic inference rules (Rips, 1988, 1995, 2003). These accounts posit that reasoning operates over amodal, abstract symbols, potentially implemented in frontal and parietal circuits (e.g., Monti et al., 2009; Reverberi et al., 2007). Perhaps the most promising candidate from this class of proposals is the probabilistic language of thought (PLOT) hypothesis (Goodman et al., 2014; Ellis et al., 2020; Rule et al., 2020), where mental algorithms are construed as symbolic programs over concepts, which encode probabilistic knowledge, including both knowledge of particular domains and abstract relational knowledge. In this way, PLOT provides a flexible medium for both domain-specific and abstract, domain-general reasoning, and captures the graded nature of mental representations.

Given that these different accounts make distinct implementation-level predictions, future brain imaging studies may help determine whether logical reasoning depends on visuo-spatial, fully abstract, or domain-specific (at least during early stages of development) representations.

Modularity of reasoning. Aside from the dissociation between language processing and logical reasoning, this study also highlights the fact that different kinds of reasoning dissociate from one another. Some forms of reasoning have been shown to be domain-specific, including social reasoning (Saxe & Kanwisher, 2003; Gallagher & Frith, 2003; Schurz et al., 2014) and intuitive physical reasoning (Fischer et al., 2016; Schwettmann et al., 2019; Pramod et al., 2022; 2025: Mitko & Fischer, 2024). Both of these systems are distinct from the Default network, implicated in episodic self-projection and constructing situation models (Hassabis & Maguire, 2007; Spreng et al., 2009; Baldassano et al., 2018; Buckner & DiNicola, 2019), and the Multiple Demand (MD) network, implicated in diverse goal-directed behaviors (Duncan, 2010; Assem et al., 2020a). We here show that inductive reasoning, similar to mathematical reasoning (e.g., simple arithmetic), draws on the MD network, but deductive reasoning does not engage the MD network, instead recruiting a distinct set of frontal and parietal areas, in line with past work (Kroger et al., 2008; Coetzee & Monti, 2018). How these deductive-reasoning areas relate to other known systems supporting complex cognition, why these different forms of reasoning dissociate, and whether there exist other specialized reasoning systems remains to be discovered.

Limitations and open questions. The current study is limited in several ways. (1) The study is limited to native English speakers. Although we are not aware of hypotheses about cross-linguistic differences in whether language and logical reasoning overlap, it remains important to generalize these findings to other linguistic populations (Blasi et al., 2022; Malik-Moraleda et al., 2022). (2) The aphasia component is limited to two participants. This small number was dictated by the requirement of a *profound* linguistic impairment, which is rare. Most individuals with aphasia retain or recover linguistic abilities to some degree (Wilson et al., 2022). Evidence of intact logical abilities in such patients would be of limited utility because it would always be possible that they are relying on the remaining portions of the language network to perform the reasoning tasks. Evidence of intact cognition in severely linguistically impaired individuals is precious and important, even if it comes from a limited number of participants. After all, at its inception, the field of cognitive neuroscience gained some of its greatest insights from case studies (Harlow, 1848; 1869; Broca, 1861; Scoville & Milner, 1957). And (3): It would be useful to extend these findings to a broader array of logical reasoning tasks, including contextualized paradigms (Cosmides, 1989; Cox & Griggs, 1982; Griggs & Cox, 1982; Gigerenzer & Hug, 1992), and to develop non-verbal versions of purely deductive tasks for patients with aphasia. Our study also leaves a number of research questions open. For example: Is the language-logic dissociation already present during early childhood, or does it emerge over development? And is the dissociation between linguistic processing and logical reasoning, or among different types of reasoning, an inevitability given the computational demands of these different cognitive tasks, or is it merely an accident of biological evolution—a question that can now be tested in large reasoning models (Yang et al., 2019; Csordás et al., 2021; Lepori et al., 2023a, 2023b; AlKhamissi et al., 2025).

Conclusions. Our results provide key evidence against the hypothesis that natural language serves as the medium of abstract reasoning and suggest that logical reasoning is underpinned by a distinct representational format.

MATERIALS AND METHODS

Study 1 (fMRI) Participants. Twenty-nine participants contributed data to the fMRI component of the study (15 female; mean age = 27.3 years, SD = 12.9; all but three participants were right-handed; the one ambidexterous and two left-handed participants had left-lateralized language systems, as determined by the language localizer task described below. All participants completed the language localizer task and the Multiple Demand (MD) localizer task (used in control analyses). Different subsets of the 29 participants completed the three logic tasks, with at least 17 participants performing each task (17 completed the induction task; 23 completed the verbal deduction task; and 17 completed the deductive matrix reasoning task). The participants were recruited from MIT and the surrounding Cambridge/Boston, MA, community and paid for their participation. All were native English speakers, had normal hearing and vision, and had no history of language impairment. All participants provided written informed consent in line with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects.

Study 2 (behavior) Participants. Two profoundly aphasic male participants took part in the study (S.A. 78 years; G.S. 50 years). Both had large lesions that had significantly damaged the left inferior frontal and left temporal brain areas. Both were native English speakers, did not present with any visual impairments, and were pre-morbidly right-handed. Both individuals were classified as severely agrammatic (**SI Table 3**), but their nonlinguistic cognitive skills were largely spared, with the exception of below-average Digit Span scores, in line with an impaired phonological loop (**SI Table 3**). We also tested 40 age-matched neurotypical control participants (20 female; mean age 56 years, SD = 10.7, range: 40 to 86 years). The mean number of years of education was 17.2 (SD = 2.3), perhaps driven up by the fact that our recruitment posters described the study as consisting of logic puzzles. All participants were native English speakers, had normal, or corrected-to-normal vision, and no history of speech or language disorders, neurological diseases, or reading impairments, and performed within the normal range on the mini mental status examination (MMSE; Folstein et al., 1975). All participants completed the experiments individually, in a quiet room, with an experimenter present throughout the testing session. Ethics approval was granted by the UCL Research Committee (LC/2023/05). All participants provided informed consent prior to taking part in the study.

Induction Task (fMRI, behavior). Participants were shown an input list of 1-5 single-digit numbers and an output list of 0-5 single-digit numbers and asked to guess the rule that transformed the input list into the output list. They were then shown another input list and asked to provide the output list. They were told whether or not their guess about the underlying rule is correct and are shown another input list. In the **fMRI version**, the responses were entered via a scanner-safe fiber-optic button response device (Nata Technologies LxPad system), which contains 12 buttons allowing input of digits 0–9, backspace, and enter (**Figure 1**, left panel). Each trial (corresponding to a rule) consisted of 8 problems. After the eighth problem, participants were told—via a schematic (**Figure 1**, left panel)—the correct rule and were asked to apply this rule to two more input lists. Because in most cases, participants guessed the correct rule part-way through the initial list of 8 problems (as indicated by generating correct output lists for the last several problems), we defined the *induction* period as the subset of the eight problems before the problem after which the participant made no more errors, and the *application* period as the subsequent problems along with the two problems after the correct rule was revealed. The induction > application contrast targets

neural processes involved in hypothesis generation and rule discovery beyond those required for rule implementation and response entry. Each participant completed 40 rules (all the materials are available at OSF: https://osf.io/jm9rd). The rules were distributed across 20 scanning runs, with 2 rules per run and each run lasting between 200s and 420s (durations are variable given the self-paced nature of the task; see **SI Figure 3A** for the timing details).

For the **behavioral version** of the task, the problems were presented on paper, and participants provided written responses. The same 40 rules were used, except that a few problems were added in order to have 14 problems (input-output pairs) per rule, and a few problems were replaced based on feedback from the fMRI participants who found some input-output pairs confusing. The stimuli were divided into four sets, with ten rules per set. GS completed all 40 rules across several sessions (~6 rules per session); SA only completed sets 1-3 also across sessions (~4 rules per session); the variability in the number of rules per session was due to age-related health conditions and because the experiment had to be completed amongst other necessary activities at routine check-in sessions. Prior to the experiment, participants with aphasia were shown an example rule, which acted as a training item. Control participants each completed one set of 10 rules within a single testing session, and 10 participants completed each set. A break was offered half-way through the testing session, during which they could rest and also complete the brief MMSE (performance on the MMSE was used as an inclusion criterion). The order of the rules within a set was kept constant across participants, but the order of the problems within a rule was randomized. For both the patients and the controls, testing for a given rule was stopped after the participant answered four problems in a row correctly, which was taken as evidence that they guessed the transformation rule correctly. Time permitting, the participants with aphasia were sometimes offered to schematically illustrate the rule they guessed (see **Figure 2** for an example).

Deduction Task 1 – Verbal deductive reasoning (fMRI). Participants were presented with classic three-sentence syllogisms and asked to judge the validity of the third sentence (the conclusion) given the first two sentences (the premises), by pressing one of two buttons on a button box. Each trial corresponded to a syllogism, and trials varied in difficulty between harder deduction (Modus Tollens) and easier deduction (Modus Ponens), and in whether the problems used real words or nonwords; the experiment also included two conditions of no interest (challenging memory conditions). The Modus Tollens > Modus Ponens contrast targets cognitive processes related to the complexity of logical deductive reasoning. Each participant completed between 16 and 32 trials of each condition (all the materials are available at OSF: https://osf.io/jm9rd). The trials were distributed across two scanning runs, with 8 trials per condition per run and each run lasting between 551 s and 1,034 s (durations are variable given the self-paced nature of the task; see **SI Figure 3B** for the timing details).

