COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13448
© 2024 The Authors. *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).
ISSN: 1551-6709 online
DOI: 10.1111/cogs.13448

Learning the Meanings of Function Words From Grounded Language Using a Visual Question Answering Model

Eva Portelance,[a,b] Michael C. Frank,[c] Dan Jurafsky[d]

[a]*Department of Linguistics, McGill University*
[b]*Mila - Quebec Artificial Intelligence Institute*
[c]*Department of Psychology, Stanford University*
[d]*Department of Linguistics, Stanford University*

Received 15 August 2023; received in revised form 5 April 2024; accepted 12 April 2024

Abstract

Interpreting a seemingly simple function word like "or," "behind," or "more" can require logical, numerical, and relational reasoning. How are such words learned by children? Prior acquisition theories have often relied on positing a foundation of innate knowledge. Yet recent neural-network-based visual question answering models apparently can learn to use function words as part of answering questions about complex visual scenes. In this paper, we study what these models learn about function words, in the hope of better understanding how the meanings of these words can be learned by both models and children. We show that recurrent models trained on visually grounded language learn gradient semantics for function words requiring spatial and numerical reasoning. Furthermore, we find that these models can learn the meanings of logical connectives *and* and *or* without any prior knowledge of logical reasoning as well as early evidence that they are sensitive to alternative expressions when interpreting language. Finally, we show that word learning difficulty is dependent on the frequency of models' input. Our findings offer proof-of-concept evidence that it is possible to learn the nuanced interpretations of function words in a visually grounded context by using non-symbolic general statistical learning algorithms, without any prior knowledge of linguistic meaning.

Keywords: Function word acquisition; Visually-grounded language; Visual question answering; Multimodal statistical learning; Reasoning skills; Logical reasoning; Neural network models; Proof of concept

Correspondence should be sent to Eva Portelance, Mila - Quebec Artificial Intelligence Institute, 6666 Rue Saint-Urbain, Montreal, QC H2S 3H1, Canada. E-mail: eva.portelance@mila.quebec

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

When studying how children learn words, researchers often make the assumption that knowing the meaning of a word *w* means having the ability to differentiate between things that are *w* and things that are not (Bloom, 2002, ch. 1). This notion of meaning, sometimes called "external meaning" is in contrast to "internal meaning"—the mental representation of meaning that a person has for *w*—the favored definition of meaning in theoretical semantics. Evaluating children's ability to understand the meaning of words by how they use them in the external world seems pretty straightforward in the case of nouns and predicates, but not so much for function words, like determiners, conjunctions, and prepositions. These closed-class words tend to have external meanings that only manifest themselves in how they modify other words or sentences as a whole, making them difficult to study without referring in some way to their internal meaning. Additionally, parsing their meaning often requires complex reasoning skills such as logical, numerical, spatial, or relational reasoning. The abstract nature and complexity of function words are what make their acquisition by children so difficult to study using conventional methods. Yet, these same qualities are also what make function words an ideal test case to compare different theories of language acquisition and their respective learning strategies.

It has been widely observed that children tend to acquire words and grammatical structures in a specific order; this is also the case for function words. For example, *and* is much more prevalent in children's linguistic input and is acquired before *or* (Jasbi, Jaggi, & Frank, 2018; Morris, 2008); Children start to correctly use the preposition *behind* before they do *in front of* and furthermore, their initial uses of these words are possibly conditioned on contextual factors like whether the referent object has the property of having a front and back, like a car or a doll (E. V. Clark, 1977; Kuczaj & Maratsos, 1975; Windmiller, 1973), and the degree of occlusion between two objects (Grigoroglou, Johanson, & Papafragou, 2019; Johnston, 1984). These differences in order of acquisition represent learning outcomes that can be used as test cases to study the impact of different types of information available in the input on learners' ability to acquire these words.

Theories for the acquisition of function words tend to fall somewhere along the spectrum between nativist explanations—for example, logical nativism (Crain, 2012)—and usage-based approaches (Tomasello, 2005). Nativist theories posit that humans are endowed with innate knowledge of some reasoning skills and that children may undergo a series of maturational stages, to reach adult-like understanding. These stage-based and symbolic learning explanations predict that conceptual differences between words may lead to asymmetries in their acquisition. On the other hand, usage-based approaches argue that the reasoning skills necessary for understanding function words are learned through experience. Children learn these words using non-symbolic general learning mechanisms which are not exclusive to language acquisition. Usage-based learning mechanisms specifically predict that frequency of exposure is a primary factor in determining the order in which new words may be learned. While frequency may also play a role in nativist theories, it is often posited to be secondary to other conceptual differences.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1.1. The current study

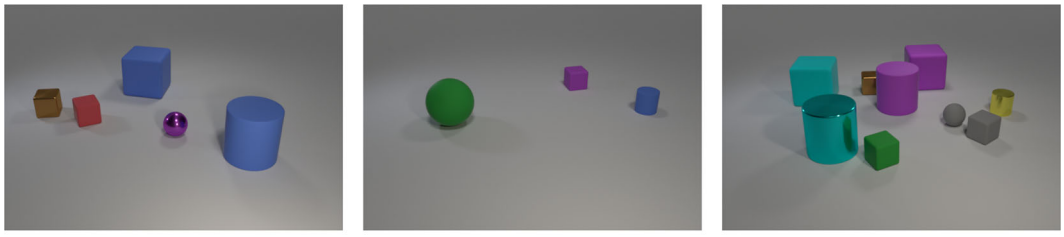
In this paper, we will consider the acquisition of three pairs of function words and their respective reasoning skills: (1) logical reasoning with the connectives *or* and *and*; (2) spatial reasoning with the prepositions *in front of* and *behind*; (3) numerical reasoning with the scalar quantifiers *more* and *fewer*. We hypothesize that these reasoning pairs can be learned using non-symbolic general learning algorithms and, furthermore, that the ordering effects seen in children's acquisition of these words are simply the result of their frequency in children's input, rather than evidence for non-symbolic or stage-based learning strategies. We propose to use computational models that learn these types of words from grounded input to test these hypotheses.

We propose to use a new modeling approach, which considers models as independent learners—in other words, like a new "species" of language learners—that can be leveraged to implement "proofs of concept" (Lappin, 2021, ch. 1.2; Pearl, 2023; Portelance & Jasbi, 2023; Tsuji, Cristia, & Dupoux, 2021; Warstadt & Bowman, 2023). A proof of concept can show us what is learnable "in practice" for models and "in principle" for humans. In doing so, models may be used to inform debates about the relative innateness of certain linguistic knowledge (A. Clark & Lappin, 2011). This approach draws on recent model interpretability work showing what kinds of grammatical knowledge language models learn (Futrell et al., 2019; J. Hu, Gauthier, Qian, Wilcox, & Levy, 2020; Lake & Baroni, 2018; Linzen, Dupoux, & Goldberg, 2016; Manning, Clark, Hewitt, Khandelwal, & Levy, 2020), and what kinds of learning biases they have (Papadimitriou & Jurafsky, 2023). With our experiments, we hope to offer proof of concept evidence showing what is in practice learnable from visually grounded language on the meaning of abstract function words requiring complex reasoning skills.

We use neural network models that learn from both linguistic and visual representations to study the effect of visual grounding on learning the meaning of function words. We can consider the interactions that may emerge from cross-modal statistical word learning, an open question developmentalists are still tackling (Saffran & Kirkham, 2018). Specifically, we experiment with neural network models learning a language in a visual question answering task, where they must come up with word representations in order to answer questions about visual scenes. The task we use is called the CLEVR (Compositional Language and Elementary Visual Reasoning) dataset (Johnson et al., 2017). It contains visual block-world scenes and corresponding questions like "Are there more red cubes than metal spheres?" Models are never given the meaning of words or any form of mapping between words and the content of images. They must deduce this information during training. Learning the meaning of words then becomes an auxiliary objective that can lead models to successfully complete their task: to generate the correct answer given some string and an image (examples from the task are given in Fig. 1).

In order to propose that a neural network learner offers additional proof that some outcome—the meaning of function words—is likely learnable in humans, it is insufficient to just show that the models can learn this outcome; we must also weigh in on what might have led the model to learn it in the first place and acknowledge that the proposed prerequisites for learning the outcome must also be available to human learners (Baroni, 2021). Furthermore, we must outline the learning assumptions on which our proof-of-concept depends.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Q: Are there more brown shiny objects behind the large rubber cylinder than gray blocks?
A: yes

Q: What color is the matte object to the right of the large block?
A: blue

Q: Are there any other things that are the same shape as the large green thing?
A: no

Q: The matte thing that is both in front of the purple cube and to the left of the blue rubber cylinder is what color?
A: green

Q: What number of cubes are the same color as the rubber ball?
A: 1

Q: Is the big cyan metal thing the same shape as the brown thing?
A: no

Fig. 1. Example images and corresponding questions taken from the CLEVR dataset.

The learning mechanisms used by visual question answering models are almost certainly different from those used by children, but they do share one high-level property: the use of indirect negative evidence. Early work suggested that children do not make use of any explicit negative evidence for word learning (Baker, 1979; Fodor & Crowther, 2002; Marcus, 1993; Pinker, 1989). However, many researchers have shown that they do rely on implicit negative evidence (R. Brown, 1970; Chouinard & Clark, 2003; A. Clark & Lappin, 2011; Farrar, 1992; Penner, 1987; Saxton, 1997; Snow & Ferguson, 1977), for example, when their desired outcomes are not met when they are misunderstood. The meanings of words may then be learned indirectly from this evidence, and the same may be said for our models. Though visual question answering models receive direct supervision on their training task—generating correct answers to questions—they do not receive direct supervision to learn abstract reasoning or the meanings of function words; these learning outcomes are incidental to the task and instead could be one of many strategies that models converge towards to answer the questions correctly. Our proof of concepts is thus conditional on the availability of some form of supervision—direct or indirect—during learning.

Indeed, we are not the first to make these assumptions. Visual question answering models have already been used to explore neural networks' capacity to learn meaningful representations of referential words, such as nouns and predicates when trained on language tasks grounded in the visual world (Jiang et al., 2023; Mao, Gan, Kohli, Tenenbaum, & Wu, 2019; Pillai, Matuszek, & Ferraro, 2021; Wang, Mao, Gershman, & Wu, 2021; Zellers et al., 2021). As for function words, Hill, Hermann, Blunsom, and Clark (2018) briefly consider how visually grounded models learn negation, and Kuhnle and Copestake (2019) studied how these models interpret the quantifier *most*. Regier's (1996) earlier extensive work also considered how neural network models can learn to map visual scenes to spatial prepositions, though his models did not learn from any linguistic input per se and predate visual question answering models. Others more recently have also used these tasks to model noun and predicate learning in children (Hill, Clark, Blunsom, & Hermann, 2020; Nikolaus & Fourtassi, 2021). However, to the best of our knowledge, no work has probed visually

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

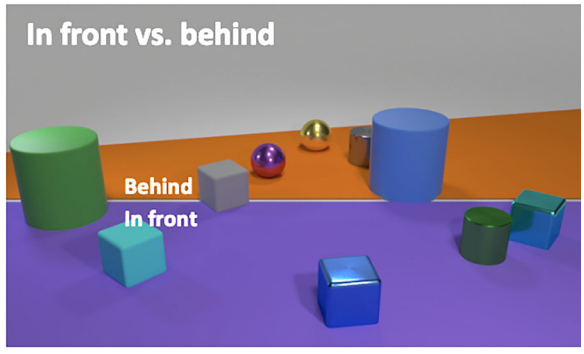


Fig. 2. Threshold-based interpretation of *behind* and *in front* relative to gray cube in CLEVR dataset. (Image from Johnson et al. 2017.)

grounded neural network models' representations of the meaning of function words in the context of children's function word learning.

Throughout this paper, we will address three major research questions:

1. How do visually grounded question answering models learn to represent and interpret function words and do these representations generalize to unseen linguistic and visual contexts?
2. Does the existence of alternative expressions in each reasoning pair affect their acquisition or are the meanings of function words acquired in isolation?
3. Do models learn these function words in a similar order to children and are these ordering effects the results of their frequency or do they follow from other conceptual explanations?

With respect to our first research question, each of our function words of interest is defined in absolute terms and mapped to a function over predicates in the CLEVR dataset we use. For example *or* is defined as the logical operator, $A \vee B$, *more* is defined as the function greater than, $|A| > |B|$, and *behind* is defined as having a y-coordinate that is strictly greater than some other referent's, as in Fig. 2. In practice however, most of these words have much more gradient meanings when used by people in naturalistic contexts. The use of language in context distinguishes semantic representations from pragmatic interpretations. We probe how models interpret these words in novel contexts to determine how their meanings may be represented. Do their interpretations suggest that they have clear-cut thresholds that distinguish the meaning of words like *more* and *fewer* or does linguistic gradience arise as a result of their learning environment when exposed to grounded language use in continuous visual settings? In the CLEVR dataset, the underlying meanings of words like *more* and *behind* are threshold-based. So, the statement "there are more As than Bs" is always interpreted as true as long as the difference between $|A|$ and $|B|$ is over some threshold, here $|A| - |B| > 0$. Linguistic gradience, on the other hand, may be thought of as allowing words to have different interpretations depending on context as a function of some gradient factor. So instead, we may expect our statement "there are more As than Bs" to be interpreted as true or false as a function of

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the magnitude of the difference $|A| - |B|$ across contexts rather than based on some context-agnostic threshold. If models can learn representations that lead to gradient interpretations in novel contexts by using simple learning algorithms, then we can offer proof-of-concept evidence that these function words are learnable from supervised data using non-symbolic learning mechanisms.

With respect to our second research question, the existence of alternative expressions or worlds is the cornerstone behind Gricean pragmatic reasoning, and what allows us to have different interpretations of the same word in different contexts (Degen, 2023; Grice, 1975). Children have been found to exhibit pragmatic reasoning skills in multiple domains, especially when alternative worlds are made salient (Baharloo, Vasil, Ellwood-Lowe, & Srinivasan, 2023; Barner, Brooks, & Bale, 2011; Horowitz & Frank, 2016; Katsos & Bishop, 2011; Stiller, Goodman, & Frank, 2015). It is, however, unclear if this ability is acquired through specific means. Following Gricean's theory, we might expect children to be able to judge the informativeness of contrasting expressions as soon as they have learned their meaning (E. V. Clark, 2003; Katsos & Bishop, 2011), suggesting that these abilities may stem from the same learning mechanisms. If visual question answering models can learn to consider alternative expressions when interpreting function words like *and* and *or* in novel contexts, then we may offer proof of concept evidence that the ability to reason about alternatives can be derived from a statistical learning mechanism applied in a contextually grounded setting.

