Learning the meanings of 'hard' words like *more* and *or* may not be so hard after all

Eva Portelance, Michael C. Frank, Dan Jurafsky

January 2023

Abstract

Function words may be one of the hardest parts on language to learn since understanding their meaning requires developing complex reasoning skills, like logical, numerical, spatial, and relational reasoning. Using both linguistic and visual context, Visual question answering (VQA) models learn to reason about abstract relations making them a good testing bed for function word learning. Given the abstract nature and complexity of these words, they represent a great test case for whether a non-symbolic learning algorithm can acquire these from data. In this paper, we propose to study how VQA models learn function words to better understand how the meanings of these words can be learnt both by models and children. Using this approach, we show that recurrent VQA models trained on visually grounded language learn gradient semantics as opposed to threshold-based semantics for a series of function words requiring spacial 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. Finally, we show that word learning difficulty is dependent on their frequency in models' input. Our findings offer evidence that (1) it is possible to learn the meanings of these words in visually grounded context by using non-symbolic general learning algorithms, without any prior knowledge of linguistic meaning, supporting a usage-based theories of their acquisition in children over innateness proposals. Additionally, our results confirm that (2) the order in which children tend to learn these words may be dependent on their frequency in the input rather than other inherent word properties.

1 Background

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* (Morris, 2008; Jasbi, Jaggi, & Frank, 2018); 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 (Windmiller, 1973; Kuczaj & Maratsos, 1975; E. V. Clark, 1977), and the degree of occlusion between two objects (Johnston, 1984; Grigoroglou, Johanson, & Papafragou, 2019). These differences in order of acquisition represent learning outcomes which 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 developmental stages, or maturation, 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 learnt through experience. Children learn these words using non-symbolic general learning mechanisms which are not exclusive to language acquisition. Usage-based learning mechanisms predict that frequency of exposure modulates the order in which new words may be learnt.

In this paper, we will consider the acquisition of three types of reasoning skills and the following corresponding function words: (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 hypothesise that these function words can be learnt 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 both whether they can be learnt using non-symbolic general learning algorithms and whether any ordering effects observed follow from the relative frequency of function words in the input. Specifically, we experiment with neural network models learning language in a visual question answering (VQA) 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?".

VQA-type 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 (Mao, Gan, Kohli, Tenenbaum, & Wu, 2019; Pillai, Matuszek, & Ferraro, 2021; Zellers et al., 2021; Wang, Mao, Gershman, & Wu, 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 VQA 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 say and predates VQA 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, probing visually grounded neural network model's representations of the meaning of function words to better understand their acquisition by children has yet to be studied.

VQA tasks use supervised learning – in other words, models have access to the correct answers to training questions and learn by trying to minimize the cross entropy between predicted answers and the correct ones. Importantly, models do not receive direct supervision to learn abstract reasoning or the meanings of function words; these learning outcomes may be incidental to the task and instead could be one of many strategies that models could converge towards to answer the questions correctly. We acknowledge that the learning mechanisms used by VQA models may be different from those used by children. If one believes that children only have access to positive feedback and no negative feedback what-so-ever, then the use of VQA models to study function word learning may seem inappropriate given the supervised nature of models' training. However, many would agree that children do get some forms of indirect or implicit negative evidence, at least tangentially – be it that their desired outcomes are not met when they are misunderstood for example – where some form of supervision is to a certain extent always available (Brown, 1970; Snow & Ferguson, 1977; Penner, 1987; Farrar, 1992; Saxton, 1997; Chouinard & Clark, 2003). The meanings of words may then be learnt indirectly from this evidence and the same may be said for our models. With our experiments, we hope to be able 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. The behaviour of models can then serve as a lower bound for what is also possible in children's acquisition of these same words.

Throughout this paper, we will address the following sets of research questions: (1) What type of semantic representations do VQA models learn for function words? And do these representations generalize to unseen linguistic and visual contexts? (2) 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?

Each of our function words of interest are defined in absolute terms in the CLEVR dataset we use – for example *or* is defined as the logical operator, $A \vee B$, and *more* is defined as greater than, $A > B$. In practice however, most of these words have much more gradient meanings when used in naturalistic contexts. Thus, we will first try to answer our first set of research questions and probe what type of semantic representations VQA models learn for these function words. If models can learn gradient representations that generalize to novel contexts using simple learning algorithms, then we can offer the first 'proof of concept' evidence – showing us what is in practice possible – in favor of usage-based theories for the acquisition of these function words.

Frequency or word predictability is a known predictor of the order in which children acquire words (Goodman, Dale, & Li, 2008; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Braginsky, Yurovsky, Marchman, & Frank, 2019). There may however be other factors in relation with or independent from frequency that make learning the meaning of certain function words harder than others. For example, 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. She suggests that 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 will try to answer our second set of research questions by exploring whether the order in which these words are learnt is a function of how frequent they are in the input or if there may be other factors which makes 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 VQA 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 distributions and train models on different subsets of the training data with various function word frequencies to analyse whether the ordering effects initially observed are solely mediated by frequency or if other more conceptual factors play a role.[1]

2 On children's acquisition of function words

For each of the word pairs and their respective reasoning skills considered in this study, we review what is currently known and debated about their acquisition in the child language learning literature.

2.1 Logical reasoning

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 learnt 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* (Morris, 2008; Jasbi et al., 2018). There is however some debate about the order

[1] All of the data, models, and experiment code presented in this paper are publicly available at `www.github.com/evaportelance/vqa-function-word-learning`.

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).

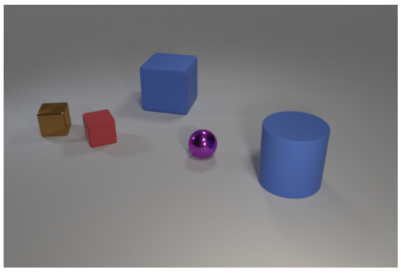
All of the experimental results showing that children understand *or* inclusively still leave unanswered the question of how they come to learn the meaning of this word in the first place. Crain (2008, 2012) argues that these results are in fact evidence in favor of a logical nativist explanation since,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

2.2 Spatial reasoning

Children learn the meaning of the locative preposition *behind* before they do *in front of* (Johnston & Slobin, 1979; Johnston, 1984). 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 analysing 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 analyse 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. 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 versus being behind a block) (Windmiller, 1973; Kuczaj & Maratsos, 1975; E. V. Clark, 1977). 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 spacial reasoning.

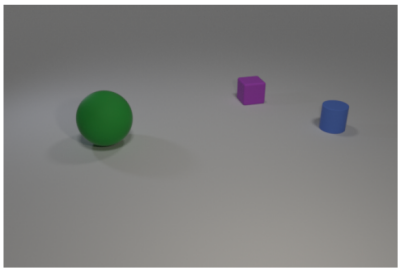
2.3 Numerical reasoning

Quantifiers have been found to follow quite robust acquisition ordering effects cross-linguistically (Katsos et al., 2016). As for the comparative quantifiers *more (than)* and *fewer (than)*, the meaning of *more* has repeatedly been found to be learnt earlier than *fewer/less* by children (Donaldson &



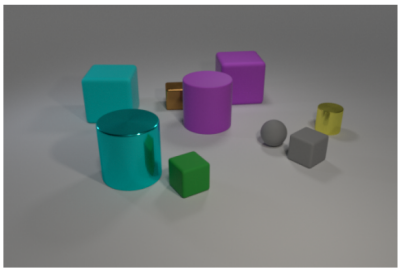
Q: Are there more brown shiny objects behind the large rubber cylinder than gray blocks?
A: yes

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

Figure 1: Example images and corresponding questions taken from CLEVR dataset.

Balfour, 1968; Palermo, 1973; Donaldson & Wales, 1970; Townsend, 1974; Geurts, Katsos, Cummins, Moons, & Noordman, 2010). Some have also found that children initially interpret *less* as a synomym 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 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.

3 On VQA tasks and the CLEVR dataset

As a testbed for the learnability of the function words listed in the previous section, we will use VQA models trained on the CLEVR dataset (Johnson et al., 2017).

VQA was proposed as a language learning task that is grounded in images and requires models to develop abstract reasoning skills (Malinowski & Fritz, 2014; Antol et al., 2015; Gao et al., 2015; Ren, Kiros, & Zemel, 2015). VQA 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 Figure 1). Generating the correct answers often necessitates using 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 VQA tasks lend themselves well to the study of function word learning using neural networks.

Initial VQA tasks used datasets that were produced by having human annotators come up with questions for images (Malinowski & Fritz, 2014; Antol et al., 2015; Gao et al., 2015; Krishna et al., 2017). However, as the first VQA 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 (Johnson et al., 2017; Hudson & Manning, 2019) and images (Johnson et al., 2017) were created. CLEVR is one such dataset, containing generated images of scenes from a 3D block-world and constructed questions.