Deduction Task 2 – Deductive matrix reasoning (fMRI). Participants were presented with sets of four images of geometrical shapes and asked to decide which image does not fit with the others, by pressing one of four buttons on a button box. The experiment used a blocked design; trials in the hard blocks used stimuli from Cattell (1949), and trials in the easy blocks used simpler problems created by Woolgar et al. (2013), who adapted this task for fMRI (e.g., three images of the same simple shape and an image of a different shape). The hard > easy contrast targets cognitive processes related to relational reasoning, hypothesis generation, and deductive inference. Each participant completed a single run of the task lasting 320 s and consisting of 8 easy blocks

and 8 hard blocks. The blocks were of fixed length (16 s), which means that only a few hard trials could be solved during this period (between one and six), and more easy trials could be solved (between 4 and 15; see **SI Figure 3C** for the timing details); furthermore, because only 24 hard items were available, some participants did not have enough items for all 8 blocks, but every participant completed at least 5 hard blocks (all the materials are available at OSF: https://osf.io/jm9rd).

Deduction Task 3 – Deductive matrix reasoning (behavior). This task is the Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II), which is conceptually parallel to the matrix reasoning fMRI task. Participants were presented with sets of several images of geometrical shapes with one image missing and asked to choose the missing image from five possible options to complete the pattern. The test was administered on paper and participants indicated their choice by pointing. The items (total number = 30) were presented in ascending order of difficulty, consistent with standard WASI-II administration procedures. Prior to the main test, participants were shown a short set of practice items to familiarize them with the task. These practice items illustrated the nature of the patterns, and participants were provided with explanations (including gestural cues, where necessary) to aid comprehension. Testing was stopped after the participant answered three problems incorrectly.

Critical Analyses - fMRI. (For the details of fMRI data acquisition, preprocessing, and modeling, see Supplementary Methods.) To test whether the language areas respond during logical reasoning tasks, we used an extensively validated language localizer (Fedorenko et al., 2010) to identify regions of interest functionally in individual participants. Participants were asked to attentively read sentences and sequences of nonwords; for details on this paradigm, see Supp. Methods. The responses in the language fROIs were statistically evaluated using linear mixed-effects models implemented in R (lme4 package; Bates et al., 2015), with random intercepts for participants and regions. P-values were approximated using the lmerTest package (Kuznetsova et al., 2017), and effect sizes (Cohen's *d*) were estimated using the EMAtools package (Kleiman, 2017). For each contrast, we fit separate models at both the network level and for individual fROIs.

The **network-level model**: *BOLD ~ Condition + (1 | Participant) + (1 | fROI).*
The **individual fROI-level model**: *BOLD ~ Condition + (1 | Participant).*

To ensure that our logical reasoning tasks elicit a strong response somewhere in the brain, we performed two analyses. *First*, we used a spatial working memory task (keeping track of 8 vs. 4 squares in a 3x4 grid) as a localizer for the Multiple Demand network (e.g., Assem et al., 2020b); for details on this paradigm see Supp. Methods. And *second*, because the verbal deductive reasoning task did not elicit a response in the MD network, we additionally performed a whole-brain search for areas that respond to the Modus Tollens > Modus Ponens contrast. We used a group-constrained subject-specific (GSS) analysis (Fedorenko et al., 2010; Julian et al., 2012), which is similar to a random-effects group analysis but allows for inter-individual variability in the precise locations of functional areas. This analysis identified a few areas in the frontal and parietal cortex that responded strongly to the deduction contrast (in left-out data). The responses in the MD and deductive-reasoning fROIs were statistically evaluated similarly to the language fROIs' responses.

Critical Analyses – Behavior. For the induction task, we used the Crawford and Howell (1998) test to compare each patient's performance to the controls. The Matrix Reasoning WASI-II data were scored according to the standard guidelines provided in the official manual. Each item was evaluated for accuracy, and raw scores (total correct out of 30) were converted into T-scores using age-normed tables ($\mu = 50$, $\sigma = 10$). These T-scores were the basis for statistical comparison with the normative control sample from the testing manual, consistent with standard single-case methods and group-level reference data. (For information on the behavioral performance of participants in the fMRI task, see Supp. Methods.)

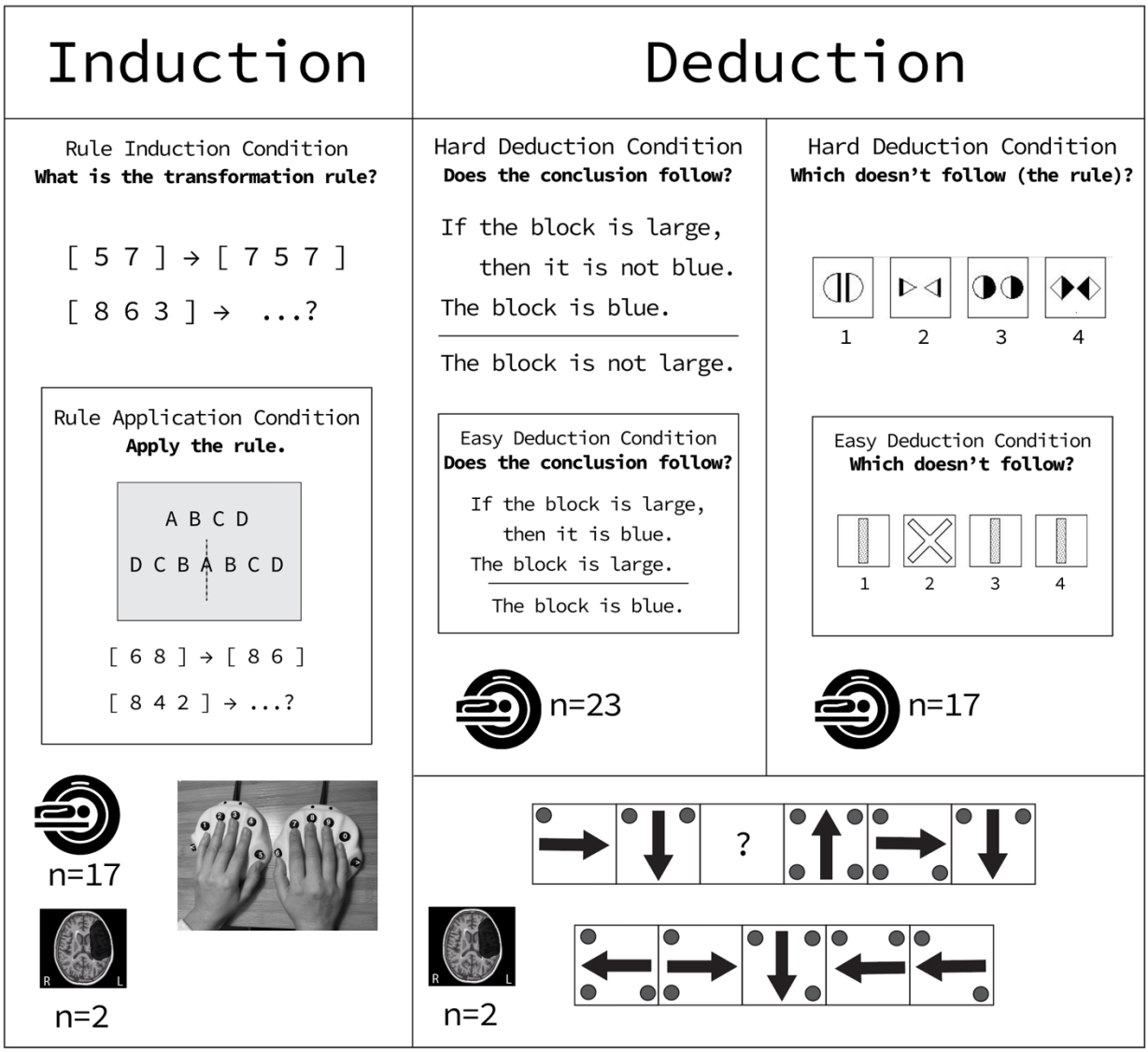


Figure 1. Inductive and deductive logical reasoning paradigms. For inductive reasoning (left), participants are shown an input list of 1-5 single-digit numbers and an output list of 0-5 single-digit numbers and asked to guess the rule that transformed the input list into the output list. They are then shown another input list and asked to provide the output list. They are told whether or not their guess about the underlying rule is correct and are shown another input list. In the fMRI version, after the eighth problem, participants were told—via a schematic— the correct rule and were asked to apply this rule to two more input lists. For deductive reasoning (right), in the verbal version, participants are presented with classic three-sentence syllogisms and asked to judge the validity of the third sentence (the conclusion) given the first two sentences (the premises), by pressing one of two buttons on a button box. Trials vary in difficulty between harder deduction (Modus Tollens) and easier deduction (Modus Ponens). In the matrix reasoning version, the fMRI participants are presented with sets of four images of geometrical shapes and asked to decide which image does not fit with the others, by pressing one of four buttons on a button box. The experiment uses a blocked design; trials in the hard blocks use stimuli from Cattell (1949), and trials in the easy blocks use simpler problems created by Woolgar et al. (2013), who adapted this task for fMRI (e.g., three images of the same simple shape and an image of a different shape). In the matrix reasoning version for the participants with aphasia, the subtest of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II) was used, which is conceptually parallel to the matrix reasoning fMRI task. Participants are presented with sets of

several images of geometrical shapes with one image missing and asked to choose the missing image from five possible options to complete the pattern. The test was administered on paper and participants indicated their choice by pointing. The items (total number = 30) are presented in ascending order of difficulty, consistent with standard WASI-II administration procedures.