With respect to our third research question, frequency or word predictability is a known predictor of the order in which children acquire words (Braginsky, Yurovsky, Marchman, & Frank, 2019; Goodman, Dale, & Li, 2008; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Portelance, Duan, Frank, & Lupyan, 2023). There may, however, be other factors—in relation to or independent from—frequency that makes learning the meaning of certain function words harder than others. For example, E. V. Clark (1993) points out that there seems to be an asymmetry in the acquisition of adjective pairs like *big* and *little*, *tall* and *short*, etc., where children tend to produce words for positive dimensions before they do negative ones. This difference in learning may be independent from frequency, since in experiments where children are exposed to nonsense word pairs like these with even frequency, they still seem to favor learning the positive words over the negative ones (Klatzky, Clark, & Macken, 1973). These results would then promote a conceptual explanation for these effects over a frequency-based explanation. Such asymmetries may also exist for similarly polarized pairs of function words. Here, we explore whether the order in which these words are learned is a function of how frequent they are in the input or if there may be other factors that make certain function words intrinsically more difficult to learn than others. We will compare the order in which children acquire words requiring similar abstract reasoning to the order in which visual question answering models learn these same words while varying their relative frequency in the models' input.

Our approach is as follows. We define a novel semantic testing task within the CLEVR block world to determine whether models understand the meanings of function words in unseen contexts. We then evaluate model performance on these novel tests throughout training to visualize how learning progresses. Next, we compare the relative order in which models learn our function words to the acquisition order we expect in children. We manipulate input

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

distributions and train models on different subsets of the training data with various function word frequencies to analyze whether the ordering effects initially observed are solely mediated by frequency or if other more conceptual factors play a role.[1] In the remainder of this introduction, we briefly review children's acquisition of function words and the visual question answering task and dataset we use.

1.2. Children's acquisition of the target function words

For each of the word pairs and their respective reasoning skills considered in this study ("and"/"or," "behind"/"in front of," "more"/"fewer"), we review what is currently known and debated about their acquisition in the child language learning literature. We note that most of the previous research on these words is exclusively about English, with a couple of exceptions, mentioned when relevant.

1.2.1. "and" / "or"

The source of the emergence of logical reasoning in children has been debated for quite some time (for a thorough review of the field, see Jasbi, 2018, Ch. 5). Proposals tend to fall somewhere along the spectrum between logical nativism (Crain, 2012) and usage-based approaches (Morris, 2008). Logical nativism posits that humans are endowed with innate logic and children then go through a series of developmental stages to reach adult-like logical understanding. As for usage-based approaches, these argue that logical reasoning is learned through experience using general learning mechanisms—as opposed to learning strategies that are specific to logical reasoning—and that frequency in children's input explains any ordering effects seen in children's learning of logical concepts.

All agree that children correctly interpret *and* before *or*; *and* is also much more frequent than *or* in children's input, and, furthermore, they are exposed to more instances of exclusive *or* than inclusive *or* (Jasbi et al., 2018; Morris, 2008). There is, however, some debate about the order in which children acquire possible meanings of *or* and what the underlying meaning of this logical connective may be in children's representations. Given its higher frequency, Morris (2008) suggests that children initially learn exclusive *or*. Similarly, early nativist approaches argued that children's early understanding of *or* was as a simple choice, making it compatible with exclusivity (Neimark, 1970). Following Grice's (1975) proposal that exclusive interpretations are the result of generalized conversational implicature, others have instead advocated that *or* is underlyingly inclusive and that children eventually learn exclusive *or* via pragmatic reasoning (Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Chierchia et al., 2004; Jasbi & Frank, 2021). Interestingly, some have also found the children often mistakenly interpret *or* as conjunction (Braine & Rumain, 1981; Singh, Wexler, Astle-Rahim, Kamawar, & Fox, 2016; Tieu et al., 2017), though it has been suggested that this finding may be an artifact to the specific experimental task designs used in these studies (Paris, 1973; Skordos, Feiman, Bale, & Barner, 2020).

All of the experimental results show that children understand *or* inclusively still leave unanswered the question of how they came to learn the meaning of this word in the first place. Crain (2008, 2012) argues that these results are in fact evidence in favor of a logical nativist

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

explanation since, though children are exposed to more instances of exclusive interpretations of *or*, they seem to instead favor inclusive interpretations initially. Currently, there is little evidence showing that inclusive *or* is learnable from more general learning mechanisms that would support a usage-based approach.

1.2.2. "Behind" / "in front of"

Children learn the meaning of the locative preposition *behind* before they do *in front of* (Johnston, 1984; Johnston & Slobin, 1979). There have been a few proposals for explaining this asymmetry, all sharing a common thread: that children do not initially encode the meaning of these words in geometric spatial terms. The semantic misanalysis hypothesis for the asymmetry in children's early understanding of these expressions suggests that children struggle to incorporate the perspective of the observer in analyzing the meanings of these words (Piaget & Inhelder, 1967), so they erroneously define the concepts of *front* and *back* in terms of visibility and occlusion (Johnston, 1984). Grigoroglou et al. (2019) also suggest that children analyze these words in terms of occlusion but not as a result of semantic misanalysis, instead as the result of pragmatic inference, where occlusion is more notable than visibility. Much of the research on the acquisition of *behind* and *in front of* then documents the stages of development between these early word representations and their adult-like geometric meanings. They conducted experiments in both English and Greek. Some researchers have found that this transition is aided by the eventual projection of the property of having a front or back on objects (e.g., being behind a doll vs. being behind a block) (E. V. Clark, 1977; Kuczaj & Maratsos, 1975; Windmiller, 1973). Again, there is currently a lack of evidence supporting the use of more general learning mechanisms behind the acquisition of these words, as opposed to learning strategies specific to spatial reasoning.

1.2.3. "More" / "fewer"

Quantifiers have been found to follow quite robust acquisition ordering effects cross-linguistically (Katsos et al., 2016, analysis over 30 languages). For the comparative quantifiers *more (than)* and *fewer (than)*, the meaning of *more* has repeatedly been found to be learned earlier than *fewer/less* by children (Donaldson & Balfour, 1968; Donaldson & Wales, 1970; Geurts, Katsos, Cummins, Moons, & Noordman, 2010; Palermo, 1973; Townsend, 1974). Some have also found that children initially interpret *less* as a synonym of *more* (Donaldson & Balfour, 1968; Palermo, 1973), but as Townsend (1974) points out, these earlier experimental studies did not have a way to truly distinguish between children interpreting *less* as *more* or simply not knowing the meaning of *less*. A few hypotheses have been put forward to explain the acquisition asymmetry between these two comparative quantifiers, all favoring conceptual explanations over frequency-based ones. Though Donaldson and Wales (1970) briefly mention that *more* is much more frequent than *less* in children's input, they quickly reject the possibility that frequency is the answer, arguing that if the asymmetry was down to frequency, we would expect children that do not know the meaning of *less* to interpret this word in a variety of ways. However, citing previous work, they suggest that *less* is instead always interpreted as *more*. They thus propose that there are a series of developmental stages for the processing of comparatives, which lead to this asymmetry, where *more* is acquired earlier

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

because children initially learn to use it in singular referent contexts like in the additive sense of *more*, for which they say a counterpart with *less* is not possible. H. H. Clark (2018) offers a similar proposal with slightly different developmental stages. Still, these results clearly suggest that word frequency might account for developmental ordering phenomena, consistent with usage-based accounts as well.

2. Evaluating function word knowledge using semantic probes

As a test bed for the learnability of function words, we use visual question answering models trained on the CLEVR dataset, a standard dataset used in the broader natural language processing community (Johnson et al., 2017). We propose a semantic probe zero-shot evaluation task based on CLEVR to determine whether models were able to learn meaningful representations for each of reasoning pairs under study: *and/or*, *behind/in front of*, and *more/fewer*.[2]

2.1. Visual question answering and the CLEVR dataset

Visual question answering was proposed as a language learning task that is grounded in images and requires models to develop abstract reasoning skills (Antol et al., 2015; Gao et al., 2015; Malinowski & Fritz, 2014; Ren, Kiros, & Zemel, 2015). Models are given images and questions about their content as input; they are then trained to answer these visually grounded questions (example image–question pairs from the CLEVR dataset are given in Fig. 1). Generating the correct answers often requires reasoning skills, such as logical reasoning, spatial reasoning, and numerical reasoning, which models must also learn. Since learning the meaning of function words requires developing these same reasoning skills, models trained to complete these types of tasks lend themselves well to the study of function word learning using neural networks.

Initial visual question answering tasks used datasets that were produced by having human annotators come up with questions for images (Antol et al., 2015; Gao et al., 2015; Krishna et al., 2017; Malinowski & Fritz, 2014). However, as the first resulting models emerged it became clear that they had shortcomings which prevented them from developing abstract reasoning, in part due to unbalanced datasets (Agrawal, Batra, & Parikh, 2016; Zhang, Goyal, Summers-Stay, Batra, & Parikh, 2016). To avoid this problem and to help parse which reasoning skills models were developing and relying on, balanced datasets with explicit generative models to produce questions (Hudson & Manning, 2019; Johnson et al., 2017) and images (Johnson et al., 2017) were created. CLEVR is one such dataset, containing generated images of scenes from a three-dimensional (3D) block-world and constructed questions.

We chose this dataset as it offered us the benefit of precisely defining the function words in the dataset by associating them to explicit functional relations, giving us a better grasp over the underlying semantics of these words. For these reasons, the CLEVR dataset (Johnson et al., 2017) serves as a good starting point for our comparison between the Visual question answering (VQA) model and children's acquisition of function words. Specifically, as mentioned in

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the introduction, *or* is defined as the inclusive logical operator, $A \vee B$, while *and* is $A \wedge B$; *more* is defined as greater than, $|A| > |B|$, while *fewer* is $|A| < |B|$; and *behind* is defined as having a y-coordinate that is strictly greater than some other referent's, $y(a) > y(b)$, while *in front* does the opposite $y(a) < y(b)$, as in Fig. 2. Additionally, it is a well-balanced dataset whose composition has been extensively described and well understood (Johnson et al., 2017).

It is composed of questions paired with images like those illustrated in Fig. 1. The images are all of complex scenes in a block-world involving static objects placed on a 3D gray plane. Objects have four varying attributes: shape, color, material, and size. The number of objects in an image varies randomly between 3 and 10, as do their relative positions and the positions of light sources in the scenes. There are a total of 70,000 distinct images in the training set and another 15,000 different images in the validation set.[3]

Each image is paired with a set of questions like those in Fig. 1. In total there are 699,989 questions in the training set and 149,991 in the validation set. There are different types of questions, including existential questions, count questions, attribute identification questions, and comparison questions, requiring a slew of reasoning skills to answer them. Questions can be compositional and require multiple reasoning steps to arrive at the right answer. For a break down of all the question types and a full definition of the generative model used to generate them, we refer the reader to the original CLEVR dataset paper (Johnson et al., 2017).

The CLEVR dataset is a standardized and highly controlled dataset intended to facilitate progress in the development of natural language processing systems, but it is not natural language; it does not have all the same properties as the speech children are exposed to. The language in CLEVR is template-based[4] and text only; by contrast, children's input is composed of a much richer signal including varied syntactic frames, prosody, social cues, and other sources of information. This fundamental difference means that our models do not have access to much of the rich information that children leverage to learn new words. On the other hand, natural environments are also noisier; a constrained learning environment may inadvertently help models learn and converge on the tasks quicker. Working within a highly controlled and simplified learning environment is a necessary first step to understand the relations that exist between models' input and their learning outcomes.

2.2. Semantic probes

Each semantic probe is a set of existential questions based on a simple template that contains one of our function words of interest. Models must know the meaning of the relevant word to answer probe questions correctly; otherwise, we would expect performance to be at or below chance on probe questions overall. Each question is associated with an image from the CLEVR validation image set that satisfies any implied presuppositions. Example image–question pairs from each probe are presented in Fig. 3. The probes are all based on unseen templates, though they are all composed of words which are part of the CLEVR vocabulary and show some similarities with existing CLEVR question templates.

For each probe, given the template, we created the set of questions such that we iterate through every possible combination of referents in the CLEVR universe, allowing us to abstract away any difficulty answering questions that may be due to other content words. For

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

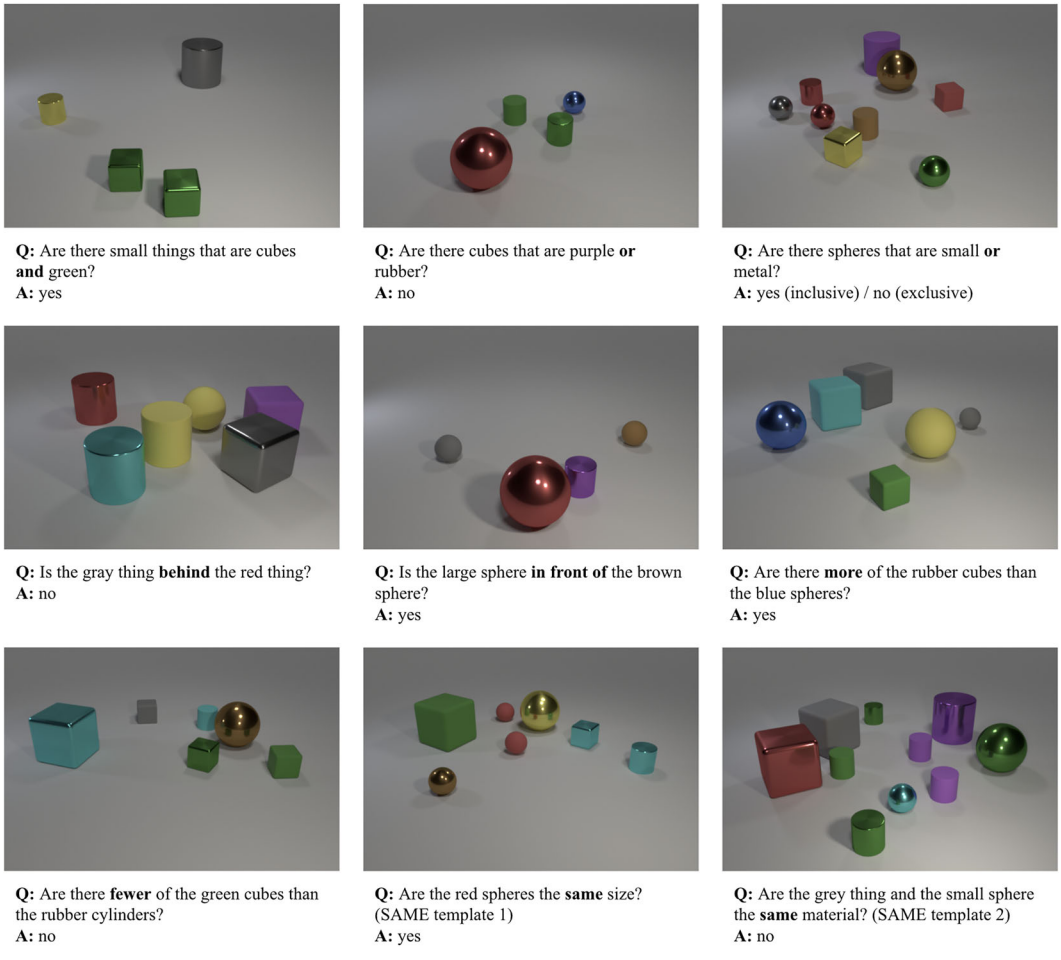


Fig. 3. Example image–question pairs from semantic probes.

each question, we then identified all the images in the validation set that met its presuppositions. If there were more than 10 such images, we randomly sampled 10 of them. Fig. 4 illustrates this procedure. In the rest of this paper, we will use the capitalized version of a word to refer to its respective semantic probes, for example, AND will refer to the semantic probe for the word *and*.