Though using a dataset like CLEVR clearly has its advantages, it is important to note that it is not natural language and does not have all the same properties as the speech children are exposed to. The language in CLEVR is template based; by contrast, children's input is composed of much richer signal. We acknowledge that these differences mean that we may be missing some of the information found in natural language that children leverage to learn new words. However, working within controlled and simplified learning environment allows us to parse the relations that exist between models' input and their learning outcomes. Additionally, it helps us place a lower bound on the necessary learning conditions for the meanings of function words we are interested in.

3.1 The CLEVR dataset

The CLEVR dataset is composed of questions paired with images like those illustrated in Figure 1. The images are all of complex scenes in a block-world involving static objects placed on a 3D grey 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 does 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[2]. For each image, there is also a 'scene' file available which contains all of the metadata that was sampled to generate the image, including the number of objects, their properties, as well as their coordinates in the image and their relative position to one another.

Each image is paired with a set of questions like those in Figure 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).

[2]The CLEVR dataset also contains a test set, but since this dataset was designed as for benchmarking task, the meta-information for test images isn't 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.

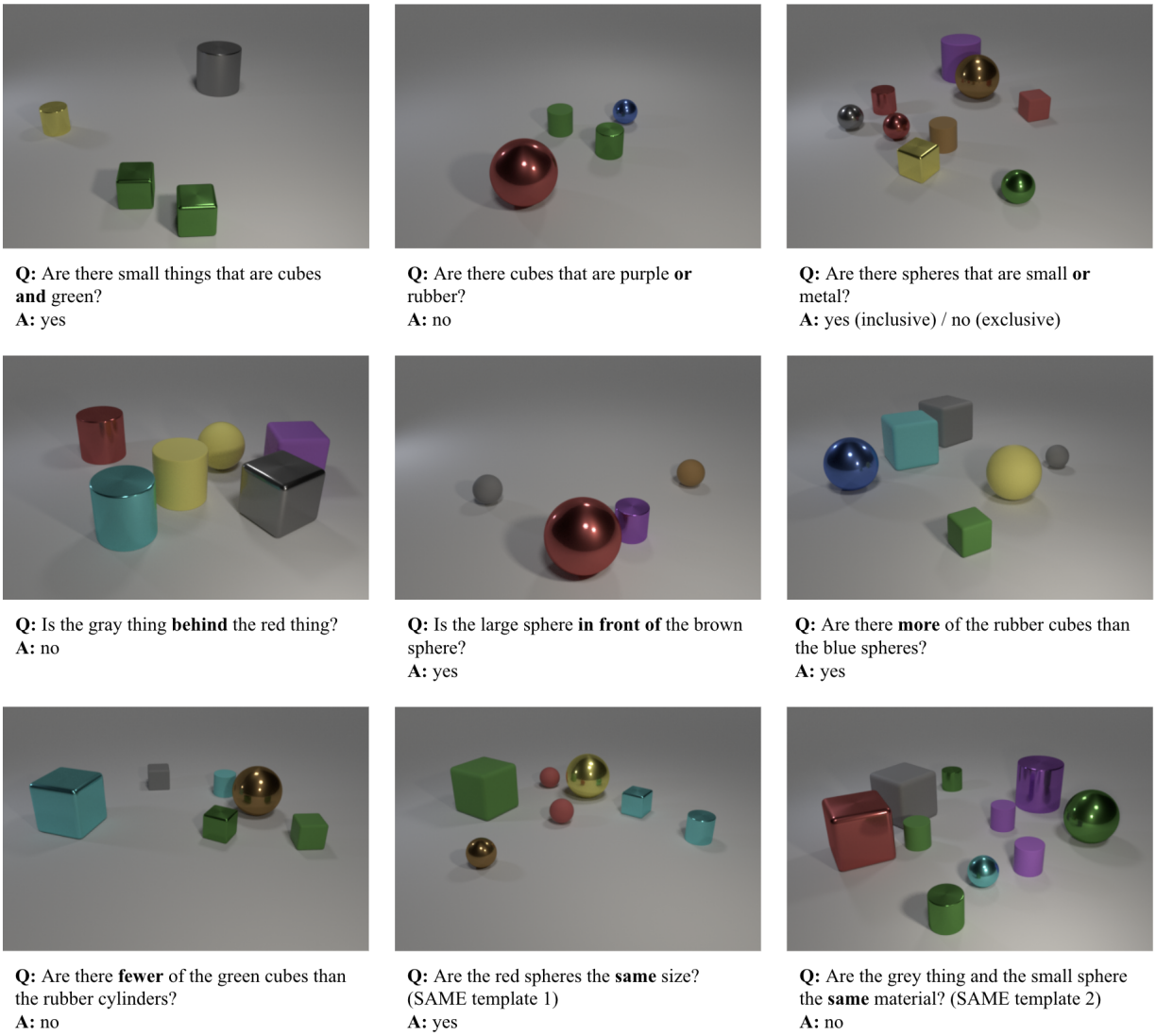


Figure 2: Example image-question pairs from semantic probes.

4 Semantic probes and the evaluation task

We consider the following three types of reasoning skills and their corresponding function words in our experiments: logical reasoning and the connectives *and* and *or*; spatial reasoning and the prepositions *behind* and *in front of*; numerical reasoning and the quantifiers *more* and *fewer*, all of which are available vocabulary words in CLEVR[3]. We designed a semantic probe evaluation task to determine whether the models were able to learn meaningful representations for each of these words. 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*.

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 an-

[3]In appendix B we also include some experiments with relational reasoning and the adjective *same*.

swer 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 Figure 2. The probes are an out-of-distribution generalization task for the models, since the questions 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 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.

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 noun (thing, sphere, cylinder, cube) optionally preceded by a modifier referring to its color, material or size[4]. These templates do not have any presuppositions, so 10 images were randomly sampled for each one, totalling 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 wasn't 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 there 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.

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, we identified all the images which 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 difference in y coordinates, we determined if an object was behind or in front of another. Image-question pairs

[4]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.

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 these same (x, y, z) coordinates.

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 to each image. We identified all of the objects which were part of X and Y referent categories and then compared their cardinality. 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 were 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.

In each of the experiments that follow, we use these probes to evaluate how much models have learnt 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 analyse what they are learning over time.

5 A recurrent reasoning model

There are a variety of models that exist for completing VQA tasks, all of which must include both visual and a linguistic processing units. For these experiments, we chose to use a the state-of-the-art VQA model which has been shown to have one of the top performance scores on the original CLEVR task, the MAC (Memory, Attention, and Composition) model from Hudson & 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 out-of-distribution questions like our probes. Additionally, the MAC model is a reccurrent attention-based neural network model with generic and homogeneous structure. It eliminates the possibility of the model itself introducing any forms of symbolic structural biases, which is important since it will serve as an example of non-symbolic learning for our hypotheses testing. Its structure is illustrated in Figure 3 and described below.

Following previous approaches to the CLEVR task (Hu, Andreas, Rohrbach, Darrell, & Saenko, 2017; Santoro et al., 2017; Perez, Strub, De Vries, Dumoulin, & Courville, 2018), the MAC model creates a representation for images by first extracting a fixed set of features from ResNet-101 (He, Zhang, Ren, & Sun, 2016), pre-trained on ImageNet (Russakovsky et al., 2015). It then processes