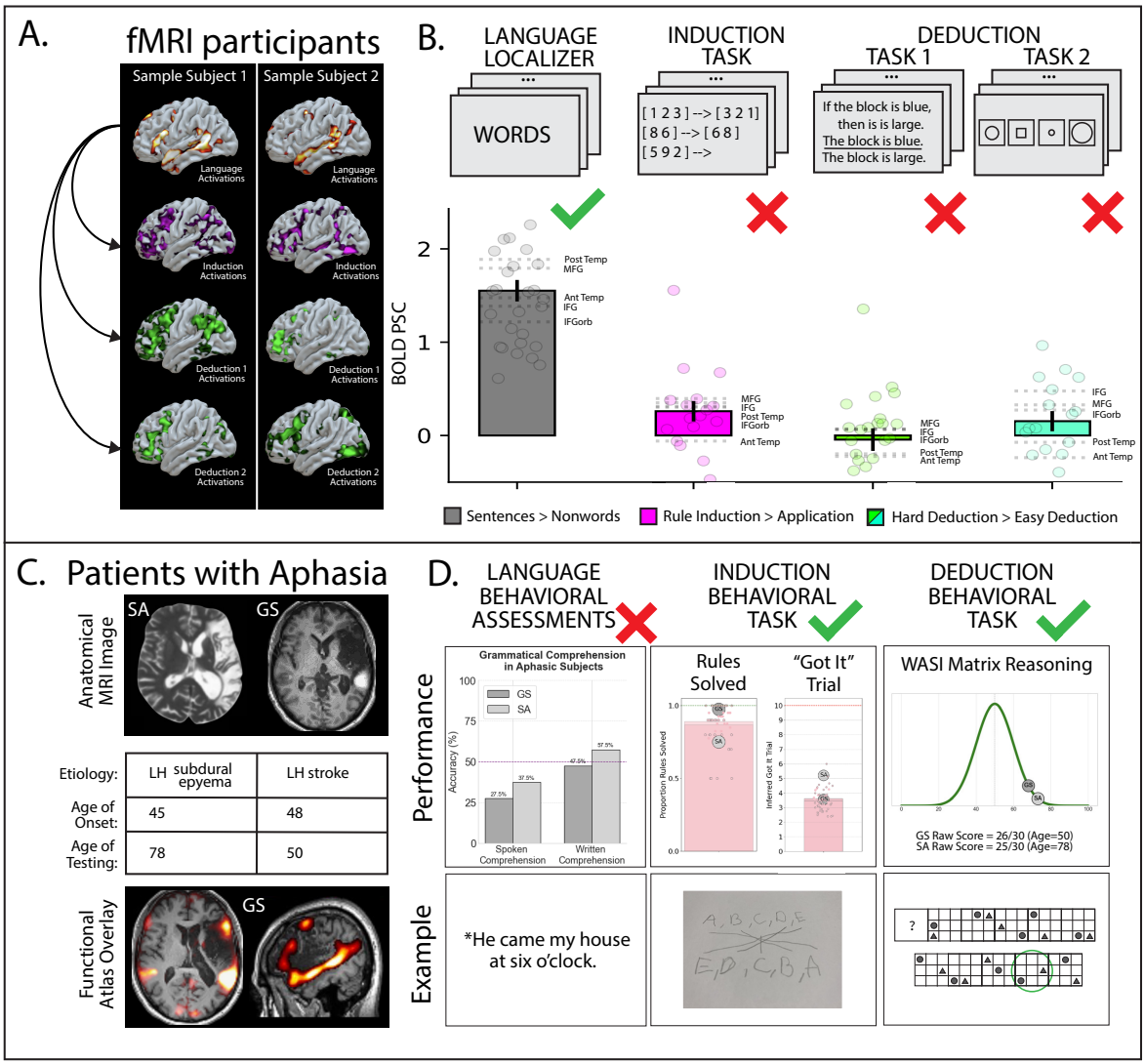
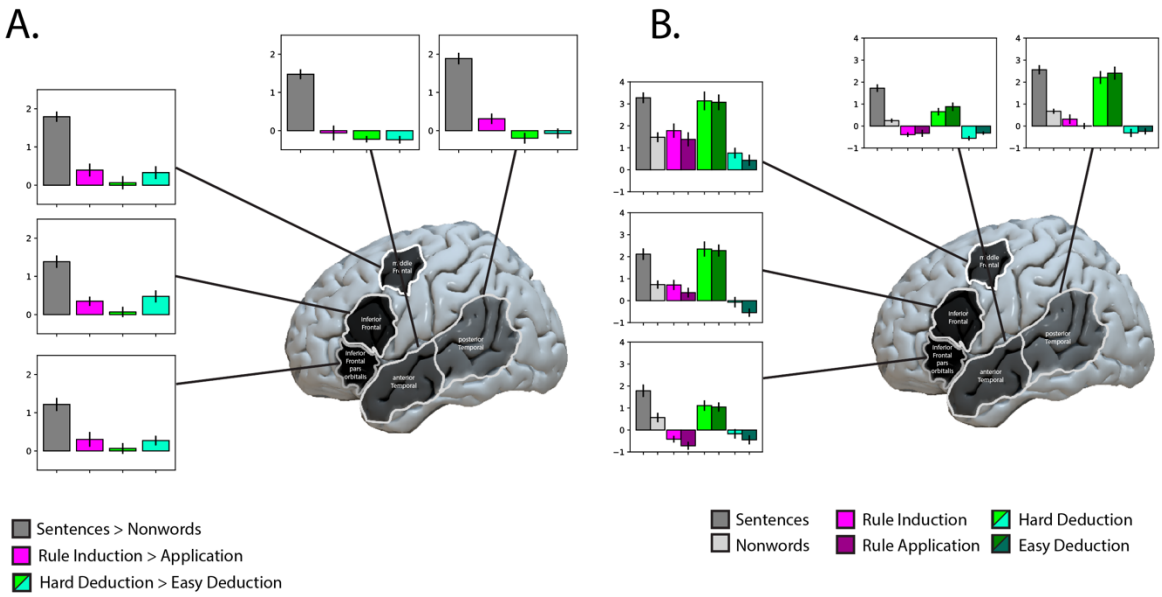


Figure 2. The language system does not support logical reasoning. A. Sample activation maps for the language contrast (1st row, yellow) and three logical reasoning contrasts (2nd-4th rows, purple and green) in two individual sample participants. **B.** Responses in the language network (individually defined in each participant; see Methods), in percent BOLD signal change, to the language contrast (sentences > nonwords; estimated in left-out data; grey), the induction contrast (induction > application; magenta), and the two deduction contrasts (hard deduction > easy deduction; green shades). The bars show the across-network averages; the means for each of the 5 language fROIs (IFGorb, IFG, MFG, AntTemp, and PostTemp) are shown with horizontal dashed lines. The error bars correspond to the standard error of the mean by participant; the light dots correspond to individual participants' responses. **C.** The anatomical scans of S.A. and G.S., brief info on the patients, and a projection of the probabilistic atlas for the language network (from Lipkin et al., 2022) into the MRI image of G.S. **D.** Behavioral performance of the patients with aphasia and—for the critical tasks—control participants. The first column shows performance on the language evaluation tasks (spoken and written comprehension; Zimmerer et al., 2014) for S.A. (light grey bars) and G.S. (dark grey bars). Both participants show at or below chance performance. In the bottom row, we show sample items. (See Methods for additional details of the linguistic evaluation of the patients.) The second and third columns show performance on the induction and deduction tasks (S.A.: light grey dots; G.S.: dark grey dots). For the induction task, the control data

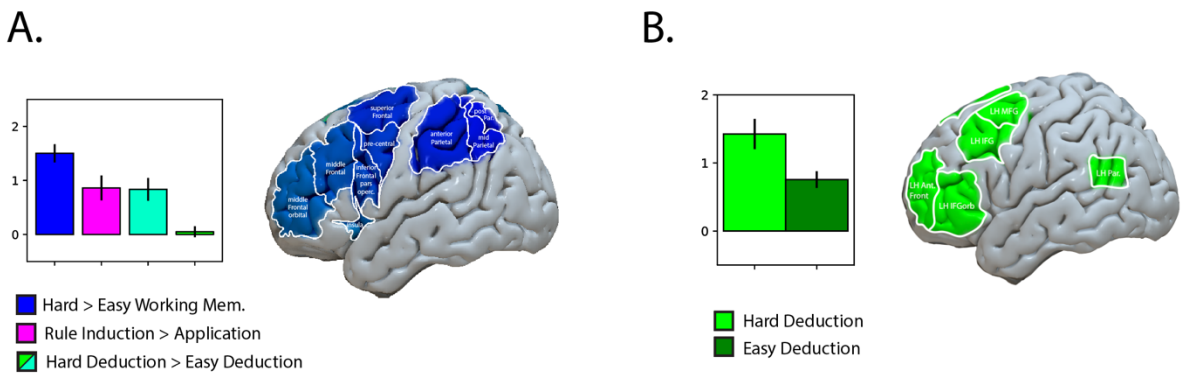
are shown with a pink bar. In the bottom row, we show a sample rule representation generated by G.S., where he nicely illustrates the “reverse the numbers” rule. For the deduction task, the control data are shown with the normal age-matched distribution, and in the bottom row, we show a sample item.