2.2.1. AND–OR

AND–OR probes templates are "Are there Xs that are α **and** β?" and "Are there Xs that are α **or** β?", where X is a referential expression (e.g., gray sphere, metal thing, big cylinder, cube) and α, β are properties (e.g., purple, small, metal). As previously mentioned, the probes iterate through every possible referent combination, where a referent is a noun (thing, sphere, cylinder, cube) optionally preceded by a modifier referring to its color, material, or size. These

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

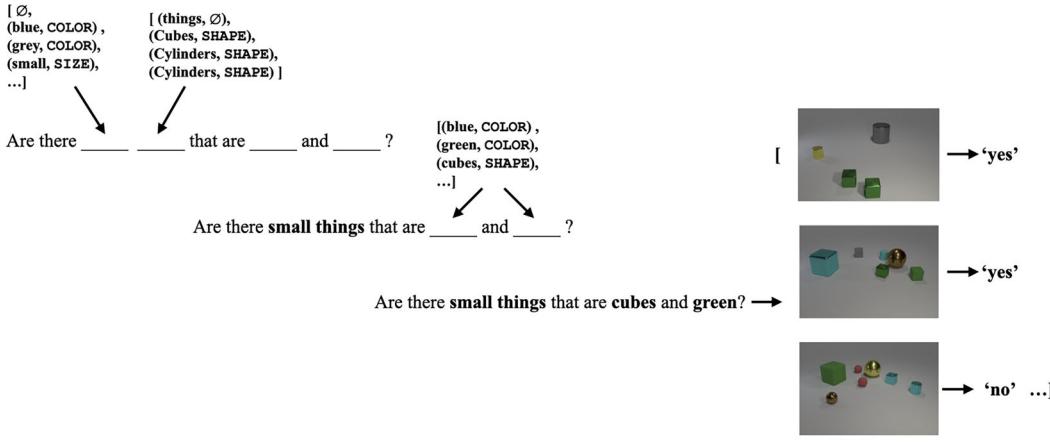


Fig. 4. Example probe creation procedure. Given a template, we cycle through every possible variable combination and then sample 10 images and determine their corresponding answers.

templates do not have any presuppositions, so 10 images were randomly sampled for each one, totaling 15,600 image–question pairs in each probe.

For the AND probe, questions which were paired with images that contained at least one X that was both α and β—$(\alpha \wedge \beta)$—had "yes" as their correct answer, while questions where this requirement was not met in the image had "no" as their answer.

We were interested in determining the prevalence of inclusive versus exclusive interpretations of the word *or* by models. For this reason, we used the following answer scheme for the OR probe. Questions which were paired with images that contained at least one X that was α but not β—$(\alpha \wedge \neg\beta)$ –, or not α but β—$(\neg\alpha \wedge \beta)$ –, expected the correct answer "yes." Questions which were paired with images that contained no Xs or only Xs that were neither α nor β—$(\neg\alpha \wedge \neg\beta)$—had "no" as their answer. As for question–image pairs where all Xs were both α and β—$(\alpha \wedge \beta)$—were ambiguous, expecting a "yes" answer if *or* was interpreted as inclusive, while a "no" answer if on the other hand *or* was interpreted as exclusive.

2.2.2. *BEHIND–IN FRONT OF*

BEHIND–IN FRONT OF probes used as templates "Is the X **behind** the Y?" and "Is the X **in front of** the Y?", where both X and Y are referential expressions. These templates presuppose that the images contain exactly one X and one Y. Again iterating over the same complete set of referent combinations,[5] we identified all the images that satisfied this presupposition. If there were more than 10, we randomly sampled 10 of them; otherwise, we included all available images. In the end, there were a total of 24,380 image–question pairs for each probe.

Using the "scene" metadata available for each image, which contains annotations as to the relative position of objects, we determined the correct answer to each question. These relative positions were determined using the (x, y, z) center point coordinates of objects. Using the underlying threshold definitions of *behind* and *in front* from CLEVR, we determined if an object was behind or in front of another by taking the difference between their y coordinates.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Image–question pairs where X was in fact behind Y received a "yes" answer for the BEHIND probe and a "no" answer for the IN FRONT OF probe. If the opposite was true, the answers were reversed. In our analyses, we additionally wanted to track probe questions performance based on the relative distance between X and Y. For these analyses, we kept track of the Euclidean distance between the two referent objects using all three of their (x, y, z) coordinates.

2.2.3. MORE–FEWER

MORE–FEWER probes follow the forms "Are there **more** of the Xs than the Ys?" and "Are there **fewer** of the Xs than the Ys?" Both these templates presuppose that the images contain at least one X and one Y. Based on this presupposition, we identified all of the compatible images for each question and, again, if more than 10 images were found we randomly sampled 10 of them for a given question. In total, there were 24,420 image–questions pairs in each of these probes.

To determine the answers to each image–question pair, we once again used the "scene" metadata, which was associated with each image. We identified all of the objects that were part of X and Y referent categories and then compared their cardinality. Based on our underlying CLEVR definitions, if the number of Xs was greater than the number of Ys, ($|X| > |Y|$), then the answer to a question in the MORE probe was "yes", while the answer to a question in the FEWER probe was "no". If on the other hand, the number of Xs was less than the number of Ys, ($|X| < |Y|$), then the opposite answering pattern applied, MORE questions had "no" for an answer, while FEWER questions - "yes". In the event that there was the exact same number of Xs and Ys, ($|X| = |Y|$), both probe question types' answer was "no." We were interested in tracking model performance on probe questions as a function of the difference in cardinality between the two referent sets, ($|X| - |Y|$), so we also kept track of this number for each image–question pair.

2.3. Evaluation

In each of the experiments that follow, we use these probes to evaluate how much models have learned about the meaning of these words and how they interpret them given different visual contexts. We test models on all probes at each epoch during model training, allowing us to analyze what they are learning over time. As we do these analyses, it is important to understand certain distributional facts about the training data our models are exposed to.

The CLEVR dataset is well-balanced in terms of the relative frequency of each function word. Table 1 shows the raw counts for words as well as their relative frequency by word pair in the training data. The total number of word tokens is 12,868,670 words, over 699,989 training questions.

Additionally, "yes" and "no" answers to questions containing these words are also generally well balanced, the exception being questions containing the word *or*. Table 2 shows the relative frequencies of these answers for questions containing each of our function words. As evident from this table, there are no questions containing the word *or* which are answered using "yes" or "no." *Or* is always used as a logical conjunct connecting referents, specifically

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
Relative frequencies of each function word pair in the CLEVR training data

Word Pairs	Raw Counts	Frequency
and	81,506	56.32%
or	63,214	43.68%
behind	147,409	49.98%
in front of	147,506	50.02%
more	11,570	49.40%
fewer	11,851	50.60%

Table 2
Frequencies of *yes* and *no* answers for questions containing each function word in the CLEVR training data

	Yes Answers		*No* Answers	
Word Pair	Raw Counts	Frequency	Raw Counts	Frequency
and	20,673	25.36 %	21,463	26.33%
or	0	0%	0	0%
behind	27,491	18.65%	28,707	19.47%
in front of	27,748	18.81%	28,563	19.36%
more	5,549	47.96 %	6,021	52.04%
fewer	5,840	49.28 %	6,011	50.72%

in count questions (e.g., "How many things are blue cubes or small cylinders?"), which all require a number as their answer. All the while, *and* is additionally used in a much wider variety of question types, sometimes connecting prepositional phrases (e.g., "What material is the blue cube that is behind the cylinder and left of the red thing?"). Cumulatively, about 52% of questions with *and* require a yes/no answer, while the rest are other words in the vocabulary. Like *and*, *behind* and *in front of* show up in a variety of question types, requiring different types of answers, while *more* and *fewer* are only used in questions that require "yes" or "no" answers. These differences in input distributions are artifacts of the CLEVR dataset generator and the question templates used by the original authors behind this dataset. Thus, in the results that follow, it is difficult to fairly compare across word pairs or across AND and OR probes; we should instead consider them somewhat independently. However, if we observe differences in results within well-balanced pairs, these are likely due to other factors beyond their frequency in the models' input. We will explore some of these factors further in the experiments that follow.

3. MAC: A recurrent reasoning model for question answering

A variety of models have been proposed for completing visual question answering tasks; all of these include both visual and linguistic processing units. For our current experiments, we chose to use a model that—at the time we began the project—had the top performance

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

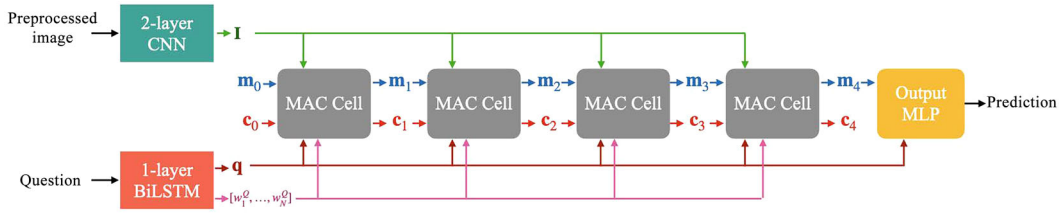


Fig. 5. The MAC model initially processes the image and question through Convolutional neural network (CNN) and biLSTM units, respectively, and then through four recurrent MAC cells, each generating output memory and control states. The final output unit takes the final memory state and the question representation to produce a prediction. The model is fully differentiable.

scores on the original CLEVR task: the MAC (Memory, Attention, and Composition) model of Hudson and Manning (2018). This model reaches an accuracy level of 98.9% on CLEVR's test set. Because it does so well on within-sample questions, we hoped that it could also generalize to our probe questions as well.

Following previous approaches to the CLEVR task (R. Hu, Andreas, Rohrbach, Darrell, & Saenko, 2017; Perez, Strub, De Vries, Dumoulin, & Courville, 2018; Santoro et al., 2017), the MAC model preprocesses images using ResNet-101 (He, Zhang, Ren, & Sun, 2016), pretrained on ImageNet (Russakovsky et al., 2015). The *conv4* layer features from ResNet-101 are then used to represent each image.

The MAC model is a recurrent reasoning model, which we illustrate in Fig. 5 and describe in what follows. It first processes the preprocessed image and question separately. The preprocessed image goes through a two-layered convolutional neural network resulting in a 3D matrix (preprocessed image width × preprocessed image height × number of channels in final convolutional layer) representing what Hudson and Manning call *the knowledge base*, $\boldsymbol{I}$. As for question Q, each word is converted to an embedding vector and then processed through a single-layered bidirectional long-short-term memory (biLSTM) network. The biLSTM yields two outputs for a question Q of length N words: (1) a vector of contextualized word embeddings $[w_1^Q, \ldots, w_N^Q]$, where each w_n^Q is the models output state for w_n; (2) a question representation $\boldsymbol{q}$ which is the concatenation of the final states of both the forward and backward passes of the biLSTM, $\boldsymbol{q} = [\overleftarrow{w_1^Q}, \overrightarrow{w_N^Q}]$. Once the image and question are processed as $\boldsymbol{I}$, $[w_1^Q, \ldots, w_N^Q]$, and $\boldsymbol{q}$, they are used as input for a set of recurrent reasoning steps.

The MAC model uses custom recurrent cells (MAC cells) which each represent one reasoning step t. The best version of the MAC model as originally reported used 12 recurrent MAC cells before the final output layer. Hudson and Manning, however, found that very similar performance could be achieved with as few as four recurrent reasoning steps (test accuracy 97.9%). Thus, we chose to use this smaller and more efficient version of the model for our experiments—see Fig. 5 for a visualization of our version of the model.[6] For each reasoning step t between 1 and 4, the MAC cell takes as input the processed image representation $\boldsymbol{I}$, the contextualized word embeddings $[w_1^Q, \ldots, w_N^Q]$, and the processed question representation $\boldsymbol{q}$.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

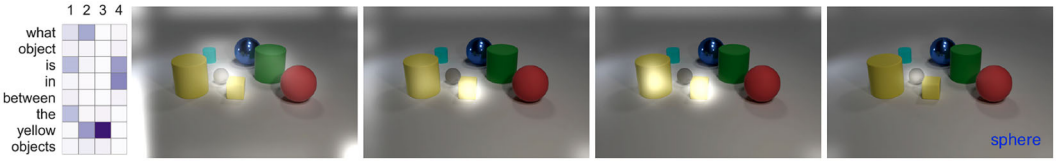


Fig. 6. Example attention maps produced by MAC model at each reasoning step taken from Hudson and Manning (2018).