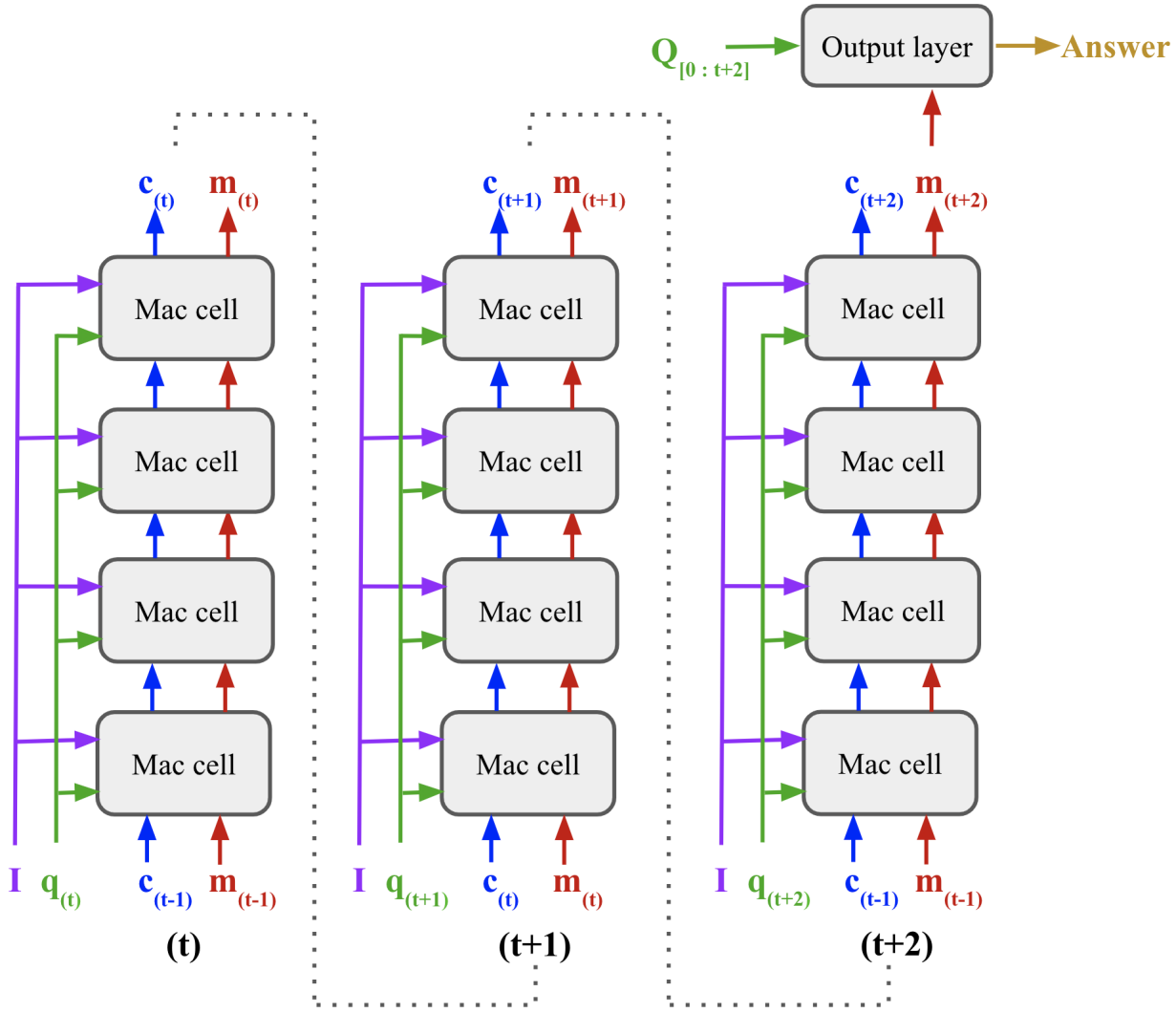


Figure 3: A 4-layer MAC model with recurrent processing states, where $\boldsymbol{I}$ is the processed image, $\boldsymbol{Q} = [\boldsymbol{q}_0, ..., \boldsymbol{q}_{t+2}]$ is the question which contains the output of a biLSTM at for each state, or word, while $\boldsymbol{c}_{(t)}$ and $\boldsymbol{m}_{(t)}$ are soft attention maps from state $\boldsymbol{t}$ over the question and the image respectively. At the final state, the final attention map over the image, here $\boldsymbol{m}_{(t+2)}$ and the output states of the biLSTM over all of the words in the question, $\boldsymbol{Q}$, go through an output cell to produce a distribution over possible answers.

images through two CNN layers, producing its final image representation $\boldsymbol{I}$. As for questions, at each state t, the current word is first converted to an embedding vector and then goes through a single layered biLSTM, the output of which is then used as a word representation $\boldsymbol{q}_t$, taking into account linguistic contextual information.

The MAC model uses custom recurrent cells (MAC cells) which incrementally process information as the model makes its way through each word in the question. Unlike previous models, the MAC model uses soft attention mechanisms (Xu et al., 2015) over both the word and the image representations to learn to prioritise and 'reason' about different parts of its input during processing. The MAC cell takes into consideration the previous state's soft attention maps over the previous word's representation $\boldsymbol{c}_{(t-1)}$ (called the control state in Hudson & Manning, 2018) as well as the image representation $\boldsymbol{m}_{(t-1)}$ (called the memory state in the original paper) – these represent the previous hidden states. At each time step, in addition to these previous hidden states, the model uses the current word and the image representations to produce new soft attention maps over the word and image to be used at the next time step. At the final step, the model integrates the final soft attention map over the image representation (the final memory state) with all of the word's final representations to generate an answer. Answers always consist of a single word which can be any word from the model's shared question and answer vocabulary. For further details about the inner workings and modules contained in each MAC cell that generate these attention maps, we refer the reader to the model's original paper (Hudson & Manning, 2018).

The best version of the MAC model originally reported used 12 layers of recurrent MAC cells before the output layer, however the authors found that very similar performance could be achieved with as few as 4 layers (test accuracy 97.9%). Thus, we chose to use this smaller and more efficient version of the model for our experiments – see Figure 3 for a visualization of the model's layers. We otherwise kept all other hyperparameters the same as the ones used in the main version of the model from the original paper[5].

6 Experiment 1: The emergence of gradient semantics

With this first experiment, we hope to answer our first set of research questions: what type of semantic representations do VQA models learn for 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?

6.1 Setup

To help answer these questions, we train five MAC models on the original CLEVR training data for 25 epochs, initialized using different random seeds. Models learn and update using back propagation with the addition of variational dropout on 15% of parameters across the model at each pass. We evaluate their performance on our semantic probes at each epoch and report the mean performance and standard deviation across models for each probe.

[5]For a complete list of hyperparameters and values see appendix A.

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 1: Relative frequencies of each function word pair 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 %

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

6.2 Results

Models reach 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 4-layered MAC models. In the rest of this subsection, we report model performance on our semantic probes for each word pair separately. Note that since our probe questions are out-of-distribution questions, models are not likely to reach as high accuracy scores as they do on within task questions like those seen in the validation set.

Chance performance on the probes is in theory near 0% accuracy since models can produce any word in their vocabulary as the answer to probe questions. However, as the reader will note, models very quickly learn after only a couple batches that existential questions are always answered with either 'yes' or 'no'. Thus, in practice, after the first epoch, we should consider chance to be 50% accuracy, since models are in fact only considering two possible answers to probe questions.

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 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 which 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 which 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 accuracy 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.

AND - OR Probe questions were all of the form 'Are there *X*s that are α and/or β?'. As a reminder, there are four possible truth-conditions associated to 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's consider the overall accuracy of models on probes in non-ambiguous contexts in Figure 4 – in other words, excluding OR probe questions in $(\alpha \wedge \beta)$ contexts, where inclusive and exclusive interpretations of *or* have opposing answers. As we can see in this figure, models perform better than chance on both the AND and OR probes.

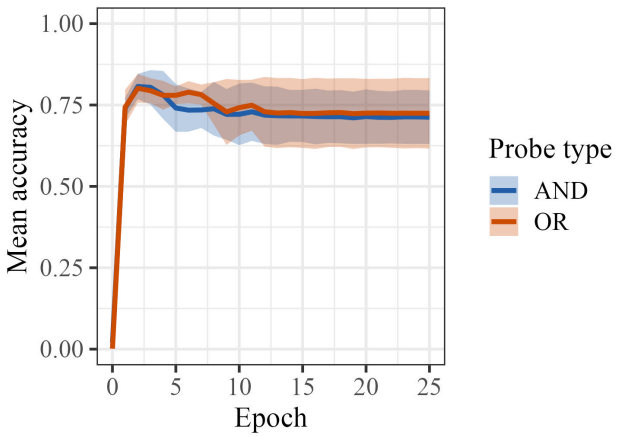


Figure 4: Experiment 1: Mean accuracy on AND - OR probes overall in non-ambiguous questions, shading represents standard deviation across 5 models.

Figure 5 shows the mean accuracy reported in the previous figure as a function of the answer type – 'yes' or 'no' – expected for each question for these probes. Here, we can observe two very interesting patterns: first, there is a clear 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 has no issue recognizing the answer in $(\neg\alpha \wedge \neg\beta)$, they struggle more when one of the conjuncts is true, or 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 Figure 6.

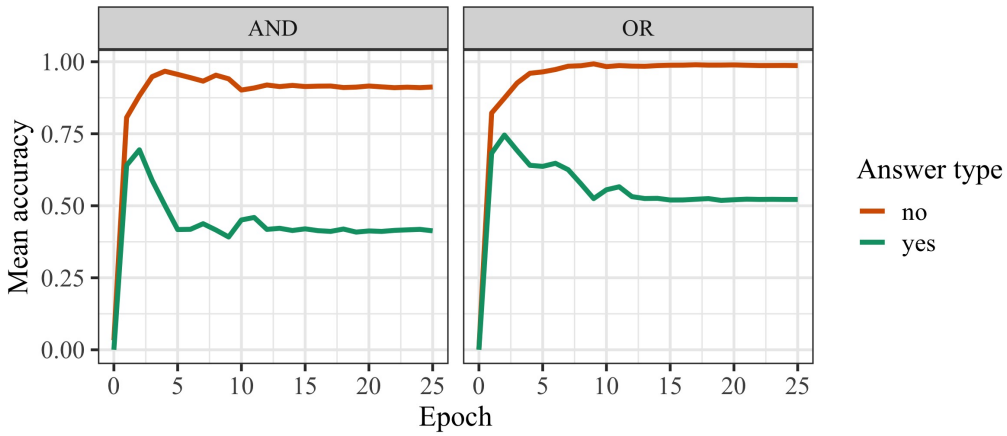


Figure 5: Experiment 1: Mean accuracy on AND - OR probes by answer type in non-ambiguous questions.