Predictors	Estimates	p-value
Language Task		
Intercept (nonwords)	0.7407	0.0666
Critical Effect (sentences)	**1.5518**	**<0.0001**
Induction Task		
Intercept (application/post-solution)	0.1408	0.7414
Critical Effect (induction/pre-solution)	0.2601	0.0224
Deduction Task 1		
Intercept (easy/Modus Ponens)	1.9361	0.0065
Critical Effect (hard/Modus Tollens)	-0.0444	0.7102
Deduction Task 2		
Intercept (easy deduction)	-0.2234	0.3527
Critical Effect (hard deduction)	0.1533	0.1185

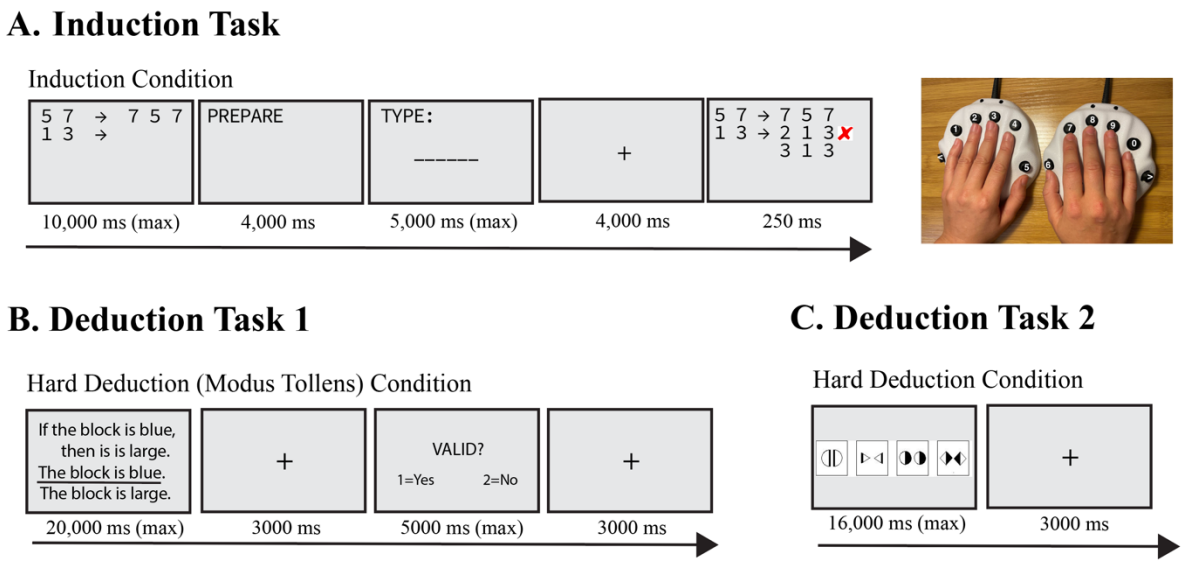
Table 1. The responses of the language system to the language and logical reasoning contrasts. The results from the linear mixed-effects models (see Methods) at the network level (see Supp. Table 1 for the results broken down by fROI).



Supplementary Figure 1. The language system does not support logical reasoning: The results broken down by fROI. A. Responses in the five language fROIs (individually defined and constituting 10% of the broad parcel masks shown on the brain; see Methods), in percent BOLD signal change, to the language (grey), induction (magenta), and deduction (green shades) contrasts. The style matches the style of main Figure 2A. **B.** Responses in the five language fROIs to the individual conditions relative to the fixation baseline (cf. the contrasts shown in panel A).



Supplementary Figure 2. The brain systems that support logical reasoning. A. Responses in the Multiple Demand (MD) system, in percent BOLD signal change, to the MD localizer contrast (a spatial working memory task; see Methods; blue), and the three logical reasoning contrasts (magenta and green shades). The MD fROIs are defined within individuals and constitute 10% of the broad parcel masks shown in blue on the brain. The MD system shows a reliable response to the induction contrast and the matrix reasoning deduction contrast, but not to the verbal syllogism-based contrast, which more narrowly isolates deductive reasoning complexity. **B.** Responses in the brain areas identified with the Hard > Easy contrast in the verbal deductive reasoning paradigm (i.e., Modus Tollens > Modus Ponens). The fROIs are defined within individuals (using a portion of the data) and constitute 10% of the broad parcel masks shown in green on the brain. The responses are shown in left-out data and reveal a robust and replicable effect.



Supplementary Figure 3. Timing details for the three reasoning tasks used in the fMRI study. Panel A shows a sample trial structure for the Induction Task, in which participants iteratively generated outputs from given input lists and received trial-by-trial feedback to discover an underlying transformation rule. Each rule block consisted of 8 initial problems followed by 2 confirmation trials after the rule was revealed. Blocks were self-paced, and each scanning run included 2 rule blocks. Panel B shows the trial timing for Deduction Task 1, which involved syllogistic reasoning using three-statement arguments. Trials varied in difficulty (e.g., Modus Ponens vs. Modus Tollens), and participants judged the logical validity of the conclusion. Each run included multiple trials and was self-paced. Panel C shows a sample trial from Deduction Task 2, which had a blocked design and which required participants to identify the odd image out among abstract shape sets. Blocks were either "easy" (simple perceptual groupings) or "hard" (Cattell-type fluid reasoning problems). Each block was 16 seconds long and fixed in length such that participants could complete as many trials in a block as they were able; all participants completed at least five hard blocks. See OSF (https://osf.io/jm9rd) for full task materials.

Task	Type	Beta	p-value
Lang	LH Inferior Frontal Gyrus (orbital)	**1.218**	**<0.0001**
Lang	LH Inferior Frontal Gyrus	**1.386**	**<0.0001**
Lang	LH Middle Frontal Gyrus	**1.793**	**<0.0001**
Lang	LH Anterior Temporal	**1.476**	**<0.0001**
Lang	LH Posterior Temporal	**1.887**	**<0.0001**
Induction	LH Inferior Frontal Gyrus (orbital)	0.302	0.1360
Induction	LH Inferior Frontal Gyrus	0.350	0.0092
Induction	LH Middle Frontal Gyrus	0.396	0.0301
Induction	LH Anterior Temporal	-0.060	0.7500
Induction	LH Posterior Temporal	0.312	0.0356
Deduction1	LH Inferior Frontal Gyrus (orbital)	0.0647	0.6466
Deduction1	LH Inferior Frontal Gyrus	0.0702	0.6059
Deduction1	LH Middle Frontal Gyrus	0.0645	0.7105
Deduction1	LH Anterior Temporal	-0.225	0.0117
Deduction1	LH Posterior Temporal	-0.197	0.1717
Deduction2	LH Inferior Frontal Gyrus (orbital)	0.272	0.0448
Deduction2	LH Inferior Frontal Gyrus	0.477	0.0077
Deduction2	LH Middle Frontal Gyrus	0.328	0.0655
Deduction2	LH Anterior Temporal	-0.238	0.0265
Deduction2	LH Posterior Temporal	-0.073	0.5734

Supplementary Table 1. The results from the linear mixed-effects models (see Methods) at the level of individual language fROIs. We report uncorrected p-values, and we mark the values that survive the Bonferroni correction for the number of fROIs (n=5) in **bold**.

Multiple Demand System			Deduction-selective fROIs		
Predictors	**Betas**	**p-value**	**Predictors**	**Betas**	**p-value**
Spatial WM Task			*Deduction Task 1*		
Intercept (easy)	**1.83**	**<0.0001**	Intercept	**0.76**	**0.0016**
Critical (hard)	**1.50**	**<0.0001**	Critical	**0.67**	**<0.0001**
Induction Task					
Intercept (application)	**1.81**	**<0.0001**			
Critical (induction)	**0.86**	**<0.0001**			
Deduction Task 1					
Intercept (Modus Ponens)	**2.26**	**<0.0001**			
Critical (Modus Tollens)	0.05	0.4822			
Deduction Task 2					
Intercept (easy deduction)	**1.67**	**<0.0001**			
Critical (hard deduction)	**0.84**	**<0.0001**			

Supplementary Table 2. The responses of the Multiple Demand system and the deduction-responsive brain areas to the logical reasoning contrasts. The results from the linear mixed-effects model (see Methods),

	GS	SA
Age	50	78
Time post-onset	2.5 years	33 years
Etiology	Left hemisphere stroke	Left subdural empyema and meningitis
Previous occupation and educational background	Graduate degree; worked in cybersecurity	Left school at 15; police sergeant
Cognitive assessments		
- Forward digit span (Wechsler Adult Intelligence Scale)	- 3 items	- 2 items
- Pyramid & Palm Trees (3-picture version)	- 50/52	- 49/52
Grammatical processing		
Comprehension spoken reversible sentences:		
- Total (chance = 40/80)	- 22/80	- 30/80
- Active (20/40)	- 20/40	- 16/40
- Passive (20/40)	- 2/40	- 14/40
Comprehension written reversible sentences:		
- Total (chance = 40/80)	- 38/80	- 46/80
- Active (20/40)	- 32/40*	- 34/40*
- Passive (20/40)	- 6/40	- 12/40
Auditory grammaticality judgment:		
- Total (chance = 32/64)	- 36/64	- 48/64*
Written grammaticality judgment:		
- Total (chance = 32/64)	- 44/64*	- 29/64
Lexical processing		
ADA spoken-word picture matching (total):	- 60/66	- 51/66
ADA written-word picture matching (total):	- 63/66	- 53/66
PALPA Picture Name (spoken or written):		
- High-frequency set	- 19/20	- 1/20
- Mid-frequency set	- 15/20	- 0/20**
- Low-frequency set	- 13/20	- 0/20**
ADA Synonym Judge (combined written and auditory presentation):		
- Total	- 137/160	- 135/160
- High imageability errors	- 2	- 4
- Low imageability errors	- 10	- 12
- High frequency errors	- 4	- 6
- Low frequency errors	- 7	- 3