Additionally, as these are recurrent cells, it also considers two hidden states as input: (1) one representing a soft attention map over the question, $\boldsymbol{c}_{(t-1)}$ (called the control state in Hudson & Manning, 2018); (2) the other representing soft attention map over the image, $\boldsymbol{m}_{(t-1)}$ (called the memory state), where $\boldsymbol{c}_0, \boldsymbol{m}_0$ would be randomly initialized dummy vectors. The output of a recurrent cell at reasoning step t is then $\boldsymbol{c}_t$ and $\boldsymbol{m}_t$, which can then be used as the hidden states for the next reasoning step. At the final step 4, the model integrates the final soft attention map over the image representation $\boldsymbol{m}_4$ with the question representation $\boldsymbol{q}$ through a basic multilayer perceptron (MLP) to predict an answer, which always consists of a single word from the model's shared question and answer vocabulary.

The control state $\boldsymbol{c}_t$ is a weighted distribution over the contextualized word embeddings. In other words, it indicates which words are most important to attend to in a given reasoning step. The memory state $\boldsymbol{m}_t$ is a weighted distribution over regions in the processed image which is conditioned on $\boldsymbol{c}_t$. Intuitively, it encodes which parts of the image to attend to given the parts of the question being considered at a given reasoning step. Example outputs of both memory and control states 1–4 for a given question image pair can be seen in Fig. 6.

For a detailed breakdown of the MAC cell's internal structure and how these attention maps are derived, we refer the reader to Hudson and Manning (2018). For the purpose of this paper, we note that the cell has a relatively simple and straightforward structure composed of separate MLPs for processing the control $\boldsymbol{c}_t$ and $\boldsymbol{m}_t$ memory states. It was designed "to capture the inner workings of an elementary, yet general-purpose reasoning step" and to "encourage the network to solve problems by decomposing them into a sequence of attention-based reasoning operations that are directly inferred from the data, without resorting to any strong supervision" (Hudson & Manning, 2018). The model's generic and simple structure eliminates the possibility of it introducing any form of symbolic structural biases, which is important since it will serve as an example of non-symbolic learning for our hypotheses testing.

4. Experiment 1: Learning to interpret and represent function words

How do visually grounded question answering models learn to represent and interpret function words? Do the representations they learn for words like *and*, *or*, *behind*, *in front of*, *more*, and *fewer* generalize to unseen linguistic and visual contexts?

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

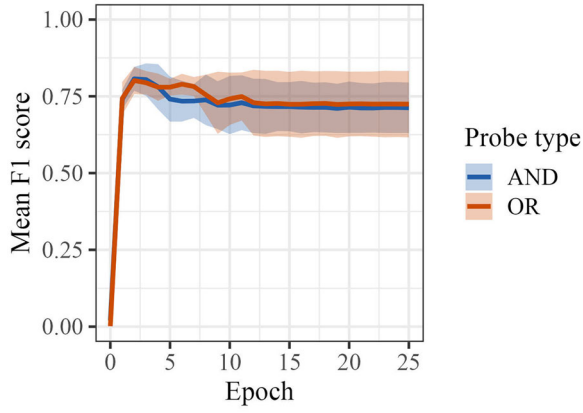


Fig. 7. Experiment 1: Mean F1 score on AND–OR probes overall in non-ambiguous questions, shading represents the standard deviation across five models.

4.1. Setup

We trained five MAC models on the original CLEVR training data for 25 epochs, initialized using different random seeds. Models learn and update using backpropagation with the addition of variational dropout on 15% of parameters across the model at each pass. Models reached an average prediction accuracy of 98.84% on the training data and of 97.74% on the validation set, reproducing the performances originally reported by Hudson and Manning (2018) for four-step MAC models. Since our probes are based on never seen question templates, we expect models' performance on probes to be lower than their performance on the CLEVR's validation set which was created using the same question templates as the training data. We report the mean F1 score and standard deviation across all five models for each probe at each epoch throughout training. Chance performance is, in theory, a near 0 F1 score, since models can produce any word in their vocabulary as the answer to probe questions. However, models very quickly learn after only a couple of batches that existential questions are always answered with either "yes" or "no." significantly reducing the number of answers they actually consider.

4.2. Results

4.2.1. AND–OR

Probe questions were all of the form "Are there Xs that are α and/or β?" As a reminder, there are four possible truth conditions associated with the images the questions are paired with: $(\alpha \wedge \beta)$, $(\alpha \wedge \neg\beta)$, $(\neg\alpha \wedge \beta)$, and $(\neg\alpha \wedge \neg\beta)$. First, let us consider the overall scores of models on probes in non-ambiguous contexts in Fig. 7—in other words, excluding OR probe questions in $(\alpha \wedge \beta)$ contexts, where inclusive and exclusive interpretations of *or* have opposing answers. As seen in the figure, models perform better than chance on both the AND and OR probes.

Next, Fig. 8 shows the mean F1 score reported in the previous figure as a function of the answer type—"yes" or "no"—expected for each question for these probes. There is a clear

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

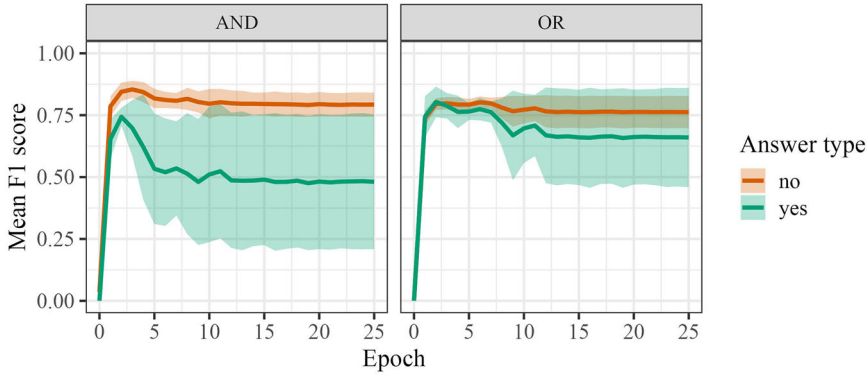


Fig. 8. Experiment 1: Mean F1 score on AND–OR probes by answer type in non-ambiguous questions.

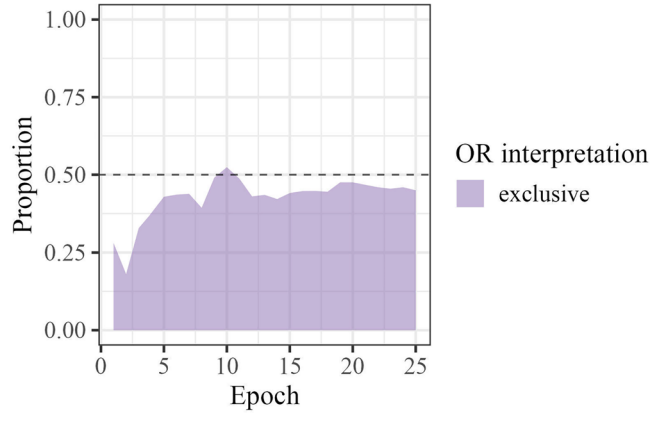


Fig. 9. Experiment 1: Average proportion of exclusive (vs, inclusive) interpretations of OR probe in ambiguous contexts, $(\alpha \wedge \beta)$. Overall standard deviation is $\pm$ 0.36, 3/5 runs learning to favor exclusive interpretation more than 50% of the time.

asymmetry for both probes between questions in contexts requiring a "no" answer versus a "yes," and, second, models performance in "yes" contexts then seems to drop after the second epoch. For AND, "yes" is expected in $(\alpha \wedge \beta)$ contexts and "no" otherwise. For OR, "yes" is expected in $(\alpha \wedge \neg\beta)$ and $(\neg\alpha \wedge \beta)$ contexts, while "no" is expected in $(\neg\alpha \wedge \neg\beta)$ contexts. Though models have no issue recognizing the answer in $(\neg\alpha \wedge \neg\beta)$, they struggle more when OR and AND expect opposing answers. This drop seems to also coincide with the rise of exclusive interpretations for OR in $(\alpha \wedge \beta)$ contexts as we see in Fig. 9.

In Fig. 9, we consider the proportion of inclusive versus exclusive interpretations of OR questions in the contexts where $(\alpha \wedge \beta)$ are both true. Importantly, the CLEVR dataset generative model hard-codes *or* to be interpreted inclusively; in other words, all answers in the training data assume an inclusive *or*. As we might expect, the models initially learn to favor

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

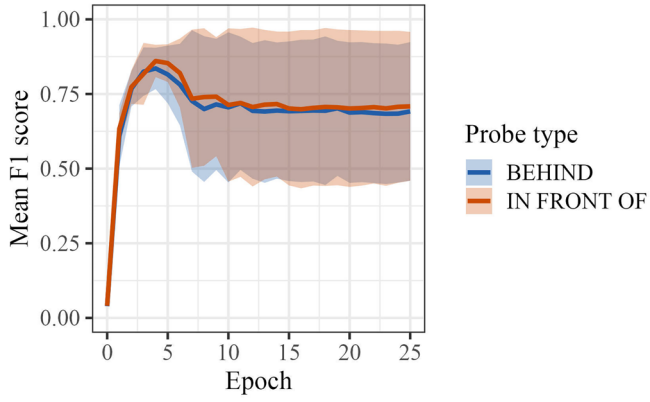


Fig. 10. Experiment 1: Mean F1 score on BEHIND–IN FRONT OF probes overall, shading represents the standard deviation across five models.

inclusive interpretations. However, as learning progresses they start to interpret OR as exclusive more and more.

The differences in performance as a function of the answer types across AND–OR probes suggest that the models struggle more in contexts where AND questions and OR questions have conflicting answers, specifically, in $(\alpha \wedge \neg\beta)$ and $(\neg\alpha \wedge \beta)$ contexts. On the other hand, the results in contexts where $(\alpha \wedge \beta)$ are both true and both AND and OR should have the same answer (assuming an inclusive interpretation of *or*), initially models seem to have no issues, but over time they start to favor exclusive interpretations for *or* and struggle more with *and* questions in the "yes" answer contexts. These results suggest that when determining the answer to a question containing *and* or *or*, models are also considering alternative questions that contain the other logical connective. In the cases where opposite answers for AND versus OR questions are expected, this attention to alternatives could lead to more uncertainty about the right answer. While in the case where the same answer is expected, it may instead be leading to a process akin to "reasoning about alternatives" where opposing logical operators should also have opposing answers, resulting in the rise of exclusive *or*. We explore this hypothesis further in Experiment 2 (see Section 5).

4.2.2. BEHIND–IN FRONT OF

Probe questions are all of the form "Is the X behind/in front of the Y?" and expect opposing answers as a function of the relative position of X to Y. Fig. 10 shows the overall F1 scores of the models on both probes. There is more variation across random seed runs, though both BEHIND and IN FRONT OF seem to be learned equally well within runs and performance is generally above chance.[7] Unlike for AND and OR, Table 2 shows us that *behind* and *in front of* are used in a similar number of questions and expect "yes/no" answers at equal frequencies; we can, therefore, fairly compare models' relative performance on these words.

As with the previous probes, we also consider models' performance as a function of the answer type. Whether the context required a "yes" or "no" answer did not seem to matter for

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

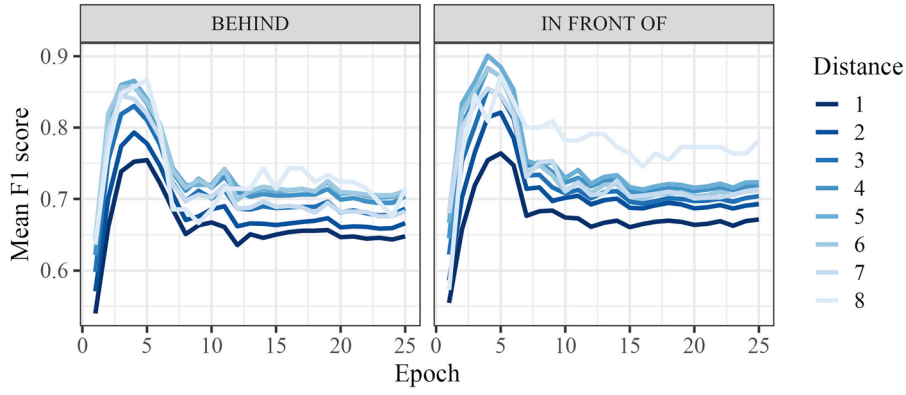


Fig. 11. Experiment 1: Mean F1 score on BEHIND–IN FRONT OF probes as a function of the Euclidean distance between referents.

these probes as much as it did for others; models performed just as well in either context overall.

In Fig. 11, we look at how well models predict the correct answer to our BEHIND–IN FRONT OF probe questions as a function of the Euclidean distance between X and Y referents. The distances were calculated based on the coordinates of the center of each object provided in the metadata of each image. We then rounded the distances to the closest integer to bin our data into distance levels. Objects that have an Euclidean distance of 1 are so close that we expect one to partially occlude the other, while distances of 8 are as far apart as objects can be within a CLEVR image. As we can see from the figure, there is a very clear gradience in performance based on the distance between X and Y, such that the further apart two objects are, the easier it is for the model to correctly interpret *behind* and *in front of*.

These results suggest the models can learn meaningful representations *behind* and *in front of* such that they can interpret them in novel contexts. Furthermore, when these prepositions are equally frequent in models' input, they are learned at the same rate. Importantly, models seem to learn a gradient semantic representation for the words as a function of the distance between referents, rather than the strict threshold-based meaning which the CLEVR generative model uses.

4.2.3. MORE–FEWER

Probes are composed of questions of the form "Are there more/fewer of the Xs than the Ys?" For this analysis, we consider three contexts: when $|X| > |Y|$, $|X| < |Y|$, and $|X| = |Y|$. In the first two contexts, MORE and FEWER questions expect opposite answers, while in the third context where there is no difference in the number of Xs and Ys, they expect the same answer, "no." Fig. 12 presents the overall F1 scores of models on both probes. This initial plot suggests that MORE is learned first and may be overall easier than FEWER.