In Figure 6, 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*. Yet, as we can see in this figure, though the models initially learn to favor inclusive interpretations as we might expect, as learning progresses, they start to interpret OR as exclusive more and more.

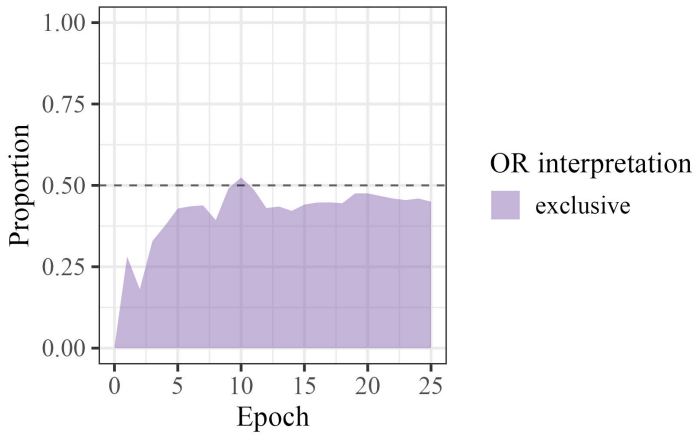


Figure 6: Experiment 1: Proportion of exclusive (versus inclusive) interpretations of OR probe in ambiguous contexts, $(\alpha \wedge \beta)$.

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 overtime 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 case of $(\alpha \wedge \neg\beta)$ and $(\neg\alpha \wedge \beta)$ 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 of $(\alpha \wedge \beta)$ where the same answer is expected, it may instead be leading to a form of pragmatic reasoning where opposing logical operators should also have opposing answers, resulting in the rise of exclusive *or*. We explore this hypothesis further in the next experiment.

BEHIND - IN FRONT OF Probe questions are all of 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*. Figure 7 shows the overall accuracy of the models on both probes. There is more variation across random seed runs, though both BEHIND and IN FRONT OF seem to be learnt equally well within runs and performance is generally above chance.[6] 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.

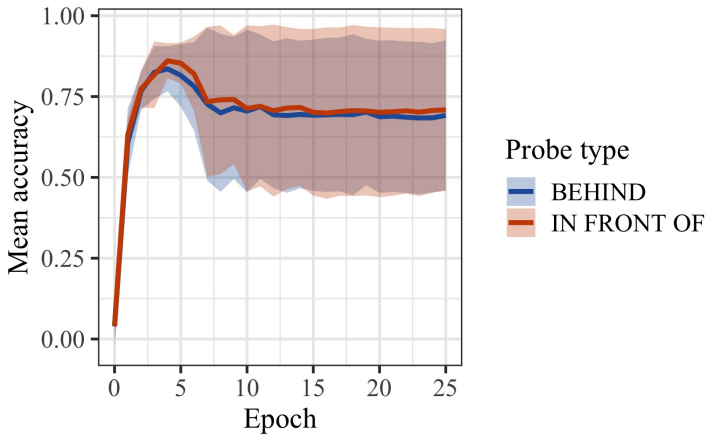


Figure 7: Experiment 1: Mean accuracy on BEHIND - IN FRONT OF probes overall, shading represents standard deviation across 5 models.

As with the previous probes, we also consider the mean accuracy of models as a function of the answer type. Whether the context required a 'yes' or 'no' answer did not seem to matter for these probes as much as it did for others; models performed just as well in either context overall.

In Figure 8, 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.

[6]We note that there is a slight drop in mean performance at the 6 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.

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 a 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 gradient 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*.

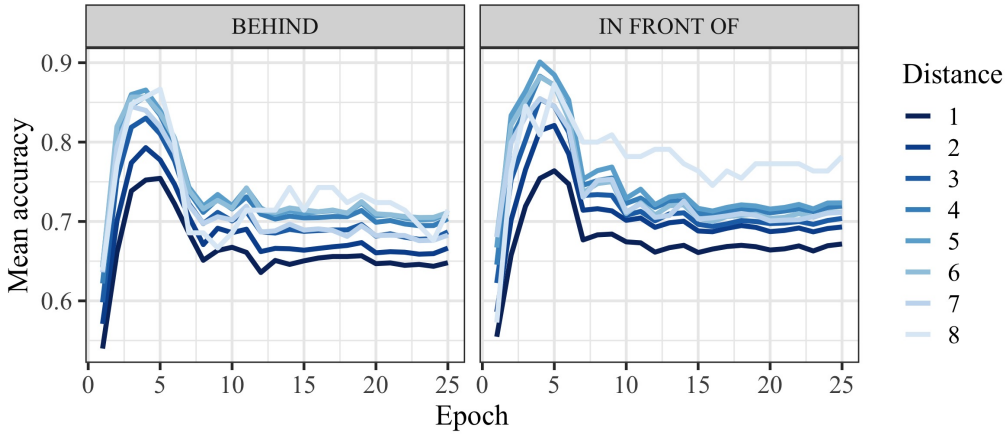


Figure 8: Experiment 1: Mean accuracy on BEHIND - IN FRONT OF probes as a function of the Euclidean distance between referents.

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 learnt 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.

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'. Figure 9 presents the overall accuracy of models on both probes. This initial plot suggests that MORE is learnt first and may be overall easier than FEWER.

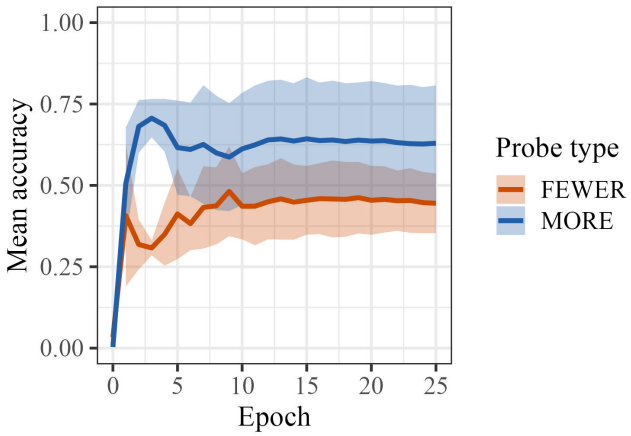


Figure 9: Experiment 1: Mean accuracy on MORE - FEWER probes overall, shading represents standard deviation across 5 models.

When we consider the average performance of models as a function of the expected answers a more nuanced story emerges, Figure 10. Clearly, contexts where 'yes' answers are expected are much easier than 'no' contexts, reaching accuracies way above chance for 'yes' and at or below chance for 'no'. As a reminder, 'no' contexts include $\neg(|X| > |Y|)$ for MORE and $\neg(|X| < |Y|)$ for FEWER, but also contexts where $|X| = |Y|$. Additionally, again we see that FEWER may be more difficult to learn than MORE.

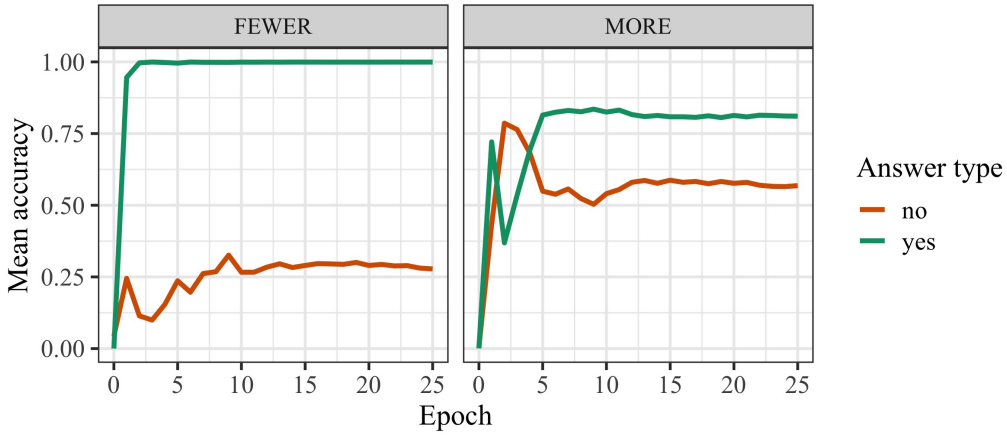


Figure 10: Experiment 1: Accuracy on MORE - FEWER probes by answer type.

Next, we plot accuracy on probes as a function of the absolute difference between the number of Xs and Ys, $absolute(|X| - |Y|)$, Figure 11. Models clearly struggle with both MORE 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 two plots seems to be isolated to the contexts where $|X| = |Y|$.