Supplementary Table 3. Extended demographic and language assessment information for participants with aphasia. S.A. was premorbidly a police sergeant and G.S. was premorbidly in cybersecurity. S.A. suffered a subdural empyema, followed by meningitis, leading to a secondary vascular lesion in the left middle cerebral artery and damage to the perisylvian areas. GS suffered a left-hemisphere stroke, also resulting in a vascular lesion in the left middle cerebral artery and damage to the perisylvian areas. This table further shows cognitive assessments include the WAIS digit span (Wechsler, 1981), Pyramid and Palm Trees test (3-picture version), identifying semantic relations between pictures (Howard and Patterson, 1992) and WASI matrices (Wechsler, 1999). Grammatical processing assessments from

Zimmerer et al., (2014) include comprehension of spoken and written reversible sentences and from Linebarger et al. (1983), grammaticality judgments, with repetition penalties for spoken/auditory tasks. Lexical processing includes the picture matching and synonym judge tasks from the ADA comprehension battery (Franklin et al. 1992) and picture naming task from the PALPA test battery (Kay et al. 1992). * indicates performance above chance under a binomial test at $p = 0.05$. ** a cut-off was applied for SA's PALPA naming test: with 1/20 on the high frequency set, the mid- and low-frequency sets were automatically scored 0/20 without further testing.

Supplementary Methods and Materials

fMRI data acquisition

Whole-brain structural and functional data were collected on a whole-body 3 Tesla Siemens Trio scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted structural images were collected in 176 axial slices with 1 mm isotropic voxels (repetition time (TR) = 2,530 ms; echo time (TE) = 3.48 ms). Functional, BOLD data were acquired using an EPI sequence with a 90° flip angle and using GRAPPA with an acceleration factor of 2; the following parameters were used: 31 4.4 mm thick near axial slices acquired in an interleaved order (with 10% distance factor), with an in-plane resolution of 2.1 × 2.1 mm, FoV in the phase encoding (A >> P) direction 200 mm and matrix size 96 × 96 voxels, TR = 2,000 ms and TE = 30 ms. The first 10 s of each run were excluded to allow for steady state magnetization.

fMRI data preprocessing

fMRI data were preprocessed and analyzed using SPM12 (release 7487), CONN EvLab module (release 19b), and other custom MATLAB scripts. Each participant's functional and structural data were converted from DICOM to NIFTI format. All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session (Friston et al., 1995). Potential outlier scans were identified from the resulting subject-motion estimates as well as from BOLD signal indicators using default thresholds in CONN preprocessing pipeline (5 standard deviations above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm; Nieto-Castañón, 2020). Functional and structural data were independently normalized into a common space (the Montreal Neurological Institute [MNI] template; IXI549Space) using SPM12 unified segmentation and normalization procedure (Ashburner and Friston 2005) with a reference functional image computed as the mean functional data after realignment across all timepoints omitting outlier scans. The output data were resampled to a common bounding box between MNI coordinates (−90, −126, −72) and (90, 90, 108), using 2 mm isotropic voxels and 4th order spline interpolation for the functional data, and 1 mm isotropic voxels and trilinear interpolation for the structural data. Last, the functional data were smoothed spatially using spatial convolution with a 4 mm FWHM Gaussian kernel.

fMRI data modeling

For all experiments, effects were estimated using a general linear model (GLM) in which each experimental condition was modeled with a boxcar function convolved with the canonical hemodynamic response function (HRF) (fixation was modeled implicitly, such that all timepoints that did not correspond to one of the conditions were assumed to correspond to a fixation period). Temporal autocorrelations in the BOLD signal timeseries were accounted for by a combination of high-pass filtering with a 128 s cutoff, and whitening using an AR (0.2) model (first-order autoregressive model linearized around the coefficient a = 0.2) to approximate the observed covariance of the functional data in the context of restricted maximum likelihood estimation. In addition to experimental condition effects, the GLM design included first-order temporal derivatives for each condition (included to model variability in the HRF delays), as well as nuisance regressors to control for the effect of slow linear drifts, subject-motion parameters, and potential outlier scans on the BOLD signal.

Details on the Language Localizer.

In this task, participants silently read sentences and lists of nonwords in a blocked design. The sentences > nonwords contrast targets cognitive processes related to high-level language comprehension, including understanding word meanings and combinatorial linguistic processing, and has been shown to effectively isolate language areas from nearby functional areas (see Fedorenko et al., 2024 for a review). The task is available for download from https://www.evlab.mit.edu/resources. Following much past work (e.g., Lipkin et a. 2022; Malik-Moraleda et al., 2022; Tuckute et al., 2024), to define the language functional regions of interest (fROIs), we used a set of group-level parcels—derived from a large group of independent

participants performing the same task—and intersected these parcels with individual activation maps; within each parcel, we selected the 10% of most responsive voxels in each participant as that participant's fROI. The responses in these language fROIs to each condition of the critical tasks were estimated; to estimate the responses to the conditions of the language localizer, we used a split-half approach where one run is used to define the fROIs, and the other run to estimate their responses, so as to avoid non-independence (e.g., Kriegeskorte, 2011).

Details on the Spatial Working Memory MD Localizer.

In this task, participants are presented with 3x4 grids within which sequences of locations flash up (one at a time for a total of four locations in the easy condition, and two at a time for a total of eight locations in the hard condition). Participants are asked to keep track of the locations, and at the end of each trial, they are shown two sets of locations and asked to choose the set they just saw. The hard > easy contrast targets cognitive processes broadly related to performing demanding tasks—what is often referred to by an umbrella term 'executive function processes'. To define the MD fROIs, we used a set of group-level parcels and, similar to the language localizer, selected the 10% of most responsive voxels in each participant as that participant's fROI.

In-scanner behavioral data.

All fMRI participants were engaged during the logical reasoning tasks: the overall response rate was 100% for the induction task, 99.63% for the verbal deductive reasoning task, and 100% for the deductive matrix reasoning task. The accuracies were high: induction task—0.77 (0.74 for the pre-solution induction trials and 0.89 for the post-solution application trials); verbal deductive reasoning task—0.81 (0.67 for the harder Modus Tollens condition and 0.95 for the easier Modus Ponens condition); deductive matrix reasoning task—72.13% (47% for the harder condition and 97.25% for the easier condition). And the reaction times (RTs) mirrored the accuracy data, with longer RTs for the harder conditions: induction task—4.17s (4.49s for the pre-solution induction trials and 2.88s for the post-solution application trials); verbal deductive reasoning task—11.31s (13.52s for the harder Modus Tollens condition and 9.10s seconds for the easier Modus Ponens condition); deductive matrix reasoning task—2.364s (3.73s for the harder condition trials and 0.997s for the easier condition trials).

References

Alfred, K. L., Connolly, A. C., Cetron, J. S., & Kraemer, D. J. M. (2020). Mental models use common neural spatial structure for spatial and abstract content. *Communications Biology*, *3*(1), 17. https://doi.org/10.1038/s42003-019-0740-8

AlKhamissi, B., Sabbata, C. N. D., Chen, Z., Schrimpf, M., & Bosselut, A. (2025). *Mixture of Cognitive Reasoners: Modular Reasoning with Brain-Like Specialization* (arXiv:2506.13331). arXiv. https://doi.org/10.48550/arXiv.2506.13331

Amalric, M., & Dehaene, S. (2016). 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. https://doi.org/10.1073/pnas.1603205113

Amalric, M., & Dehaene, S. (2019). A distinct cortical network for mathematical knowledge in the human brain. *NeuroImage*, *189*, 19–31. https://doi.org/10.1016/j.neuroimage.2019.01.001

Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. (2006). Intact first- and second-order false belief reasoning in a patient with severely impaired grammar. *Social Neuroscience*, *1*(3–4), 334–348. https://doi.org/10.1080/17470910601038693

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience*, *16*(10), 1773–1784. https://doi.org/10.1162/0898929042947928

Aristotle. (1938). *Categories and De Interpretatione* (H. Cooke & H. Tredennick, Trans.). Harvard University Press.

Assem, M., Blank, I. A., Mineroff, Z., Ademoğlu, A., & Fedorenko, E. (2020). Activity in the fronto-parietal multiple-demand network is robustly associated with individual differences in working memory and fluid intelligence. *Cortex*, *131*, 1–16. https://doi.org/10.1016/j.cortex.2020.06.013

Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020). A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. *Cerebral Cortex*, *30*(8), 4361–4380. https://doi.org/10.1093/cercor/bhaa023

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of Real-World Event Schemas during Narrative Perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(45), 9689–9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology* (pp. xix, 317). Cambridge University Press.