Next, we plot accuracy on probes as a function of the absolute difference between the number of Xs and Ys, $absolute(|X| - |Y|)$ (Fig. 13). Models clearly struggle with both MORE

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

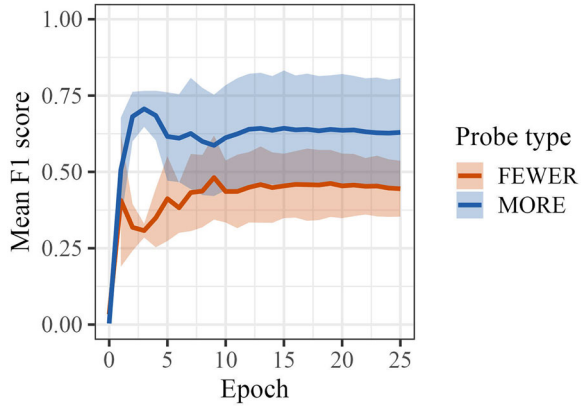


Fig. 12. Experiment 1: Mean F1 score on MORE–FEWER probes overall, shading represents the standard deviation across five models.

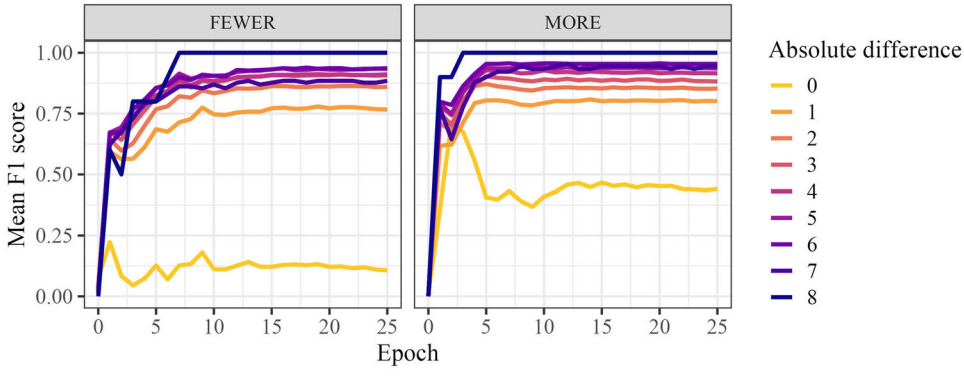


Fig. 13. Experiment 1: Mean F1 score on MORE–FEWER probes by absolute difference in the number of objects in each referent class.

and FEWER questions specifically when the difference is 0, or $|X| = |Y|$, performing below chance in this context. In all other cases, whether the answer is "yes" or "no," models correctly answer questions over 75% of the time. Yet again, performance for these probes is gradient. Models correctly interpret both *more* and *fewer* more often as a function of the difference in number between the two referent classes. The larger the difference, the easier it is for the model to correctly judge whether there are *more* or *fewer* of a given class of referents. Additionally, models poorer performance on FEWER probe questions overall seen in the previous plot seems to be isolated to the contexts where $|X| = |Y|$.

In fact, if we remove all probe questions where $|X| = |Y|$ and consider the overall of models again in Fig. 14, we see a very different picture than our original Fig. 12. Models have almost equally high performance on both probes, still learning *more* slightly earlier than *fewer*.

These results suggest that models learn reasonable meaning representations for both *more* and *fewer* and, furthermore, that these representations are gradient as a function of the dif-

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

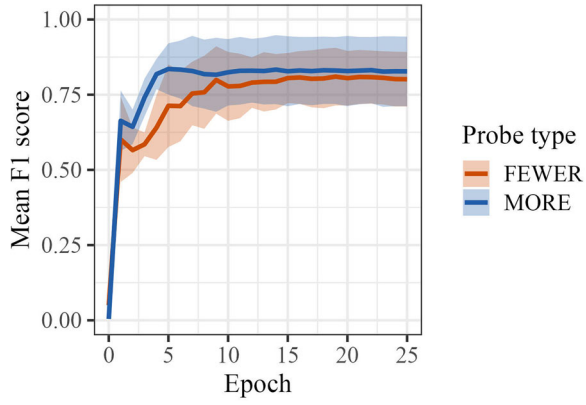


Fig. 14. Experiment 1: Mean F1 score on MORE–FEWER probes overall excluding context where $|X| = |Y|$, shading represents the standard deviation across five models.

ference in number between two referent categories, rather than being based on strict thresholds, which the CLEVR generative model uses. However, models struggled in contexts where $|X| = |Y|$ specifically. We hypothesize that this may be because they are exposed to a third alternative numerical reasoning expression during training, *equal/same number*. This alternative expression expects an opposing answer in these contexts. Like with AND and OR, models may be considering the existence of alternative propositions when trying to answer these questions, leading to more uncertainty in the context where the difference in number between Xs and Ys is the smallest. We explore this hypothesis further in the next experiment.

4.3. Interim conclusion

Our first experiment examined the first research question: How do visually grounded question answering models learn to represent and interpret function words and do the representations they learn generalize to unseen linguistic and visual contexts? We found that models learned gradient interpretations for function words requiring spatial and numerical, *behind*, *in front of*, *more*, and *fewer*. Additionally, we found early evidence that models consider alternative logical connectives when determining the meaning of expressions containing *and* and *or*. This behavior may be leading models to interpret *or* as exclusive in an increasing number of contexts. Further experimentation is necessary to test this hypothesis.

5. Experiment 2: The effect of alternatives on reasoning

Does the existence of alternative expressions in each reasoning pair affect their acquisition or are the meanings of function words acquired in isolation? Following our results from the previous experiment, we hypothesized that models could be considering alternative questions and answers that use opposing or parallel function words when they compute the probability of the answer to a given question. This process akin to "reasoning about alternatives" could

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

then explain the performance patterns we observed specifically for the AND–OR probes as well as the MORE–FEWER probes. Unlike BEHIND–IN FRONT OF which always expects opposing answers, AND–OR and MORE–FEWER pairs both have contexts where they expect the same answer and others where they do not.

The existence of alternative expressions may lead to uncertainty in model predictions in one of two ways. First, if models observe that *and* and *or* are interpreted the same in a set of contexts, then they may begin to expect them to also mean something similar in contexts where they actually should have opposing answers. Second, if models instead observe that they have opposing answers in a set of contexts, then they may instead begin to expect them to mean something different also in contexts where in fact they should be interpreted the same way. In either case, the existence of the alternative expression (*and* in the case of *or*, and *or* in the case of *and*) is what leads models to answer incorrectly, showing evidence of this process. Our second experiment tests our theory and answers our second research question.

5.1. Setup

As in Experiment 1, we train five MAC models initialized using different random seeds. Unlike the previous experiment, however, we manipulate the training data to remove alternative function words which we believe affected the probe performance for OR, AND, MORE, and FEWER. Specifically, we remove all questions from the training data which contain the word *and* and then evaluate model performance on the OR probe. We repeat this process and create a version of CLEVR where we remove all instances of *or* and then evaluate models on the AND probe. Finally, we create a version without an *equal/same number of* and evaluate the models on MORE and LESS probes. By removing *and*, we want to see if the model will correctly learn the semantics of *or* and favor inclusive interpretations when the alternative logical connective is not present. By removing *or*, we want to make sure models learn to correctly interpret *and* regardless of the answer context. Finally, by removing the *equal/same number of* and its derivatives, we would like to see if the models can correctly learn to use *more* and *fewer* in contexts where $|X| = |Y|$, when the alternative proposition that there is an *equal* amount of them is no longer available. For each of these different subsampled training datasets and evaluation probes, we train models for 25 epochs and evaluate the performance of probes at each epoch.

5.2. Results

Since the results for the OR probe and AND probe come from different models trained on different subsampled datasets, we report their results separately for this experiment.

5.2.1. OR

Models reach a higher overall mean F1 score on unambiguous OR probe questions when trained on data without *and*. Comparing model performance when trained with and without *and* as a function of the answer type expected in Fig. 15, it is clear that when we remove the alternative expression models no longer struggle in contexts expecting a "yes" answer as they did in Experiment 1.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

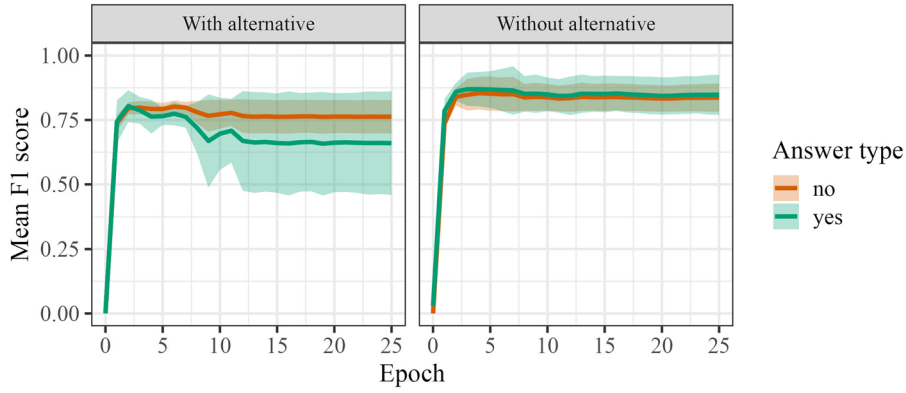


Fig. 15. Experiment 2: Mean F1 score on OR probe by answer type in non-ambiguous questions when trained on data with the alternative expression *and* from Experiment 1 versus without this alternative in Experiment 2.

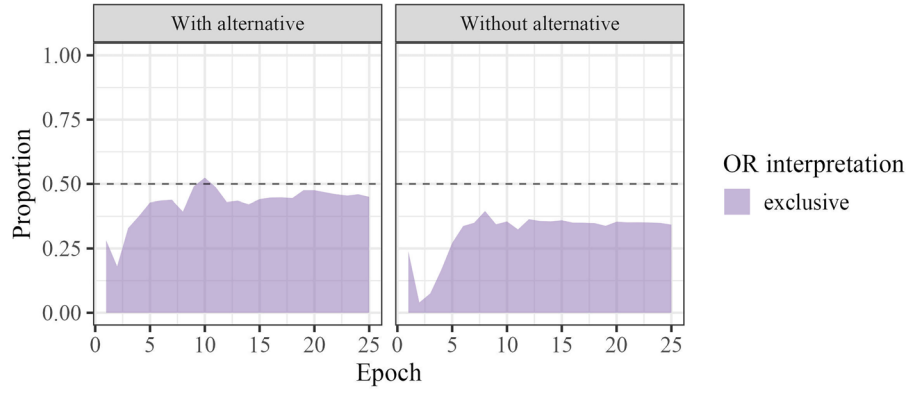


Fig. 16. Experiment 2: Proportion of exclusive (vs, inclusive) interpretations of OR probe in ambiguous contexts, $(\alpha \wedge \beta)$, when trained on data with the alternative expression *and* from Experiment 1 versus without this alternative in Experiment 2. Overall standard deviation for Experiment 2 is $\pm$ 0.3, 4/5 runs favoring inclusive interpretations.

As for probe questions containing *or* in ambiguous contexts where inclusive-or and exclusive-or interpretations predict opposing answers, we no longer see as strong of a progressive rise in exclusive interpretations, instead settling on average with around 70% of ambiguous questions being answered with inclusive "yes" answers (Fig. 16).

When the alternative logical connective *and* is not present, models have no difficulty learning the semantics of *or*. Since the CLEVR generative model defines *or* as inclusive, when no pragmatic alternative is present, models also learn to interpret *or* inclusively. These results support the hypothesis that the rise in exclusive interpretations seen in Experiment 1 is due to some form of competition between *or* and the available alternative *and*.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

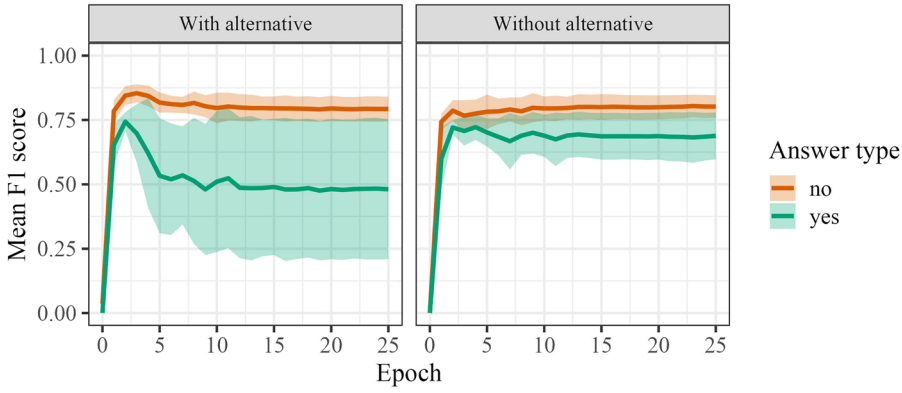


Fig. 17. Experiment 2: Mean F1 score on AND probe by answer type when trained on data with the alternative expression *or* from Experiment 1 versus without this alternative in Experiment 2.

5.2.2. *AND*

Probe results come from models trained on a subsampled version of CLEVR where all instances of *or* have been removed. Models had better F1 scores on AND probe questions when the alternative logical connective was removed than when both were present in Experiment 1. In the absence of *or*, models learned to correctly interpret *and* regardless of the truth value context (Fig. 17).

Models can learn the meaning of the logical connective *and* correctly and then generalize it to interpret this word in novel contexts. If the alternative logical connective for disjunction is present, like in Experiment 1, then the models may struggle more, as they seem to consider the existence of this alternative when trying to determine the intended meaning of *and*. This difficulty disappears if the alternative is no longer present.

5.2.3. *MORE–FEWER*

Probe results from Experiment 1 showed that models struggled to correctly interpret *more* and *fewer* in the context where there were an equal number of the two referent categories being compared. We hypothesized that models may have struggled in this context because there existed alternative questions that asked whether there were an *equal* number of Xs and Ys in the training data. To test this hypothesis, we trained models on a subsampled version of CLEVR where we removed all questions that asked about number equality. Fig. 18 shows the overall performance of these models on both probes when trained with and without this alternative *equal* expression. F1 scores on FEWER questions have definitely risen in comparison to Experiment 1, though results for MORE look quite similar.

However, when we consider model performance on questions as a function of the absolute difference in number between the compared referent categories in Fig. 19, models still struggle in contexts where $|X| = |Y|$. They do better overall in all other contexts.