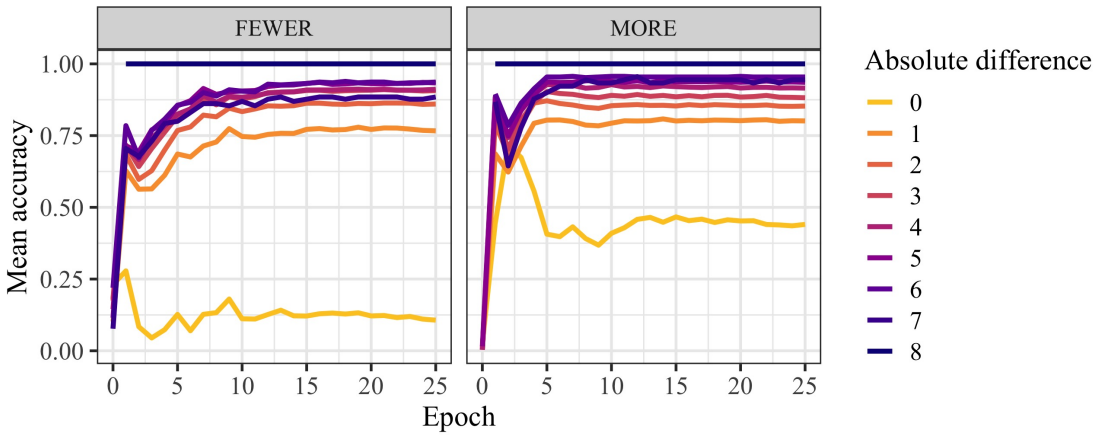


Figure 11: Experiment 1: Mean accuracy on MORE - FEWER probes by absolute difference in the number of objects in each referent class.

In fact, if we remove all probe questions where $|X| = |Y|$ and consider the overall accuracy of models again in Figure 12, we see a very different picture than our original Figure 9. Models have almost equally high performance on both probes, still learning *more* slightly earlier than *fewer*.

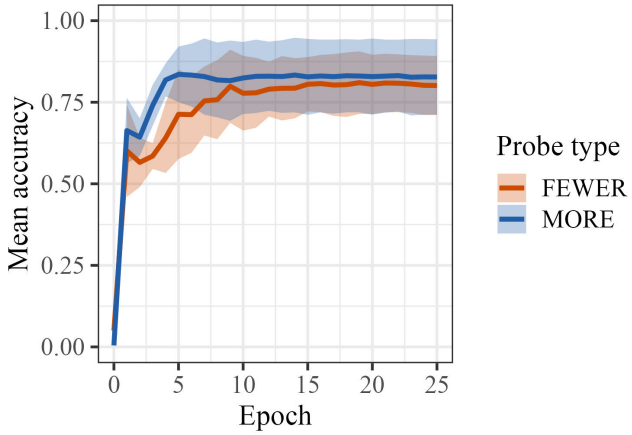


Figure 12: Experiment 1: Mean accuracy on MORE - FEWER probes overall excluding context where $|X| = |Y|$, shading represents standard deviation across 5 models.

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 difference 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 hypothesise 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.

7 Experiment 2: The effect of alternatives on reasoning

Following our results from the previous experiment, we hypothesized that models could be considering alternative questions and answers which use opposing or parallel function words when they compute the probability of the answer to a given question. This form of 'reasoning about alternatives' could 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 expect opposing answers, AND - OR and MORE - FEWER pairs both have contexts where they expect the same answer and others where they do not. Given their similar distributions in the input as well as the partial overlap in their answer spaces, models may consider the existence of alternative expressions leading to uncertainty in their predictions in one of two ways. First, if models observe that say *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 'reasoning' about alternative propositions. Thus, this second experiment will help us test this theory and serves as a follow-up to the first experiment.

7.1 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 *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 *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 are 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 performance on probes at each epoch.

7.2 Results

We report the mean performance and standard deviation across models for each probe. Since the results for the OR probe and AND probe come from different models trained on different subsampled datasets, we report their performances separately for this experiment.

OR Probe results were obtained by training models on a version of CLEVR where we removed all questions containing the word *and*. Figure 13 compares the overall results from experiment 1 where the alternative *and* was included, and the results from this experiment without the alternative expression. Models reach a higher overall accuracy on non-ambiguous OR probe questions when trained on data without *and*.

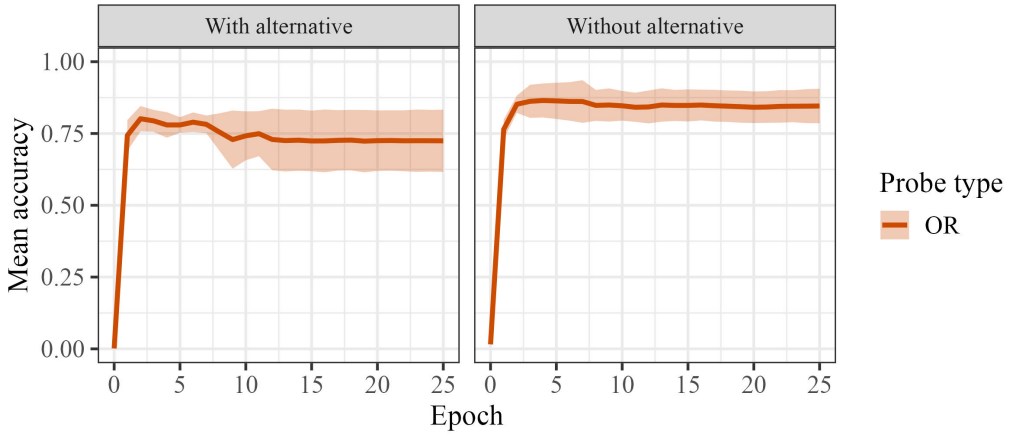


Figure 13: Experiment 2: Mean accuracy on OR probe overall in non-ambiguous questions when trained on data with the alternative expression *and* from experiment 1 versus without this alternative in experiment 2. Shading represents standard deviation across 5 models.

Comparing model performance when trained with and without *and* as a function of the answer type expected in Figure 14, 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, instead showing high accuracy regardless of the truth value contexts.

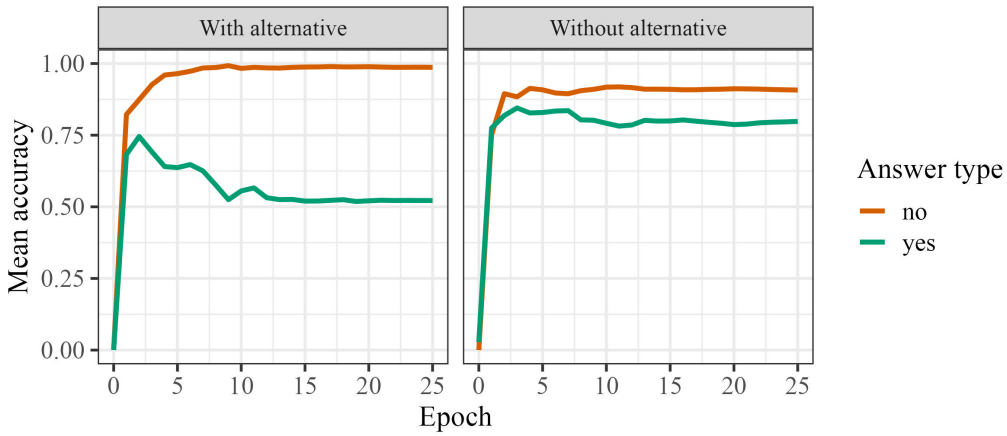


Figure 14: Experiment 2: Mean accuracy on OR probe by answer type in non-ambiguous questions when trained on data with the alternative expression *and* from experiment 1 versus without this alternative in experiment 2.

As for probe questions containing *or* in ambiguous contexts where inclusive-or and exclusive-or interpretations predict opposing answers, we no longer see a strong progressive rise in exclusive interpretations, instead settling with around 70% of ambiguous questions being answered with inclusive 'yes' answers, Figure 15.

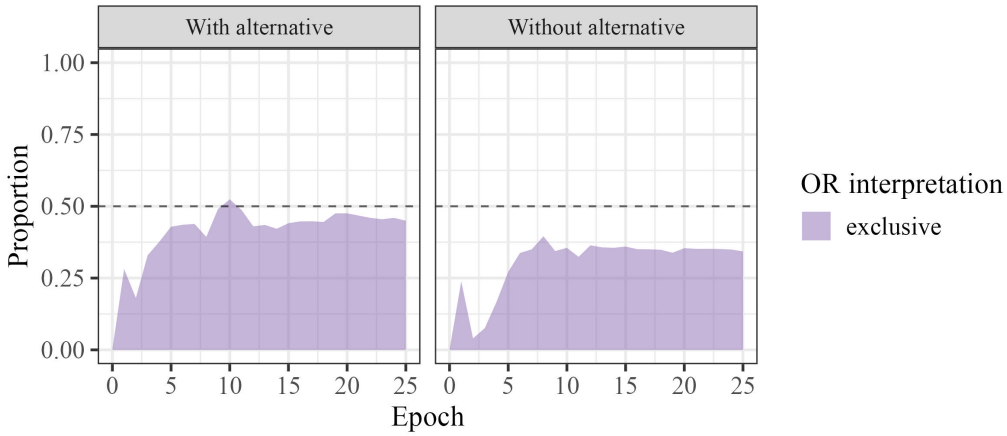


Figure 15: Experiment 2: Proportion of exclusive (versus inclusive) interpretations of OR probe in ambiguous contexts, $(\alpha \wedge \beta)$, when trained on data with the alternative expression *and* from experiment 1 versus without this alternative in experiment 2.