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, *144*(1), 77–110. https://doi.org/10.1037/bul0000130

Bediou, B., Rodgers, M. A., Tipton, E., Mayer, R. E., Green, C. S., & Bavelier, D. (2023). Effects of Action Video Game Play on Cognitive Skills: A Meta-Analysis. *Technology, Mind, and Behavior*, *4*(1: Spring 2023). https://doi.org/10.1037/TMB0000102

Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., & Grodzinsky, Y. (2003). The Neural Reality of Syntactic Transformations: Evidence From Functional Magnetic Resonance Imaging. *Psychological Science*, *14*(5), 433–440. https://doi.org/10.1111/1467-9280.01459

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

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

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Boix-Adsera, E., Bengio, S., Saremi, O., & Littwin, E. (2024). *WHEN CAN TRANSFORMERS REASON WITH ABSTRACT SYMBOLS?*

Borazjanizadeh, N., & Piantadosi, S. T. (2024). *Reliable Reasoning Beyond Natural Language* (arXiv:2407.11373). arXiv. https://doi.org/10.48550/arXiv.2407.11373

Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y., & Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies during language comprehension. *NeuroImage*, *26*(1), 221–233. https://doi.org/10.1016/j.neuroimage.2005.01.032

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*(2), 177–220. https://doi.org/10.1016/0010-0285(79)90009-4

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin de la Société Anatomique de Paris*, *6*, 330–357.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (2017). *A Study of Thinking* (1st ed.). Routledge. https://doi.org/10.4324/9781315083223

Buckner, R. L., & DiNicola, L. M. (2019). The brain's default network: Updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience*, *20*(10), 593–608. https://doi.org/10.1038/s41583-019-0212-7

Caplan, D., Chen, E., & Waters, G. (2008). Task-dependent and task-independent neurovascular responses to syntactic processing. *Cortex*, *44*(3), 257–275. https://doi.org/10.1016/j.cortex.2006.06.005

Carey, S. (2009). *The origin of concepts* (pp. viii, 598). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195367638.001.0001

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, *25*(6), 657–726. https://doi.org/10.1017/S0140525X02000122

Cattell, R. B. (1940). A culture-free intelligence test. I. *Journal of Educational Psychology*, *31*(3), 161–179. https://doi.org/10.1037/h0059043

Cattell, R. B. (1949). The dimensions of culture patterns by factorization of national characters. *The Journal of Abnormal and Social Psychology*, *44*(4), 443–469. https://doi.org/10.1037/h0054760

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science (New York, N.Y.)*, *359*(6381), 1263–1266. https://doi.org/10.1126/science.aao3539

Chen, X., Affourtit, J., Ryskin, R., Regev, T. I., Norman-Haignere, S., Jouravlev, O., Malik-Moraleda, S., Kean, H., Varley, R., & Fedorenko, E. (2023). The human language system, including its inferior frontal component in "Broca's area," does not support music perception. *Cerebral Cortex (New York, N.Y.: 1991)*, *33*(12), 7904–7929. https://doi.org/10.1093/cercor/bhad087

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. (2024). *ARC Prize 2024: Technical Report*.

Chomsky, N. (1957). *Syntactic structures* (p. 116). Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. M.I.T. Press.

Chomsky, N. (1993). A Minimalist Program for Linguistic Theory. In K. L. Hale & S. J. Keyser (Eds.), *The View From Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. MIT Press.

Chomsky, N. (1995). *The Minimalist program*. The MIT Press.

Coetzee, J. P., Johnson, M. A., Lee, Y., Wu, A. D., Iacoboni, M., & Monti, M. M. (2023). Dissociating Language and Thought in Human Reasoning. *Brain Sciences*, *13*(1), Article 1. https://doi.org/10.3390/brainsci13010067

Coetzee, J. P., & Monti, M. M. (2018). At the core of reasoning: Dissociating deductive and non-deductive load. *Human Brain Mapping*, *39*(4), 1850–1861. https://doi.org/10.1002/hbm.23979

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276. https://doi.org/10.1016/0010-0277(89)90023-1

Cox, J. R., & Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. *Memory & Cognition*, *10*(5), 496–502. https://doi.org/10.3758/BF03197653

Crawford, J. R., & Howell, D. C. (1998). Comparing an Individual's Test Score Against Norms Derived from Small Samples. *The Clinical Neuropsychologist*, *12*(4), 482–486. https://doi.org/10.1076/clin.12.4.482.7241

Csordás, R., Steenkiste, S. van, & Schmidhuber, J. (2021). *Are Neural Nets Modular? Inspecting Functional Modularity Through Differentiable Weight Masks* (arXiv:2010.02066). arXiv. https://doi.org/10.48550/arXiv.2010.02066

Davidson, D. (1967). Truth and meaning. *Synthese*, *17*(1), 304–323. https://doi.org/10.1007/bf00485035

Davidson, D. (1975). Thought and Talk. In S. D. Guttenplan (Ed.), *Mind and language* (pp. 1975–1977). Clarendon Press.

de Villiers, J. G., & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false beliefs. In *Children's reasoning and the mind* (pp. 191–228). Psychology Press/Taylor & Francis (UK).

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* (arXiv:2501.12948). arXiv. https://doi.org/10.48550/arXiv.2501.12948

Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.

Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness* (pp. viii, 184). Basic Books.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds* (1st ed). W. W. Norton & Company, Incorporated.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Diachek, E., Blank, I., Siegelman, M., Affourtit, J., & Fedorenko, E. (2020). The Domain-General Multiple Demand (MD) Network Does Not Support Core Aspects of Language Comprehension: A Large-Scale fMRI Investigation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *40*(23), 4536–4550. https://doi.org/10.1523/JNEUROSCI.2036-19.2020

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Du, J., DiNicola, L. M., Angeli, P. A., Saadon-Grosman, N., Sun, W., Kaiser, S., Ladopoulou, J., Xue, A., Yeo, B. T. T., Eldaief, M. C., & Buckner, R. L. (2024). Organization of the human cerebral cortex estimated within individuals: Networks, global topography, and function. *Journal of Neurophysiology*, *131*(6), 1014–1082. https://doi.org/10.1152/jn.00308.2023

Duncan, J. (2010a). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Duncan, J. (2010b). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Duncan, J., Assem, M., & Shashidhara, S. (2020). Integrated Intelligence from Distributed Brain Activity. *Trends in Cognitive Sciences*, *24*(10), 838–852. https://doi.org/10.1016/j.tics.2020.06.012

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475–483. https://doi.org/10.1016/S0166-2236(00)01633-7

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). *Faith and Fate: Limits of Transformers on Compositionality* (arXiv:2305.18654). arXiv. https://doi.org/10.48550/arXiv.2305.18654

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (2020). *DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning* (arXiv:2006.08381). arXiv. https://doi.org/10.48550/arXiv.2006.08381

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. https://doi.org/10.1037/0096-3445.128.3.309

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16428–16433. https://doi.org/10.1073/pnas.1112937108

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110

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

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, *25*(5), 289–312. https://doi.org/10.1038/s41583-024-00802-4

Fedorenko, E., & Shain, C. (2021). Similarity of computations across domains does not imply shared implementation: The case of language comprehension. *Current Directions in Psychological Science*, *30*(6), 526–534. https://doi.org/10.1177/09637214211046955

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*(34), E5072–E5081. https://doi.org/10.1073/pnas.1610344113

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*(1), 87–104. https://doi.org/10.1111/nyas.12406

Flanagan, D. P., & Harrison, P. L. (Eds.). (2012a). *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed). Guilford Press.

Flanagan, D. P., & Harrison, P. L. (Eds.). (2012b). *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed). Guilford Press.

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/0198236360.001.0001

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind." *Trends in Cognitive Sciences*, *7*(2), 77–83. https://doi.org/10.1016/S1364-6613(02)00025-6

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought* (pp. x, 382). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195154061.001.0001

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Geschwind, N. (1970). The organization of language and the brain. *Science (New York, N.Y.)*, *170*(3961), 940–944. https://doi.org/10.1126/science.170.3961.940

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*(2), 127–171. https://doi.org/10.1016/0010-0277(92)90060-U

Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and Asymmetries in the Neurobiological Infrastructure for Language Production and Comprehension. *Cerebral Cortex (New York, N.Y.: 1991)*, *32*(7), 1405–1418. https://doi.org/10.1093/cercor/bhab287

Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. *Trends in Cognitive Sciences*, *21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013

Goddu, M. K., & Gopnik, A. (2024). The Development of Human Causal Learning and Reasoning. *Nature Reviews Psychology*, *3*, 319–339.

Goodglass, H. (1993). *Understanding aphasia* (pp. xii, 297). Academic Press.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). *Concepts in a Probabilistic Language of Thought* [Technical Report]. Center for Brains, Minds and Machines (CBMM). https://dspace.mit.edu/handle/1721.1/100174

Gopnik, A. (1982). Words and plans: Early language and the development of intelligent action. *Journal of Child Language*, *9*(2), 303–318. https://doi.org/10.1017/s0305000900004736

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, *111*(1), 3–32. https://doi.org/10.1037/0033-295X.111.1.3

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Grice, H. P. (1968). Utterer's Meaning, Sentence-Meaning, and Word-Meaning. *Foundations of Language*, *4*(3), 225–242.