Unlike with AND and OR, removing the pragmatic alternative did not solve our issue with FEWER and MORE. After carefully scrutinizing the training data from CLEVR, it became

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

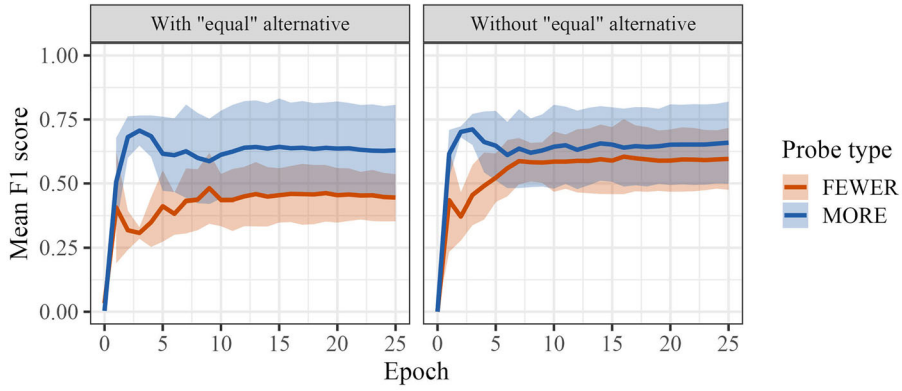


Fig. 18. Experiment 2: Mean F1 score on MORE–FEWER probes overall when trained on data with the alternative expression *equal* from Experiment 1 versus without this alternative in Experiment 2. Shading represents the standard deviation across five models.

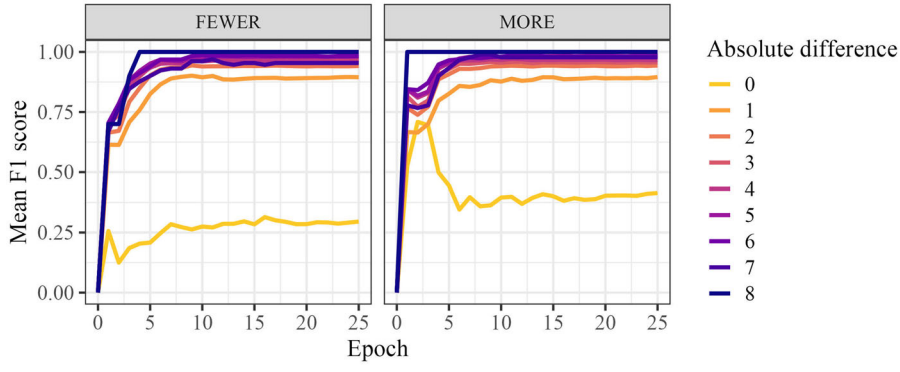


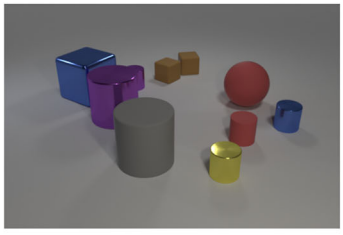
Fig. 19. Experiment 2: Mean F1 score on MORE–FEWER probes by absolute difference in the number of objects in each referent class when trained without the alternative *equal* expression.

apparent that *more/fewer* rarely appeared in contexts where $|X| = |Y|$ and only when they were part of more complex question templates. Fig. 20 shows example questions with *more* in the context where $|X| = |Y|$ taken from the CLEVR train data. Thus, the issues we see with probe performance in this context may simply be due to our choice of template and the idiosyncrasies in the distribution of *more* and *fewer* in the CLEVR training data.

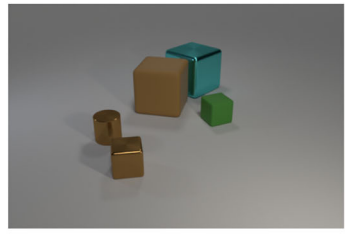
5.3. Interim conclusion

This experiment examined our second research question: Does the existence of alternative expressions in each reasoning pair affect their acquisition or are the meanings of function words acquired in isolation? We found that in the absence of a logical alternative, models correctly learned to generalize the meaning of conjunction and disjunction. Our findings confirm our hypothesis that the presence, or absence, of a pragmatic alternative, can affect

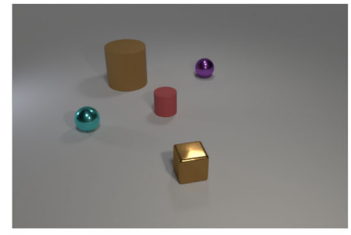
15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Q: Are there more cylinders that are behind the large purple shiny cylinder than tiny yellow cylinders?
A: no (1 = 1)



Q: Are there more big cyan blocks in front of the large cyan thing than blue cylinders?
A: no (0 = 0)



Q: Are there more balls that are in front of the brown rubber object than tiny cyan things?
A: no (1 = 1)

Fig. 20. Example CLEVR training questions with the word *more* in the context where $|X| = |Y|$.

how models learn to interpret logical connectives *and* and *or*. Next, we will evaluate how the frequency of different function words may also affect how models learn their meanings.

6. Experiment 3: The effect of frequency on learning

Our third and final experiment considers the effect of word frequency on the order in which function words are learned. We address our third research question: Does the order in which function words are acquired by models resemble that of children—and are some of these ordering effects simply the result of frequency in the input or are there other conceptual factors at play?

6.1. Setup

We again trained five MAC models initialized using different random seeds for a total of 25 epochs and considered their performance on semantic probes throughout training. Our main manipulation that differentiates this experiment from the others is the training data. As in Experiment 2, we use a subsampled version of the CLEVR training questions. This time, we created a version of CLEVR where the relative frequencies of the target function words matched their relative frequencies across all English child-directed utterances from the CHILDES repository (MacWhinney, 2000).

The CHILDES repository is a collection of open-source transcripts, recordings, and videos of child-caregiver/experimenter interactions from a wide range of studies dating as far back as the 1950s. Children in these studies vary in age between 9 months and 5 years old, the median being about 3 years. Using the childes-db API (Sanchez et al., 2019) to access the data, we isolated all of the English transcript corpora available. We then filtered each to isolate all utterances that were not said by the child, representing a sample of the linguistic input the child was exposed to. We used this corpus to calculate the relative frequencies in children's input of the function words we are interested in. The corpus contained a total of 16,062,386 word tokens.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Relative frequencies of each function word pair in the CHILDES and subsampled CLEVR training data for Experiment 3

	CHILDES		CLEVR Subsampled	
Word Pair	Raw Counts	Frequency	Raw Counts	Frequency
and	217,497	90.45%	81,506	90.45%
or	22,975	9.55%	8,610	9.55%
behind	2,954	79.62%	113,881	74.36%
in front of	756	20.38%	39,260	25.64%
more	23,406	99.10%	11,570	99.10%
fewer/less	212	0.90%	105	0.90%

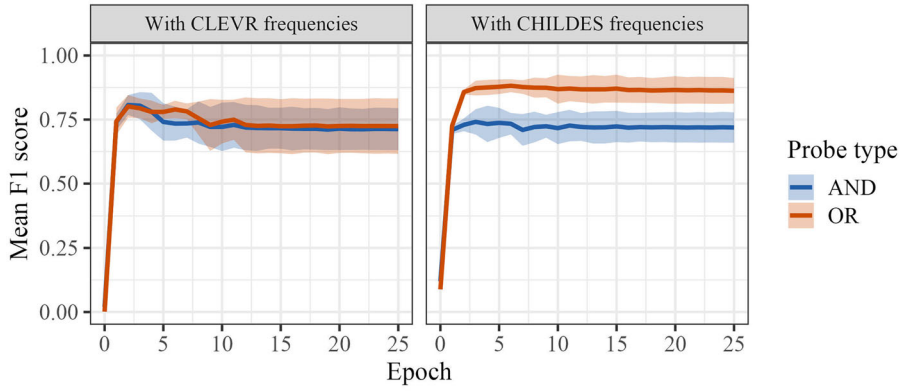


Fig. 21. Experiment 3: Mean F1 score on AND–OR probes overall in non-ambiguous questions when trained on the original CLEVR dataset and the subsampled version with CHILDES-like frequencies. Shading represents the standard deviation across five models.

We considered the relative frequencies of our function words within each contrasting pair rather than their relative frequencies overall as it would not have been possible to extract a reasonably sized subsampled version of the CLEVR training data otherwise. One of the main difficulties we ran into when trying to subsample from the CLEVR dataset was that these function words often appeared in overlapping sets of questions, so changing the frequency of one word by subsampling questions would inadvertently affect another's frequency. Nonetheless, we managed to create a version of the CLEVR training data that almost reproduced the relative frequencies of the CHILDES data and was of a reasonable size, containing 545,681 training questions (9,652,086 tokens). Table 3 shows the exact word counts and frequencies of both the CHILDES and subsampled CLEVR training datasets.

6.2. Results

6.2.1. AND–OR

Fig. 21 compares the overall performance of the models on each of these probes in non-ambiguous questions (i.e., excluding OR questions in $\alpha \wedge \beta$ contexts) in Experiment 1 with

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

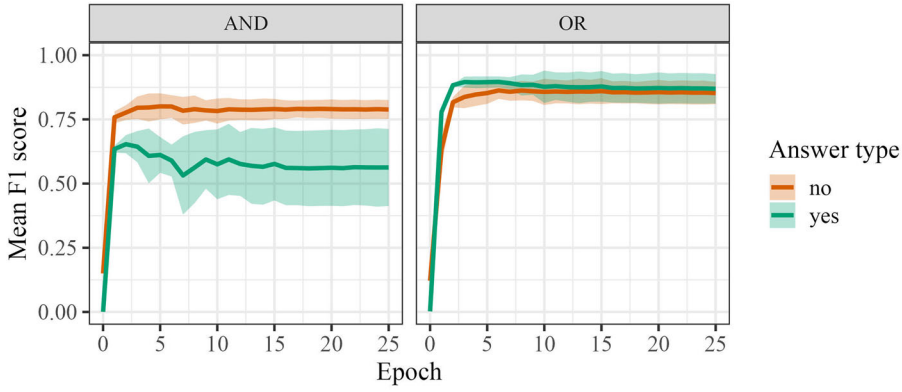


Fig. 22. Experiment 3: Mean F1 score on AND–OR probes by answer type in non-ambiguous questions when trained of subsampled CLEVR.

the original CLEVR dataset frequencies and the current experiment with CHILDES-like relative frequencies. These words have a very uneven distribution in the subsampled CLEVR version like CHILDES, *and* are much more prominent than *or* in children's input. Interestingly, even with this frequency imbalance, models seem to do quite well on both our AND and OR semantic probes, suggesting that even with a reduced number of training examples containing *or*, they are still learning a reasonable representation for this word that allows them to generalize its meaning to unseen contexts.

This observation is confirmed when we consider the current models' mean F1 score by answer type, "yes" or "no," where OR probe performance is the same regardless of context (Fig. 22). As for the AND probe, models seem to be performing as it did in Experiment 1, struggling more in contexts requiring "yes" as an answer.

In the case of ambiguous OR questions, in $\alpha \wedge \beta$ contexts, models clearly prefer inclusive answers; we see no rise in exclusive interpretations like the one seen in Experiment 1 (see Fig. 23).

If performance on these probes were solely a function of the frequency of these words in models' input, we would expect their performance on the OR probe to decrease between Experiment 1 and Experiment 3, but as we saw in Fig. 21 this is not what happens. Furthermore, if the effect of being sensitive to possible alternative expressions was also proportional to the frequency of these alternatives, we might also expect to see a stronger effect of the alternative *and* on OR probe results and an increase in inclusive interpretations for *or*, but again, we do not see this effect. It seems to have been stronger in Experiment 1 when *and* and *or* were about equally frequent. The more uniform distribution between these words in Experiment 1 could have led to more uncertainty overall. This explanation is further supported by the smaller standard deviations we see in Fig. 21 for models trained on the CHILDES-like frequencies versus those trained on the original CLEVR dataset. Another possible explanation that should not be discounted is that in downsampling questions containing *or* in the training set, we may have simply reduced the diversity of contexts seen for *or* in favor of con-

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

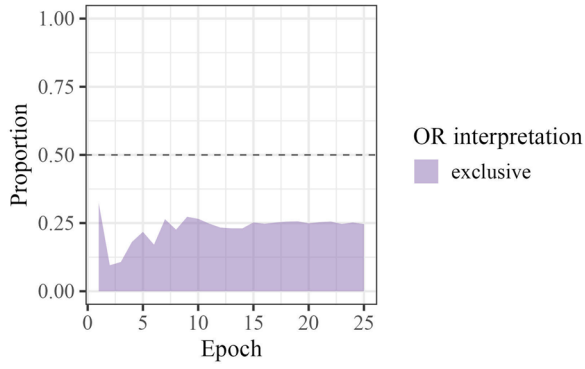


Fig. 23. Experiment 3: Proportion of exclusive versus inclusive interpretations of OR probe in ambiguous contexts, $(\alpha \wedge \beta)$, when trained of subsampled CLEVR. The overall standard deviation is $\pm$ 0.19, all runs favoring inclusive interpretations.

texts that resembled our probe template more, such that the models now had less uncertainty specifically about the meaning of *or*.

As for AND, models' performance in Experiment 1 and this experiment is very similar, albeit with a little less variation across runs in the current experiment. Models still struggle in contexts where "yes" answers are expected. The fact that they seem to do better on the OR probe than the AND probe in this experiment does not necessarily mean that *or* is easier to learn than *and*, since as we noted in Section 2.3, unlike with the other two contrasting function word pairs, *and* and *or* have very different input distributions. *Or* is always used as a logical conjunct connecting referents in count questions, while *and* is used in a much wider variety of question types, connecting different types of phrases. Some of the difficulty with AND probe questions in "yes" contexts may simply be due to the distribution over input questions the models see for *and* and how different these questions are from our out-of-distribution probe questions. Frequency is clearly not the only factor at play in determining how and when models come to learn these words.

6.2.2. BEHIND–IN FRONT OF

These words are also not evenly distributed in children's input in CHILDES and consequently in our subsampled dataset. Both the number of instances of *behind* and *in front of* had to be reduced to create the training data used in this experiment, but we had to decrease the number of *in front of* instances significantly more to reproduce their relative frequencies from CHILDES. As we can see in Fig. 24, these changes had an effect on the overall performance of models on the IN FRONT OF probe which now finds itself on average around chance with much more variation across runs. The performance on the BEHIND probe is about the same in both conditions.