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 pragmatic competition between *or* and the available alternative *and*.

AND Probe results come from models trained on a subsampled version of CLEVR where all instances of *or* have been removed. Once again, models had better overall accuracy on AND probe questions when the alternative logical connective was removed than when both were present in experiment 1, see Figure 16.

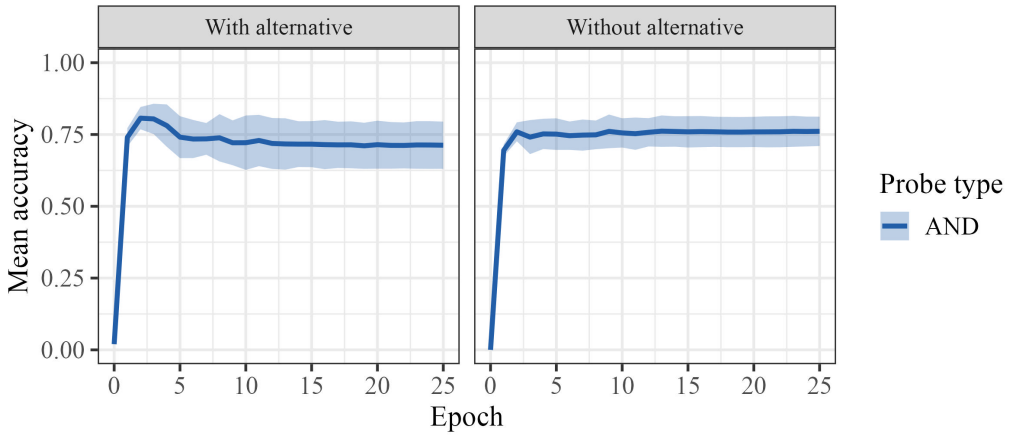


Figure 16: Experiment 2: Mean accuracy on AND probe overall when trained on data with the alternative expression *or* from experiment 1 versus without this alternative in experiment 2. Shading represents standard deviation across 5 models.

Performance as a function of answer type expected also show us that in the absence of *or*, models learn to correctly interpret *and* regardless of the truth value context, Figure 17.

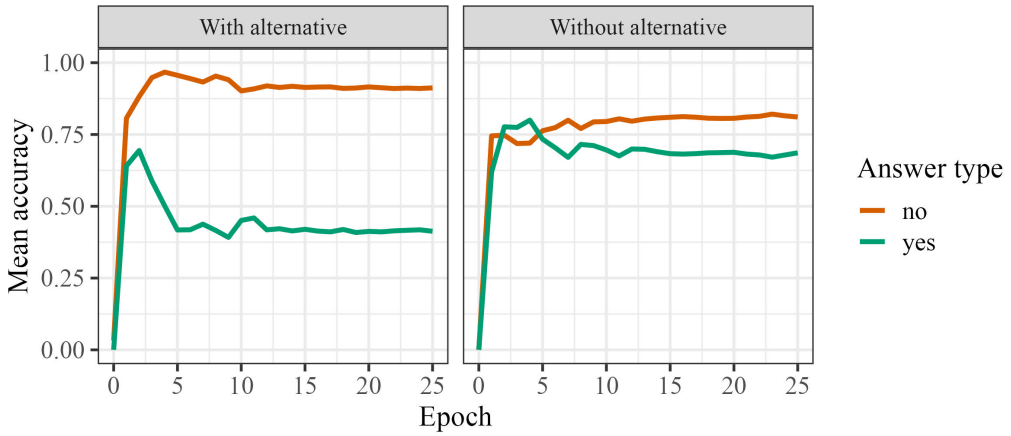


Figure 17: Experiment 2: Mean accuracy on AND probe by answer type when trained on data with the alternative expression *or* from experiment 1 versus without this alternative in experiment 2.

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.

MORE - FEWER Probe results from experiment 1 both 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 inthe 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. Figure 18 shows the overall performance of these models on both probes when trained with and without this alternative *equal* expression. Accuracy on FEWER questions has definitely risen in comparison to experiment 1, though results for MORE look quite similar.

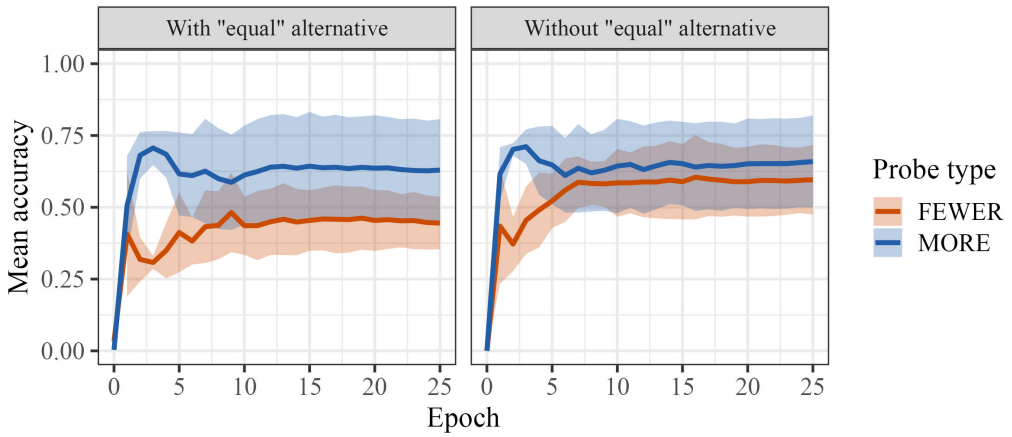


Figure 18: Experiment 2: Mean accuracy on MORE - FEWER probes overall when trained on data with the alternative expression *equal* from experiment 1 versus without this alternative in experiment 2. Shading represents standard deviation across 5 models.

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

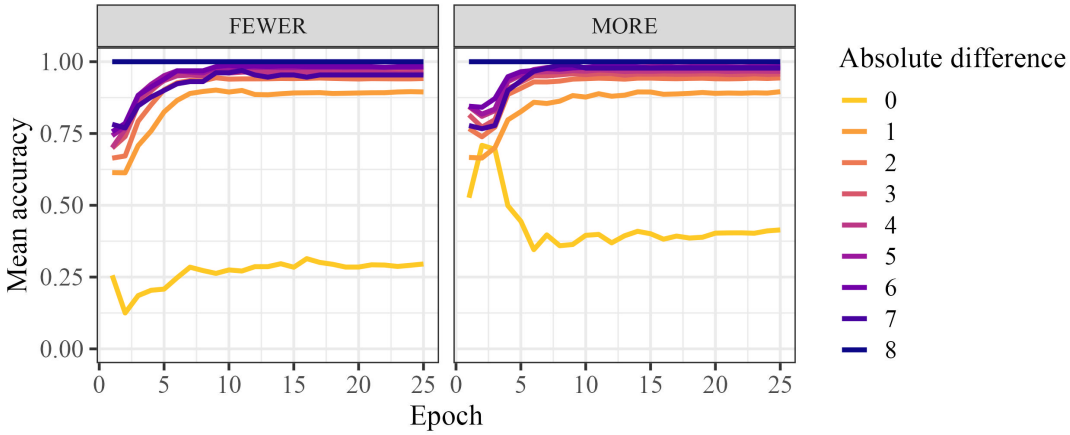
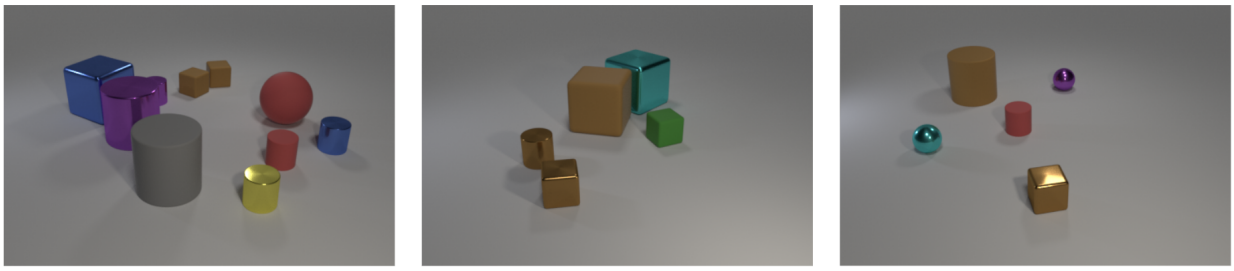


Figure 19: Experiment 2: Mean accuracy on MORE - FEWER probes by absolute difference in the number of objects in each referent class when trained without the alternative *equal* expression.