Grice, H. P. (1975). Logic and Conversation. In D. Davidson (Ed.), *The logic of grammar* (pp. 64–75). Dickenson Pub. Co.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, *73*(3), 407–420. https://doi.org/10.1111/j.2044-8295.1982.tb01823.x

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, *366*(6461), 55–58. https://doi.org/10.1126/science.aax0289

Harlow, J. (1848). Passage of an Iron Rod through the Head. *The Boston Medical and Surgical Journal*, *39*(20), 389–393. https://doi.org/10.1056/NEJM184812130392001

Harlow, J. M. (1869). *Recovery from the Passage of an Iron Bar Through the Head*. D. Clapp.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, *11*(7), 299–306. https://doi.org/10.1016/j.tics.2007.05.001

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (arXiv:2009.03300). arXiv. https://doi.org/10.48550/arXiv.2009.03300

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). *Measuring Mathematical Problem Solving With the MATH Dataset* (arXiv:2103.03874). arXiv. https://doi.org/10.48550/arXiv.2103.03874

Hiersche, K. J., Schettini, E., Li, J., & Saygin, Z. M. (2024). Functional dissociation of the language network and other cognition in early childhood. *Human Brain Mapping*, *45*(9), e26757. https://doi.org/10.1002/hbm.26757

Hofstadter, D. R., & Mitchell, M. (1994). The Copycat project: A model of mental fluidity and analogy-making. In *Analogical connections* (pp. 31–112). Ablex Publishing.

Hugdahl, K., Raichle, M. E., Mitra, A., & Specht, K. (2015). On the existence of a generalized non-specific task-dependent network. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00430

Ivanova, A. A., Srikant, S., Sueoka, Y., Kean, H. H., Dhamala, R., O'Reilly, U.-M., Bers, M. U., & Fedorenko, E. (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife*, *9*, e58906. https://doi.org/10.7554/eLife.58906

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and Syntactic Ambiguity Resolution in Broca's Area. *Journal of Cognitive Neuroscience*, *21*(12), 2434–2444. https://doi.org/10.1162/jocn.2008.21179

Johnson-Laird, P. N. (1975). Models of Deduction. In *Reasoning: Representation and Process*. Psychology Press.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge University Press.

Johnson-Laird, P., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*(2), 134–148. https://doi.org/10.1016/0010-0285(70)90009-5

Jouravlev, O., Zheng, D., Balewski, Z., Le Arnz Pongos, A., Levan, Z., Goldin-Meadow, S., & Fedorenko, E. (2019). Speech-accompanying gestures are not processed by the language-processing mechanisms. *Neuropsychologia*, *132*, 107132. https://doi.org/10.1016/j.neuropsychologia.2019.107132

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, *60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science (New York, N.Y.)*, *274*(5284), 114–116. https://doi.org/10.1126/science.274.5284.114

Kaan, E., & Swaab, T. Y. (2002). The brain circuitry of syntactic comprehension. *Trends in Cognitive Sciences*, *6*(8), 350–356. https://doi.org/10.1016/S1364-6613(02)01947-2

Kean, H. H., Fung, A., Pramod, R. T., Chomik-Morales, J., Kanwisher, N., & Fedorenko, E. (2025). Intuitive physical reasoning is not mediated by linguistic nor exclusively domain-general abstract representations. *Neuropsychologia*, *213*, 109125. https://doi.org/10.1016/j.neuropsychologia.2025.109125

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of Abnormal Psychology*, *126*(6), 726–738. https://doi.org/10.1037/abn0000273

Knauff, M. (2013). *Space to reason: A spatial theory of human thought* (pp. xvii, 290). The MIT Press.

Knauff, M., Mulack, T., Kassubek, J., Salih, H. R., & Greenlee, M. W. (2002). Spatial imagery in deductive reasoning: A functional MRI study. *Brain Research. Cognitive Brain Research*, *13*(2), 203–212. https://doi.org/10.1016/s0926-6410(01)00116-1

Kowalski, R., & Sergot, M. (1986). A logic-based calculus of events. *New Generation Computing*, *4*(1), 67–95. https://doi.org/10.1007/BF03037383

Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, *1243*, 86–103. https://doi.org/10.1016/j.brainres.2008.07.128

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lepori, M. A., Pavlick, E., & Serre, T. (2023). *NeuroSurgeon: A Toolkit for Subnetwork Analysis* (arXiv:2309.00244). arXiv. https://doi.org/10.48550/arXiv.2309.00244

Lepori, M., Serre, T., & Pavlick, E. (2023). Break it down: Evidence for structural compositionality in neural networks. *Advances in Neural Information Processing Systems*, *36*, 42623–42660.

Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., & Choi, Y. (2024). *WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild* (arXiv:2406.04770). arXiv. https://doi.org/10.48550/arXiv.2406.04770

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., Jouravlev, O., Rakocevic, L., Pritchett, B., Siegelman, M., Hoeflin, C., Pongos, A., Blank, I. A., Struhl, M. K., Ivanova, A., Shannon, S., Sathe, A., Hoffmann, M., Nieto-Castañón, A., & Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, *9*(1), 529. https://doi.org/10.1038/s41597-022-01645-3

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., & Zhang, Y. (2020). *LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning* (arXiv:2007.08124). arXiv. https://doi.org/10.48550/arXiv.2007.08124

Liu, Y.-F., Kim, J., Wilson, C., & Bedny, M. (2020). Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *eLife*, *9*, e59340. https://doi.org/10.7554/eLife.59340

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014–1019. https://doi.org/10.1038/s41593-022-01114-5

Malik-Moraleda, S., Taliaferro, M., Shannon, S., Jhingan, N., Swords, S., Peterson, D. J., Frommer, P., Okrand, M., Sams, J., Cardwell, R., Freeman, C., & Fedorenko, E. (2025). Constructed languages are processed by the same brain mechanisms as natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(12), e2313473122. https://doi.org/10.1073/pnas.2313473122

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science* (pp. xiii, 224). The MIT Press.

Masís-Obando, R., Norman, K. A., & Baldassano, C. (2022). Schema representations in distinct brain networks support narrative memory during encoding and retrieval. *eLife*, *11*, e70445. https://doi.org/10.7554/eLife.70445

McCarthy, J., & Hayes, P. (1969). Some Philosophical Problems From the Standpoint of Artificial Intelligence. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence 4* (pp. 463–502). Edinburgh University Press.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve* (arXiv:2309.13638). arXiv. https://doi.org/10.48550/arXiv.2309.13638

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared language: Overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40–48. https://doi.org/10.1016/j.cognition.2016.05.012

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2009). The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences*, *106*(30), 12554–12559. https://doi.org/10.1073/pnas.0902422106

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought Beyond Language: Neural Dissociation of Algebra and Natural Language. *Psychological Science*, *23*(8), 914–922. https://doi.org/10.1177/0956797612437427

Nezhurina, M., Cipolina-Kun, L., Cherti, M., & Jitsev, J. (2025). *Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models* (arXiv:2406.02061). arXiv. https://doi.org/10.48550/arXiv.2406.02061

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective & Behavioral Neuroscience*, *12*(2), 241–268. https://doi.org/10.3758/s13415-011-0083-5

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2014). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language, Cognition and Neuroscience*, *29*(2), 186–217. https://doi.org/10.1080/01690965.2012.758297

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective & Behavioral Neuroscience*, *5*(3), 263–281. https://doi.org/10.3758/cabn.5.3.263

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., & Odena, A. (2021). *Show Your Work: Scratchpads for Intermediate Computation with Language Models* (arXiv:2112.00114). arXiv. https://doi.org/10.48550/arXiv.2112.00114

Olausson, T., Gu, A., Lipkin, B., Zhang, C., Solar-Lezama, A., Tenenbaum, J., & Levy, R. (2023). LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5153–5176). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.313

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., … Li, Z. (2024). *OpenAI o1 System Card* (arXiv:2412.16720). arXiv. https://doi.org/10.48550/arXiv.2412.16720

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

Pan, L., Albalak, A., Wang, X., & Wang, W. (2023). Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 3806–3824). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.248

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–681. https://doi.org/10.1038/nn1082

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *Journal of Neurophysiology*, *121*(4), 1244–1265. https://doi.org/10.1152/jn.00619.2018

Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., & Fedorenko, E. (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*(3), 413–440. https://doi.org/10.1162/nol_a_00071

Piaget, J. (1926). *The Language and Thought of the Child* (M. Gabain, Trans.). Kegan Paul, Trench, Trubner & Co.

Pinker, S. (1994). *The language instinct* (p. 494). William Morrow & Co.