The most interesting results can be seen in Fig. 25 where we have plotted model performance on probe questions as a function of the Euclidean distance between the two referents in probe questions. Again, the results from Experiment 1 for the BEHIND probe are reproduced,

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

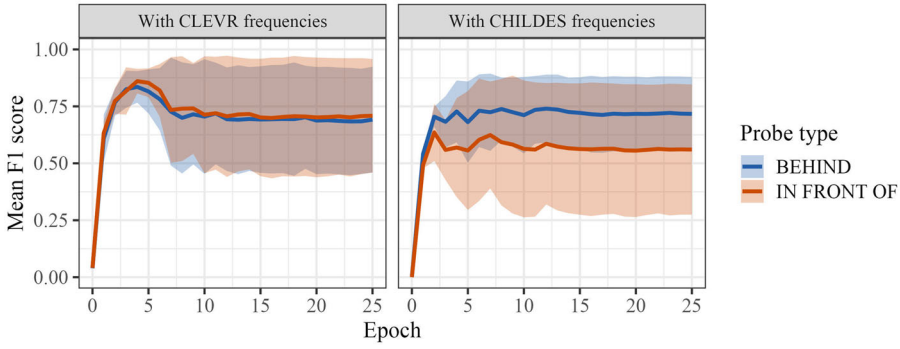


Fig. 24. Experiment 3: Mean F1 score on BEHIND–IN FRONT OF probes overall when trained on the original CLEVR dataset and the subsampled version with CHILDES-like frequencies. Shading represents the standard deviation across five models.

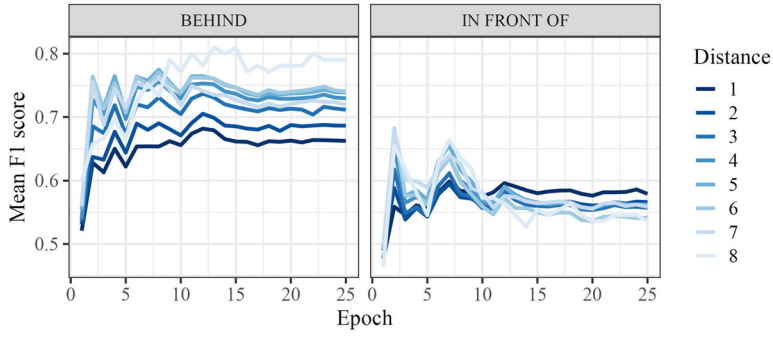


Fig. 25. Experiment 3: Mean F1 score on BEHIND–IN FRONT OF probes as a function of the Euclidean distance between referents when trained of subsampled CLEVR.

showing a clear gradience in interpretations of *behind* as a function of distance. However, in the case of IN FRONT OF, the gradience has completely disappeared.

All of these results suggest that when models are trained on a CLEVR training dataset that reproduces the relative frequencies of *behind* and *in front of* seen in children's input, they learn the most frequent word of the pair, *behind*, but struggle to learn the meaning of the less frequent opposing word, *in front of*. This pattern differs from that of *and* and *or*, since for *behind* and *in front of*, frequency does seem to be the most important factor in determining their relative learning order and difficulty.

6.2.3. MORE–FEWER

These words are an interesting case to consider because *fewer* is extremely rare in children's input while *more* is quite common. There are few different senses of the word *more*, the most common in children's input being its adverbial form as in "do you want more?", which is quite different from the comparative quantifier *more* seen in CLEVR as in "more than." Since we could not easily differentiate all the senses of *more*, we decided to also include its counterpart

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

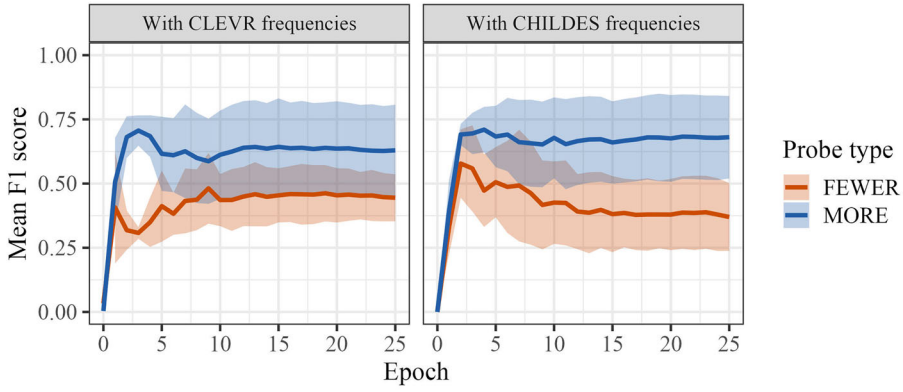


Fig. 26. Experiment 3: Mean F1 score on MORE–FEWER probes overall when trained on the original CLEVR dataset and the subsampled version with CHILDES-like frequencies. Shading represents the standard deviation across five models.

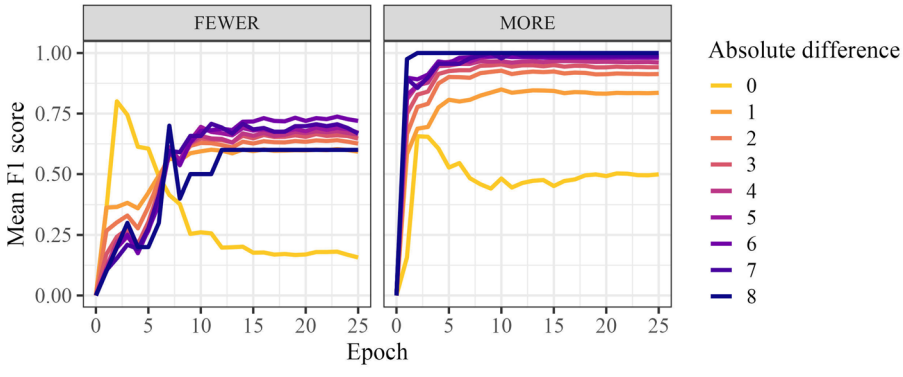


Fig. 27. Experiment 3: Mean F1 score on MORE–FEWER probes by absolute difference in the number of objects in each referent class when trained of subsampled CLEVR.

less in addition to *fewer* when determining their relative frequencies. Nonetheless, *more* was much more frequent than *fewer* and *less* combined (see Table 3).

Fig. 26 compares the overall performance of models on both probes when trained on the original CLEVR dataset and the subsampled version with CHILDES-like frequencies. Performance on MORE is about the same, while on FEWER seems a little lower in the current experiment.

Further probing with Fig. 27 shows that errors are isolated specifically to contexts where $|X| = |Y|$—yet again. Surprisingly given the very small number of exemplars of *fewer* seen during training—only 105 cases—models still seem to learn to use *fewer* in unseen contexts as long as the absolute difference in number between referent classes is greater than zero. Additionally, unlike our results for *in front of*, models still show some gradience in interpretation for *fewer* as a function of number difference. Questions with *fewer* are all answered with "yes" or "no," while questions with *in front of* expect a much broader set of answers in

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

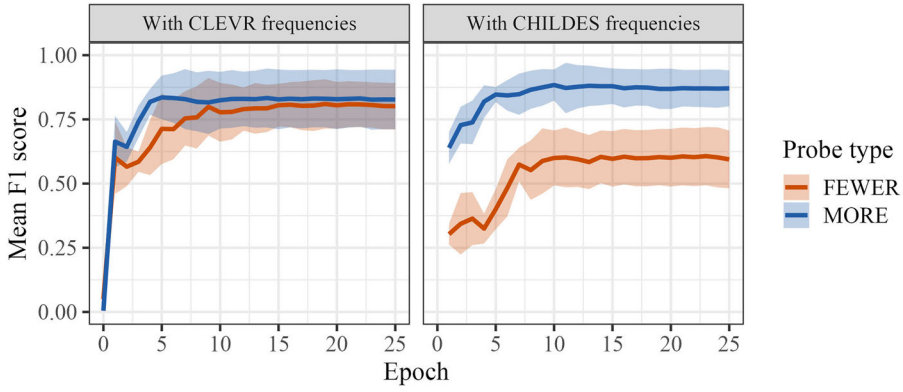


Fig. 28. Experiment 3: Mean F1 score on MORE–FEWER probes overall excluding context where $|X| = |Y|$, when trained on the original CLEVR dataset and the subsampled version with CHILDES-like frequencies. Shading represents the standard deviation across five models.

the original CLEVR dataset (see Tables 1 and 2). This difference in input distribution might explain why models can still learn a reasonable representation for *fewer* with so few examples.

By removing the probe questions where $|X| = |Y|$ and replotting models' F1 scores for all other cases in Fig. 28, we can clearly see that they learn to properly use both MORE and FEWER most of the time, though the performance on FEWER questions has definitely decreased in comparison to the results from Experiment 1 when trained on the original CLEVR data.

Even with only a few exemplars of *fewer*, models are able to learn reasonable meaning representations for this word, showing gradient interpretation as a function of the difference in number between compared classes. Models are not as accurate on the FEWER probe as they are on MORE questions. Once again, in contexts where $|X| = |Y|$, models struggle to answer both MORE and FEWER questions correctly. These results suggest that relative word frequency in the input also affects how models learn these function words.

6.3. Interim conclusion

With this experiment, we addressed our third research question: Do models learn these function words in a similar order to children and are these ordering effects the results of their frequency or do they follow from other conceptual explanations? When trained on a corpus with similar relative frequencies to children's input, the MAC models had difficulty learning less frequent function words. For our logical reasoning targets, however, there are factors beyond frequency that influenced our models' ability to learn the meanings of *and* and *or*.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

7. General discussion

How children learn "hard" words like *and/or*, *behind/in front of*, and *more/fewer* is still an open question. Proposals for their acquisition range along a spectrum between children having innate knowledge of the reasoning skills required to understand these words (a nativist perspective) to having to learn them from scratch using general learning mechanisms (a usage-based perspective). In this paper, we used a recurrent neural network model exposed to the visually grounded language as a test bed to evaluate the learnability of these function words, providing a proof of concept that such words can be learned from data.

First, we asked whether models were able to learn the meaning of these words using their non-symbolic general learning mechanisms. We found that they did learn to interpret function words along the way to succeeding in the visual question answering task they were trained on, the CLEVR dataset. Models favored learning linguistic and visual representations that allowed for gradient interpretations for function words requiring spatial and numerical reasoning rather than threshold-based interpretations, showing that gradience in meaning may emerge from exposure to language in visually grounded contexts. Models also learned to interpret logical connectives *and* and *or* without any prior knowledge of logical reasoning. Additionally, in answer to our second question, we found that models showed evidence of being sensitive to alternative possible expressions when inferring the meaning of these words, which led to a rise in exclusive interpretations for *or* in Experiment 1. Finally, we wondered whether the relative difficulty of acquisition of words for children could be replicated in models and if it varied as a function of frequency rather than conceptual factors. We found that word learning difficulty was indeed dependent on word frequency in models' input, with more frequently seen words generally being easier to learn in the case of spatial and numerical reasoning expressions. When exposed to these words at similar frequencies to children, models showed similar ordering effects for both *behind/in front of* and *more/fewer* word pairs. As for our logical reasoning targets, there seemed to be factors beyond frequency that influence our models' ability to learn them. One possible explanation for this difference may be that it is an artifact of the CLEVR dataset, which presented very different context distributions for *and* and *or* as opposed to other function word pairs.

We acknowledge that this work has its limitations. First, the CLEVR dataset is template-based and has a limited vocabulary. Its relative distribution of function words to content words like nouns and verbs is different from natural language, which may change the essence of what it means to be a function word or closed-class word. Additionally, the function words in question appear in a much wider variety of syntactic and semantic contexts in natural language. Though the dataset remains a good test bed for considering the acquisition of the reasoning skills necessary for interpreting these particular words in context, children acquiring these words may face challenges in naturalistic contexts that cannot be modeled with CLEVR data. Second, our probes do not allow us to determine where gradience in interpretations originates. We can only conclude that gradience arises from the integration of both visual and linguistic representations in the model. Third, our probes are a zero-shot evaluation looking at model generalization for a limited number of templates. To strengthen our conclusions, we would need to see the models response patterns extended to more tem-

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

plates. Still, this work exemplifies how we can probe the nuanced linguistic interpretations of visually grounded models for future studies.

It is possible to learn these complex and abstract reasoning skills and to map them to interpretations of function words without any prior knowledge. Our results offer proof-of-concept evidence that sophisticated statistical learning mechanisms, when applied to visually grounded language, may be enough to explain the acquisition of these function words and related reasoning skills supporting more usage-based theories. Congruently, word learning difficulty was found to be mainly affected by frequency of exposure rather than conceptual factors.

Our work converges with other recent work suggesting that a variety of non-symbolic neural networks can learn logical operators from sufficiently rich data. For example, Geiger, Carstensen, Frank, and Potts (2023) showed that the logical operator "same" could be learned from data. Although our work here focused on a supervised learning regime, Geiger et al. showed learning successes across supervised and unsupervised contexts, supporting the idea that supervision does not necessarily play a key role in the emergence of symbolic structure. More broadly, the successes of large language models on large-scale reasoning tasks (T. Brown et al., 2020; Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022; Wei et al., 2022) suggest that unsupervised learning may be sufficient for the emergence of functional representations supporting reasoning, though more work is needed to probe such models (Mahowald et al., 2023).

The unprecedented success of neural network models offers an opportunity for cognitive science researchers to reevaluate questions about the learnability of language (Lappin, 2021; Piantadosi, 2023; Warstadt & Bowman, 2023) and provides a new set of tools for comparisons between machine learning and child learning (Frank, Monaghan, & Tsoukala, 2019; Portelance & Jasbi, 2023). We hope that our work here contributes to this broader enterprise.

Open Research Badges


This article has earned Open Data and Open Materials badges. Data and materials are available at https://github.com/evaportelance/vqa-function-word-learning.