Unlike with AND and OR, removing the pragmatic alternative did not solve our issue with FEWER and MORE. After carefully scrutinising the training data from CLEVR, it became apparent that *more/fewer* rarely appeared in contexts where $|X| = |Y|$ and only when they were part of more complex question templates. Figure 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.



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)

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

In this experiment, we found that the presence, or absence, of a pragmatic alternative can affect how models learn the meaning of logical connectives *and* and *or*. Next, we will evaluate how the frequency of different function words may also affect how models learn their meanings.

8 Experiment 3: The effect of frequency on learning

This third and final experiment considers the effect of word frequency in children's and models' input on the order in which function words are learnt. We address our second set of research questions: 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?

8.1 Setup

Again, like in the previous experiments, we train five MAC models initialized using different random seeds for a total of 25 epochs and consider their performance on semantic probes throughout training. Our main manipulation which 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 function words we are interested in 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

	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%

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

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 one to isolate all utterances that were not said by the child, but instead represented the linguistic input the child was exposed to, resulting in a corpus of child-directed and child-adjacent utterances. 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, while the new subsampled version of CLEVR contained 9,652,086.

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 inadvertedly 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. Table 3 shows the exact word counts and frequencies of both the CHILDES and subsampled CLEVR training datasets.

8.2 Results

For each of the function word pairs, we plot the mean accuracy on semantic probes as well as standard deviation across five random seed runs for models trained on this new subsampled CLEVR dataset.

AND - OR These words have a very uneven distribution in the training data, *and* being much more prominent than *or* in children's input. Figure 21 shows 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). 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.

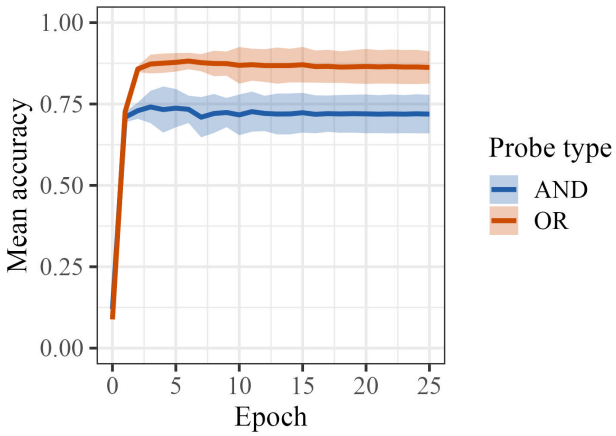


Figure 21: Experiment 3: Mean accuracy on AND - OR probes overall in non-ambiguous questions when trained of subsampled CLEVR, shading represents standard deviation across 5 models.

This observation is confirmed when we consider models' mean accuracy as a function of the answer type expected, 'yes' or 'no', where OR probe performance is eventually about the same regardless of context, Figure 22. As for the AND probe, the models seem to be performing better than they were in experiment 1, though we still see an imbalance in performance between 'yes' answer contexts and 'no' answer contexts, where the model struggles more in the former.

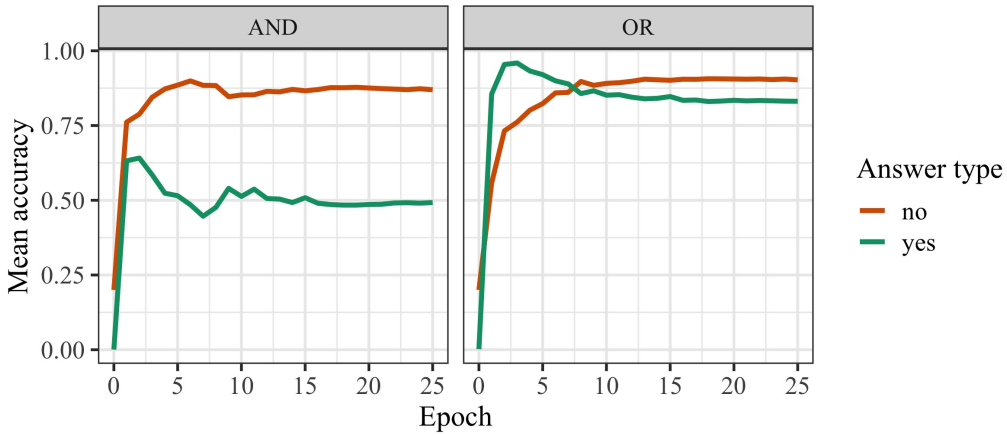


Figure 22: Experiment 3: Mean accuracy on AND - OR probes by answer type in non-ambiguous questions when trained of subsampled CLEVR.

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 Figure 23.

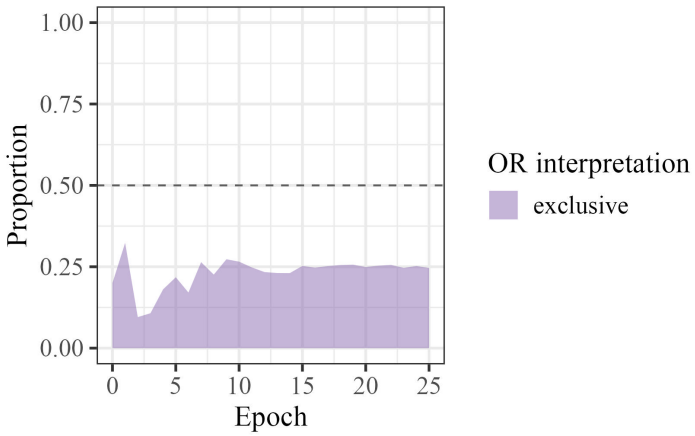


Figure 23: Experiment 3: Proportion of exclusive versus inclusive interpretations of OR probe in ambiguous contexts, $(\alpha \wedge \beta)$, when trained of subsampled CLEVR.

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 this is not what we see. Furthermore, if the effect of having possible pragmatic alternatives was also proportional to the frequency of these alternatives in the input for the models, we might also expect to see a stronger effect of pragmatic reasoning 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 lead to more uncertainty overall. This explanation is further supported by the much smaller standard deviations we see in Figure 21 for both AND and OR performance as opposed to the same plot from experiment 1, Figure 4. 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 contexts that resembled our probe template more, such that the models now had less uncertainty specifically about the meaning of *or*.

As for AND, model performance looks very similar between experiment 1 and experiment 3, albeit with a little less variation across runs. 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 experiment 1, 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 determining how and when models come to learn these words.

BEHIND - IN FRONT OF are once again 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 Figure 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 as it was in experiment 1 when *behind* and *in front of* where evenly probable.

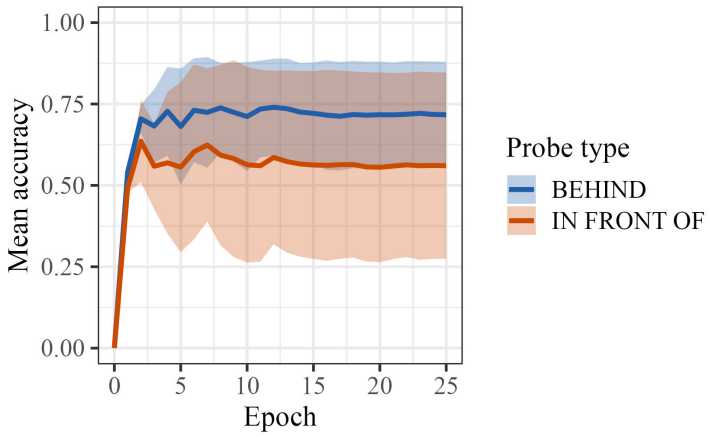


Figure 24: Experiment 3: Mean accuracy on BEHIND - IN FRONT OF probes overall when trained of subsampled CLEVR. Shading represents standard deviation across 5 models.

The most interesting results can be seen in Figure 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, showing a clear gradient representation for the meaning of *behind* as a function of distance. However, in the case of IN FRONT OF, the gradient has completely disappeared.