Pinto, J., & Reiter, R. (1993). *Temporal Reasoning in Logic Programming: A Case for the Situation Calculus*. https://doi.org/10.7551/mitpress/4305.003.0023

Poesia, G., Gandhi, K., Zelikman, E., & Goodman, N. D. (2023). *Certified Deductive Reasoning with Language Models* (arXiv:2306.04031). arXiv. https://doi.org/10.48550/arXiv.2306.04031

Pramod, R., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *eLife*, *11*, e71736. https://doi.org/10.7554/eLife.71736

Pramod, R. T., Mieczkowski, E., Fang, C. X., Tenenbaum, J. B., & Kanwisher, N. (2025). Decoding predicted future states from the brain's "physics engine." *Science Advances*, *11*(22), eadr7429. https://doi.org/10.1126/sciadv.adr7429

Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences*, *20*(5), 446–451. https://doi.org/10.1016/j.lindif.2010.05.001

Pritchett, B. L., Hoeflin, C., Koldewyn, K., Dechter, E., & Fedorenko, E. (2018). High-level language processing regions are not engaged in action observation or imitation. *Journal of Neurophysiology*, *120*(5), 2555–2570. https://doi.org/10.1152/jn.00222.2018

Qiu, L., Jiang, L., Lu, X., Sclar, M., Pyatkin, V., Bhagavatula, C., Wang, B., Kim, Y., Choi, Y., Dziri, N., & Ren, X. (2023, October 13). *Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement*. The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=bNt7oajl2a

Quillen, I. A., Yen, M., & Wilson, S. M. (2021). Distinct Neural Correlates of Linguistic and Non-Linguistic Demand. *Neurobiology of Language*, *2*(2), 202–225. https://doi.org/10.1162/nol_a_00031

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, *41*(1), 1–48. https://doi.org/10.1006/cogp.1999.0735

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark* (arXiv:2311.12022). arXiv. https://doi.org/10.48550/arXiv.2311.12022

Reverberi, C., Cherubini, P., Rapisarda, A., Rigamonti, E., Caltagirone, C., Frackowiak, R. S. J., Macaluso, E., & Paulesu, E. (2007). Neural basis of generation of conclusions in elementary deduction. *NeuroImage*, *38*(4), 752–762. https://doi.org/10.1016/j.neuroimage.2007.07.060

Rips, L. J. (1983). Cognitive Processes in Propositional Reasoning. *Psychological Review*, *90*(1), 38–71. https://doi.org/10.1037/0033-295x.90.1.38

Rips, L. J. (1988). Deduction. In *The psychology of human thought* (pp. 116–152). Cambridge University Press.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

Rips, L. J. (1995). Deduction and cognition. In *Thinking: An invitation to cognitive science, Vol. 3, 2nd ed* (pp. 297–343). The MIT Press.

Rogalsky, C., Rong, F., Saberi, K., & Hickok, G. (2011). Functional anatomy of language and music perception: Temporal and structural factors investigated using functional magnetic resonance imaging. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(10), 3843–3852. https://doi.org/10.1523/JNEUROSCI.4515-10.2011

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Rule, J. S. (n.d.). *The child as hacker: Building more human-like models of learning*.

Rule, J. S., Piantadosi, S. T., Cropper, A., Ellis, K., Nye, M., & Tenenbaum, J. B. (2024). Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications*, *15*(1), 6847. https://doi.org/10.1038/s41467-024-50966-x

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/s1053-8119(03)00230-1

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures* (p. 248). Lawrence Erlbaum.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain. *eLife*, *8*, e46619. https://doi.org/10.7554/eLife.46619

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, *20*, 11–21. https://doi.org/10.1136/jnnp.20.1.11

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust Effects of Working Memory Demand during Naturalistic Language Comprehension in Language-Selective Cortex. *Journal of Neuroscience*, *42*(39), 7412–7430. https://doi.org/10.1523/JNEUROSCI.1894-21.2022

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

Shain, C., Kean, H., Casto, C., Lipkin, B., Affourtit, J., Siegelman, M., Mollica, F., & Fedorenko, E. (2024). Distributed Sensitivity to Syntax and Semantics throughout the Language Network. *Journal of Cognitive Neuroscience*, *36*(7), 1427–1471. https://doi.org/10.1162/jocn_a_02164

Shain, C., Paunov, A., Chen, X., Lipkin, B., & Fedorenko, E. (2023). No evidence of theory of mind reasoning in the human language network. *Cerebral Cortex*, *33*(10), 6299–6319. https://doi.org/10.1093/cercor/bhac505

Shojaee*†, P., Mirzadeh*, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf

Sneha, S., Daniel, J. M., Yaara, E., & John, D. (2019). Progressive Recruitment of the Frontoparietal Multiple-demand System with Increased Task Complexity, Time Pressure, and Reward. *Journal of Cognitive Neuroscience*, *31*(11), 1617–1630. https://doi.org/10.1162/jocn_a_01440

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell.

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, *21*(3), 489–510. https://doi.org/10.1162/jocn.2008.21029

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (arXiv:2206.04615). arXiv. https://doi.org/10.48550/arXiv.2206.04615

Tenenbaum, J. B. (Joshua B. (1999). *A Bayesian framework for concept learning* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/16714

Tervo-Clemmens, B., Calabro, F. J., Parr, A. C., Fedor, J., Foran, W., & Luna, B. (2023). A canonical trajectory of executive function maturation from adolescence to adulthood. *Nature Communications*, *14*(1), 6922. https://doi.org/10.1038/s41467-023-42540-8

Thompson-Schill, S. L. (2005). Dissecting the Language Organ: A New Look at the Role of Broca's Area in Language Processing. In *Twenty-first century psycholinguistics: Four cornerstones* (pp. 173–189). Lawrence Erlbaum Associates Publishers.

Tyler, L. K., Marslen-Wilson, W. D., Randall, B., Wright, P., Devereux, B. J., Zhuang, J., Papoutsi, M., & Stamatakis, E. A. (2011). Left inferior frontal cortex and syntax: Function, structure and behaviour in patients with left hemisphere damage. *Brain*, *134*(2), 415–431. https://doi.org/10.1093/brain/awq369

Varley, R. A., Klessinger, N. J. C., Romanowski, C. A. J., & Siegal, M. (2005). Agrammatic but numerate. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(9), 3519–3524. https://doi.org/10.1073/pnas.0407470102

Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and "theory of mind" in an agrammatic aphasic patient. *Current Biology: CB*, *10*(12), 723–726. https://doi.org/10.1016/s0960-9822(00)00538-8

Wason, P. C. (1966). Reasoning. In P. C. Wason (Ed.), *New Horizons in Psychology* (pp. 135–151). Penguin Books.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content* (p. 264). Harvard U. Press.

Wechsler, D. (2011). *WASI-II - Wechsler Abbreviated Scale of Intelligence | Second Edition | Pearson Assessments US*. https://www.pearsonassessments.com/en-us/Store/Professional-Assessments/Cognition-%26-Neuro/Wechsler-Abbreviated-Scale-of-Intelligence-%7C-Second-Edition/p/100000593

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cerebral Cortex (New York, N.Y.: 1991)*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

Weiss, L. G., Saklofske, D. H., Holdnack, J. A., & Prifitera, A. (2016). *WISC-V assessment and interpretation: Scientist-practitioner perspectives*. Elsevier/AP, Academic Press is an imprint of Elsevier.

Willems, R. M., Benn, Y., Hagoort, P., Toni, I., & Varley, R. (2011). Communicating without a functioning language system: Implications for the role of language in mentalizing. *Neuropsychologia*, *49*(11), 3130–3135. https://doi.org/10.1016/j.neuropsychologia.2011.07.023

Wilson, S. M., Entrup, J. L., Schneck, S. M., Onuscheck, C. F., Levy, D. F., Rahman, M., Willey, E., Casilio, M., Yen, M., Brito, A. C., Kam, W., Davis, L. T., de Riesthal, M., & Kirshner, H. S. (2022). Recovery from aphasia in the first year after stroke. *Brain*, *146*(3), 1021–1039. https://doi.org/10.1093/brain/awac129

Wittgenstein, L. (1921). *Tractatus Logico-Philosophicus (Trans. Pears and Mcguinness)* (L. Bazzocchi & P. M. S. Hacker, Eds.). Routledge.

Woolgar, A., Bor, D., & Duncan, J. (2013). Global Increase in Task-related Fronto-parietal Activity after Focal Frontal Lobe Lesion. *Journal of Cognitive Neuroscience*, *25*(9), 1542–1552. https://doi.org/10.1162/jocn_a_00432

Woolgar, A., Duncan, J., Manes, F., & Fedorenko, E. (2018). The multiple-demand system but not the language system supports fluid intelligence. *Nature Human Behaviour*, *2*(3), 200–204. https://doi.org/10.1038/s41562-017-0282-3

Woolgar, A., Parr, A., Cusack, R., Thompson, R., Nimmo-Smith, I., Torralva, T., Roca, M., Antoun, N., Manes, F., & Duncan, J. (2010). Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *107*(33), 14899–14902. https://doi.org/10.1073/pnas.1007928107

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Ye, X., Chen, Q., Dillig, I., & Durrett, G. (2023). SatLM: Satisfiability-Aided Language Models Using Declarative Prompting. *Advances in Neural Information Processing Systems*, *36*, 45548–45580.

Zelazo, P. D., & Carlson, S. M. (2020). The neurodevelopment of executive function skills: Implications for academic achievement gaps. *Psychology & Neuroscience*, *13*(3), 273–298. https://doi.org/10.1037/pne0000208

Zimmerer, V. C., Cowell, P. E., & Varley, R. A. (2014). Artificial grammar learning in individuals with severe aphasia. *Neuropsychologia*, *53*, 25–38. https://doi.org/10.1016/j.neuropsychologia.2013.10.014

Zorowitz, S., Chierchia, G., Blakemore, S.-J., & Daw, N. D. (2024). An item response theory analysis of the matrix reasoning item bank (MaRs-IB). *Behavior Research Methods*, *56*(3), 1104–1122. https://doi.org/10.3758/s13428-023-02067-8