Notes

1 All of the data, models, and experiment code presented in this paper are publicly available at https://github.com/evaportelance/vqa-function-word-learning.
2 In Appendix C in the Supporting Information, we also include some experiments with relational reasoning and the adjective *same*.
3 The CLEVR dataset also contains a test set, but since this dataset was designed as a benchmarking task, the meta-information for test images is not publicly available, nor are the answers to the test questions. We tried contacting the authors of the original paper to gain access to the test images' meta-information in order to use them for our probe design, but we were unsuccessful. For these reasons, the images from the validation set were used in designing our semantic probe-testing task.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4 Examples of the templates in question containing our function words are given in Appendix A in the Supporting Information.
5 There is an exception for the noun "thing" in the case of BEHIND–IN FRONT OF and MORE–FEWER probe templates which obligatorily requires a modifier, for example, "Is the *blue* thing behind the sphere?" We must include some modifier like *blue* or the referent cannot be uniquely identified.
6 We kept all other hyperparameters the same as the ones used in the main version of the MAC model in Hudson and Manning (2018) (see Appendix B in the Supporting Information).
7 We note that there is a slight drop in mean performance at the six epoch mark. Two of the five random seed runs seem to be causing this drop, while the other three continue increasing. In run 0, the model's performance on both BEHIND and IN FRONT OF drops specifically in the context of questions requiring "yes" answers, while in run 4 the opposite is true, dropping in the context of "no" answers. We do not know why this might be happening in these specific runs, but since most runs do not seem to have this problem, it may be safe to assume that these drops are due to the randomness introduced by different model initializations.

References

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1955–1960). Stroudsburg, PA: Association for Computational Linguistics.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the 2015 IEEE international conference on computer vision* (pp. 2425–2433).

Baharloo, R., Vasil, N., Ellwood-Lowe, M. E., & Srinivasan, M. (2023). Children's use of pragmatic inference to learn about the social world. *Developmental Science*, *26*(3), e13333.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533–581.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93.

Baroni, M. (2021). *On the proper role of linguistically-oriented deep net analysis in linguistic theorizing*. arXiv. https://doi.org/10.48550/arXiv:2106.08694.

Bloom, P. (2002). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

Braine, M. D., & Rumain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology*, *31*, 46–70.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). New York: John Wiley & Sons, Inc.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In A. H.-J. Do, L. Domínguez, & A. Johansen (Eds.),

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Proceedings of the 25th Boston University Conference on language development (pp. 157–168). Somerville, MA: Cascadilla Press.

Chierchia, G., Guasti, M. T., Gualmini, A., Meroni, L., Crain, S., & Foppolo, F. (2004). Semantic and pragmatic competence in children's and adults' comprehension of or. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics. Palgrave Studies in Pragmatics, Language and Cognition* (pp. 283–300). London: Palgrave Macmillan.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637–669.

Clark, A., & Lappin, S. (2011). *Linguistic nativism and the poverty of the stimulus*. New York: John Wiley & Sons.

Clark, E. V. (1977). Strategies and the mapping problem in first language acquisition. In J. Macnamara (Ed.), *Language Learning and Thought* (pp. 147–168). New York: Academic Press.

Clark, E. V. (1993). The mapping problem. *The Lexicon in Acquisition* (pp. 43–66). Cambridge, England: Cambridge University Press.

Clark, E. V. (2003). *First language acquisition*. Cambridge, England: Cambridge University Press.

Clark, H. H. (2018). The primitive nature of children's relational concepts. In J. R. Hayes & R. Brown (Eds.), *Cognition and the development of language* (pp. 260–278). Hoboken, NJ: John Wiley & Sons

Crain, S. (2008). The interpretation of disjunction in universal grammar. *Language and Speech*, *51*, 151–169.

Crain, S. (2012). *The emergence of meaning*. Cambridge, England: Cambridge University Press.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

Donaldson, M., & Balfour, G. (1968). Less is more: A study of language comprehension in children. *British Journal of Psychology*, *59*, 461–471.

Donaldson, M., & Wales, R. J. (1970). On the acquisition of some relational terms. In J. R. Hayes & R. Brown (Eds.), *Cognition and the development of language* (pp. 235–268). New York: John Wiley & Sons.

Farrar, M. J. (1992). Negative evidence and grammatical morpheme acquisition. *Developmental Psychology*, *28*(1), 90.

Fodor, J. D., & Crowther, C. (2002). Understanding stimulus poverty arguments. *The Linguistic Review*, *19*(1-2), 105–145.

Frank, S. L., Monaghan, P., & Tsoukala, C. (2019). Neural network models of language acquisition and processing. In *Human language: From genes and brain to behavior* (pp. 277–293). Cambridge, MA: MIT Press.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (pp. 32–42). Stroudsburg, PA: Association for Computational Linguistics.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? Dataset and methods for multilingual image question answering. In *Proceedings of the conference on neural information processing systems* (pp. 2296–2304). Cambridge, MA: MIT Press.

Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2023). Relational reasoning and generalization using non-symbolic neural networks. *Psychological Review*, *130*(2), 308.

Geurts, B., Katsos, N., Cummins, C., Moons, J., & Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, *25*, 130–148.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3, Speech acts* (pp. 41–58). New York: Academic Press.

Grigoroglou, M., Johanson, M., & Papafragou, A. (2019). Pragmatics and spatial language: The acquisition of front and back. *Developmental Psychology*, *55*, 729–744.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Piscataway, NJ: IEEE.

Hill, F., Clark, S., Blunsom, P., & Hermann, K. M. (2020). Simulating early word learning in situated connectionist agents. In *Proceedings of CogSci 2020* (pp. 875–881). Austin, TX: Cognitive Science Society.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hill, F., Hermann, K. M., Blunsom, P., & Clark, S. (2018). Understanding grounded language learning agents. https://openreview.net/forum?id=ByZmGjkA-

Horowitz, A. C., & Frank, M. C. (2016). Children's pragmatic inferences as a route for learning about the world. *Child Development*, *87*(3), 807–819.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association of Computational Linguistics* (pp. 1725–1744). Stroudsburg, PA: Association for Computational Linguistics.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., & Saenko, K. (2017). Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 804–813). Piscataway, NJ: IEEE.

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. Paper presented at *Proceedings of the International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada*.

Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6700–6709). Piscataway, NJ: IEEE.

Jasbi, M. (2018). *Learning disjunction* [Unpublished doctoral dissertation]. Stanford University, Stanford, CA.

Jasbi, M., & Frank, M. C. (2021). Adults' and children's comprehension of linguistic disjunction. *Collabra: Psychology*, *7*, 27702.

Jasbi, M., Jaggi, A., & Frank, M. C. (2018). Conceptual and prosodic cues in child-directed speech can help children learn the meaning of disjunction. In *Proceedings of CogSci 2018* (pp. 554–559). Austin, TX: Cognitive Science Society.

Jiang, G., Xu, M., Xin, S., Liang, W., Peng, Y., Zhang, C., & Zhu, Y. (2023). Mewl: Few-shot multimodal word learning with referential uncertainty. In *Proceedings of 40th international conference on machine learning* (pp. 15144–15169). Cambridge, MA: JMLR.org.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2901–2910). Piscataway, NJ: IEEE.

Johnston, J. R. (1984). Acquisition of locative meanings: Behind and in front of. *Journal of Child Language*, *11*, 407–422.

Johnston, J. R., & Slobin, D. I. (1979). The development of locative expressions in English, Italian, Serbo-Croatian and Turkish. *Journal of Child Language*, *6*, 529–545.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kraljevič, J. K., Hrzica, G., Grohmann, K. K., Skordi, A., de López, K. J., Sundahl, L., van Hout, A., Hollebrandse, B., Overweg, J., Faber, M., van Koert, M., Smith, N., Vija, M., Zupping, S., Kunnari, S., Morisseau, T., Rusieshvili, M., Yatsushiro, K., Fengler, A., Varlokosta, S., Konstantzou, K., Farby, S., Guasti, M. T., Vernice, M., Okabe, R., Isobe, M., Crosthwaite, P., Hong, Y., Balčiūnienė, I., Nizar, Y. M. A., Grech, H., Gatt, D., Cheong, W. N., Asbjørnsen, A., von Koss Torkildsen, J., Haman, E., Miękisz, A., Gagarina, N., Puzanova, J., Anđelković, D., Savić, M., Jošić, S., Slančová, D., Kapalková, S., Barberán, T., Özge, D., Hassan, S., Chan, C. Y. H., Okubo, T., van der Lely, H., Sauerland, U., & Noveck, I. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, *113*, 9244–9249.

Klatzky, R. L., Clark, E. V., & Macken, M. (1973). Asymmetries in the acquisition of polar adjectives: linguistic or conceptual? *Journal of Experimental Child Psychology*, *16*, 32–46.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Proceedings of the 35th Conference on Neural Information Processing Systems* (Vol. 35, pp. 22199–22213). Curran Associates, Inc. Red Hook, NY.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*, 32–73.

Kuczaj, S. A., & Maratsos, M. P. (1975). On the acquisition of front, back, and side. *Child Development*, 202–210.

Kuhnle, A., & Copestake, A. (2019). The meaning of "most" for visual question answering models. In *Proceedings of the 2019 Association for Computational Linguistics Workshop BlackboxNLP: Analyzing and Interpreting neural networks for NLP* (pp. 46–55). Stroudsburg, PA: Association for Computational Linguistics.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the international conference on machine learning* (pp. 2873–2882). Cambridge, MA: MIT Press.

Lappin, S. (2021). *Deep learning and linguistic representation*. Boca Raton, FL: CRC Press.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissociating language and thought in large language models: A cognitive perspective*. arXiv. https://doi.org/10.48550/arXiv:2301.06627

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 27th international conference on neural information processing systems* (pp. 1682–1690). Cambridge, MA: MIT Press.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*, 30046–30054.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. Paper presented at the *Proceedings of the International Conference on Learning Representations 2019, New Orleans, LA*.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*, 53–85.

Morris, B. J. (2008). Logically speaking: Evidence for item-based acquisition of the connectives AND & OR. *Journal of Cognition and Development*, *9*(1), 67–88.

Neimark, E. D. (1970). Development of comprehension of logical connectives: Understanding of "or". *Psychonomic Science*, *21*, 217–219.

Nikolaus, M., & Fourtassi, A. (2021). Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on computational natural language learning* (pp. 391–407). Stroudsburg, PA: Association for Computational Linguistics.

Palermo, D. S. (1973). More about less: A study of language comprehension. *Journal of Verbal Learning and Verbal Behavior*, *12*, 211–221.

Papadimitriou, I., & Jurafsky, D. (2023). Injecting structural hints: Using language models to study inductive biases in language learning. In H. Bouamor, J. Pino, & K. Bali, (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8402–8413). Stroudsburg, PA: Association for Computational Linguistics.

Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology*, *16*, 278–291.

Pearl, L. (2023). Computational cognitive modeling for syntactic acquisition: Approaches that integrate information from multiple places. *Journal of Child Language*, *50*(6), 1353–1373.

Penner, S. G. (1987). Parental responses to grammatical and ungrammatical child utterances. *Child Development*, *58*(2), 376–384.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence* (Vol 32, pp. 3942–3951). Palo Alto, CA: AAAI Press.

Piaget, J., & Inhelder, B. (1967). *The child's conception of space*. New York: W. W. Norton and Company.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz. lingbuzz*, *7180*. https://lingbuzz.net/lingbuzz/007180

Pillai, N., Matuszek, C., & Ferraro, F. (2021). Neural variational learning for grounded language acquisition. In *Proceedings of the 30th IEEE international conference on robot & human interactive communication (RO-MAN)* (pp. 633–640). Piscataway, NJ: IEEE.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive Science*, *47*, e13334.

Portelance, E., & Jasbi, M. (2023). *The roles of neural networks in language acquisition*. PsyArXiv. b6978. https://osf.io/preprints/psyarxiv/b6978

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge, MA: MIT Press.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* (pp. 2953–2961). Cambridge, MA: MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181–203.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the Child Language Data Exchange System. *Behavior Research Methods*, *51*(4), 1928–1941.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, *30* (pp. 4974–4983). Red Hook, NY: Curran Associates Inc.

Saxton, M. (1997). The contrast theory of negative input. *Journal of Child Language*, *24*(1), 139–161.

Singh, R., Wexler, K., Astle-Rahim, A., Kamawar, D., & Fox, D. (2016). Children interpret disjunction as conjunction: Consequences for theories of implicature and child development. *Natural Language Semantics*, *24*, 305–352.

Skordos, D., Feiman, R., Bale, A., & Barner, D. (2020). Do children interpret 'or' conjunctively? *Journal of Semantics*, *37*, 247–267.

Snow, C. E., & Ferguson, C. A. (1977). *Talking to children: Language input and acquisition*. Cambridge, England: Cambridge University Press.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176–190.

Tieu, L., Yatsushiro, K., Cremers, A., Romoli, J., Sauerland, U., & Chemla, E. (2017). On the role of alternatives in the acquisition of simple and complex disjunctions in French and Japanese. *Journal of Semantics*, *34*, 127–152.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Townsend, D. J. (1974). Children's comprehension of comparative forms. *Journal of Experimental Child Psychology*, *18*, 293–303.

Tsuji, S., Cristia, A., & Dupoux, E. (2021). SCALa: A blueprint for computational models of language acquisition in social context. *Cognition*, *213*, 104779.

Wang, R., Mao, J., Gershman, S. J., & Wu, J. (2021). Language-mediated, object-centric representation learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2033–2046). Stroudsburg, PA: Association for Computational Linguistics.

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Warstadt, A., & Bowman, S. R. (2023). What artificial neural networks can tell us about human language acquisition. In S. Lappin & J.-P. Bernardy (Eds.), *Algebraic structures in natural language* (pp. 17–60). Boca Raton, FL: CRC Press, Taylor & Francis Group.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 35th Conference on Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). Curran Associates, Inc. Red Hook, NY.

Windmiller, M. (1973). *The relationship between a child's conception of space and his comprehension and production of spatial locatives* [Unpublished doctoral dissertation], University of California, Berkeley.

Zellers, R., Holtzman, A., Peters, M., Mottaghi, R., Kembhavi, A., Farhadi, A., & Choi, Y. (2021). PIGLeT: Language grounding through neuro-symbolic interaction in a 3D World. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 2040–2050). Stroudsburg, PA: Association for Computational Linguistics.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and Yang: Balancing and answering binary visual questions. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 5014–5022). Piscataway, NJ: IEEE.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1: Mean F1 score on previous SAME probes overall. Shading represents standard deviation across 3 models.

Figure S2: Mean F1 score on previous SAME probes by answer type.

Figure S3: Mean F1 score on SAME probe overall. Shading represents standard deviation across 5 models.

Figure S4: F1 score on SAME probe by answer type. Shading represents standard deviation across 5 models

15516709, 2024, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cogs.13448 by University Of California - Irvine, Wiley Online Library on [11/09/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License