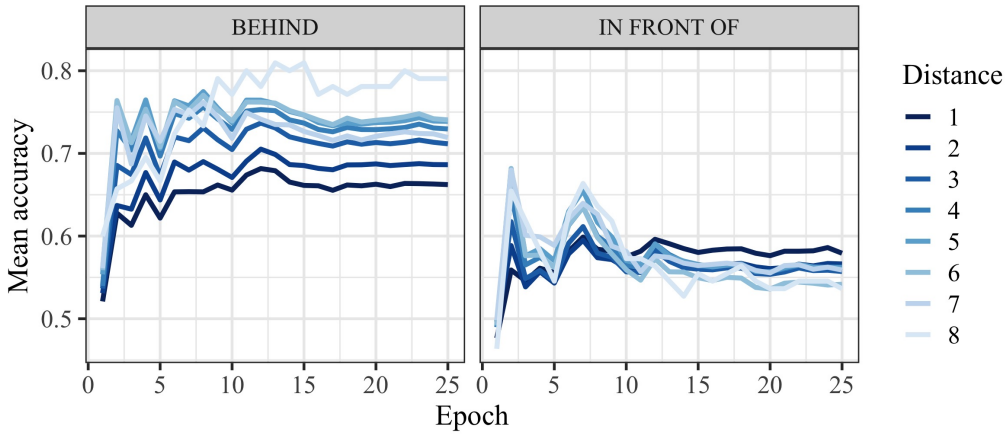


Figure 25: Experiment 3: Mean accuracy on BEHIND - IN FRONT OF probes as a function of the Euclidean distance between referents when trained of subsampled CLEVR.

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.

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

Figure 26 shows the overall performance of models on both probes. Performance on MORE is generally above chance, while on FEWER it is below chance.

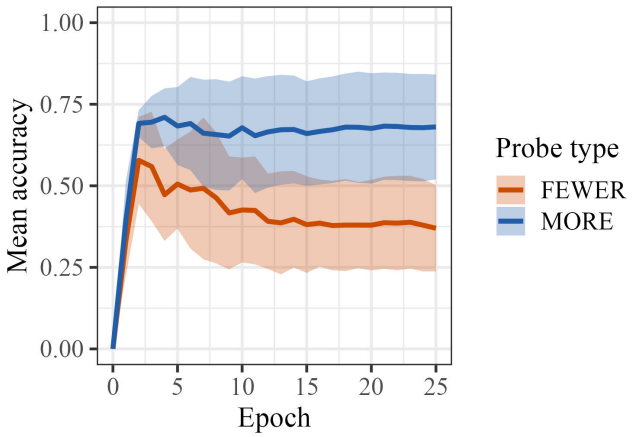


Figure 26: Experiment 3: Mean accuracy on MORE - FEWER probes overall when trained of subsampled CLEVR. Shading represents standard deviation across 5 models.

Our plots of accuracy as a function of answer type Figure 27 resembles its counterpart from experiment 1. Models do quite well in 'yes' contexts for both MORE and FEWER probe questions. The issue seems to be in contexts where a 'no' answer is required, especially for FEWER.

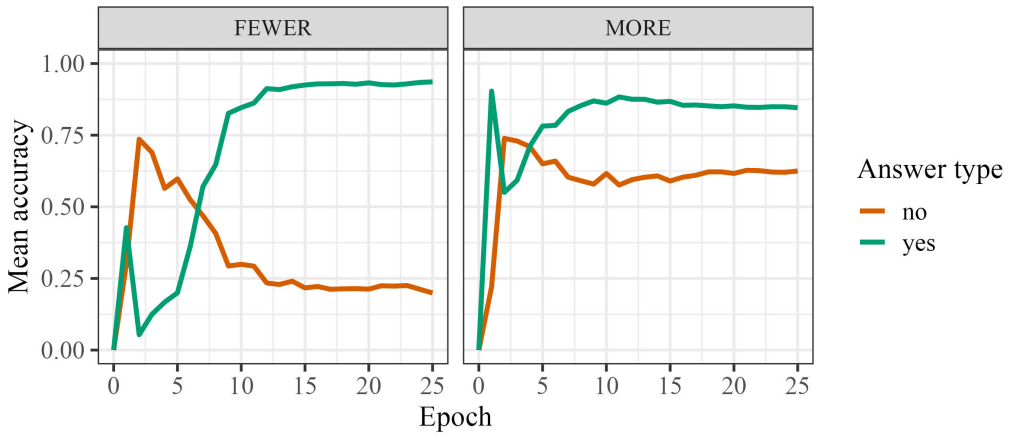


Figure 27: Experiment 3: Accuracy on MORE - FEWER probes by answer type.

Further probing with Figure 28 shows us 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 learn a gradient representation for the meaning *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 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.

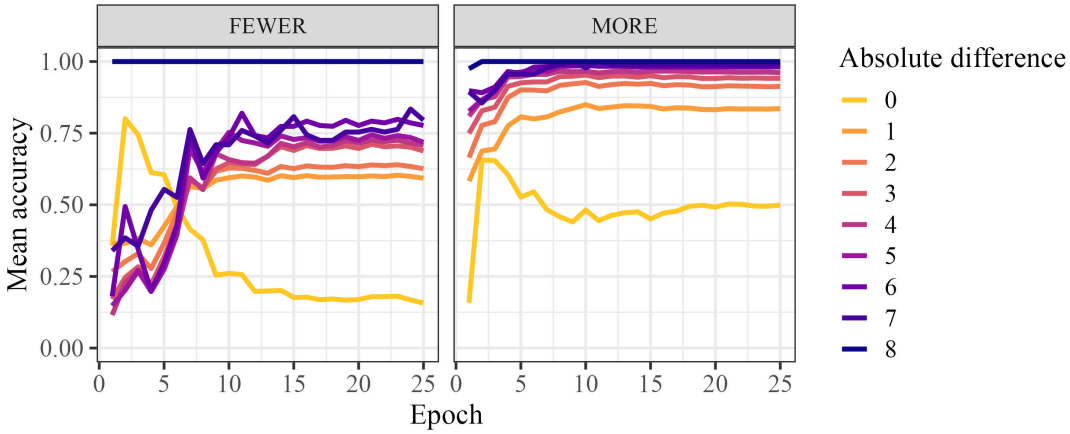


Figure 28: Experiment 3: Mean accuracy on MORE - FEWER probes by absolute difference in the number of objects in each referent class when trained of subsampled CLEVR.

By removing the probe questions where $|X| = |Y|$ and replotting the overall accuracy of the models on all other cases in Figure 29, 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 seen in experiment 1.

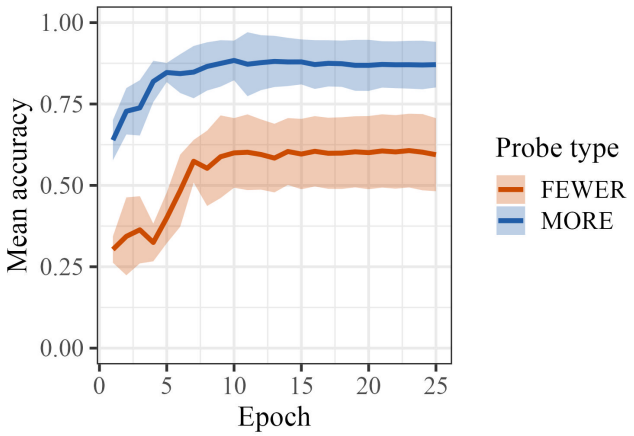


Figure 29: Experiment 3: Mean accuracy on MORE - FEWER probes overall excluding context where $|X| = |Y|$, when trained of subsampled CLEVR. Shading represents standard deviation across 5 models.

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 their are on MORE questions, but still perform above chance if we exclude contexts where $|X| = |Y|$. Once again, in these contexts, 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.

If trained on a corpus with similar frequencies to children's input, models will have difficulty learning lesser frequent function words then their more frequent counterparts. When learning logical reasoning, though frequency likely has an impact, there are clearly other factors that influence our models' ability to learn the meanings of *and* and *or*. In the next section, we discuss how our results from all three experiments can inform our understanding of function word learning in children.

9 Discussion

In this final section, we review how our experiments have helped answer our research questions. First we address what we have learnt about how models learn these words, answering our first set of research questions. Second, we discuss what we have learnt from our models about the acquisition of these function words by children, answering our second set of questions.

How models learn function words Our first set of research questions asked: What type of semantic representations do VQA models learn for function words? And do these representations generalize to unseen linguistic and visual contexts?

For *and* and *or*, we found that recurrent neural network models exposed to visually grounded language could learn the meaning of both conjunction and disjunction. Furthermore, we found evidence of models considering the availability of logical alternative questions and answers when

determining the intended meaning of a word, suggesting that they perform some form of 'pragmatic reasoning' about alternatives. This type of 'reasoning' led models to increase their exclusive interpretations of *or* over time when they were exposed to *and* and *or* at similar frequency rates in their input training data.

For *behind* and *in front of*, our results showed that models learnt gradient geometric meaning representations for *behind* and *in front of* even when their input data used threshold based representations. They had more ease interpreting the meaning of these words as a function of the Euclidean distance between referents, such that when objects were further apart, the models did better at interpreting these words – the opposite behavior than what we would expect if they payed attention to object occlusion first, as some have suggested children might (Johnston, 1984; Grigoroglou et al., 2019).

Finally, for *more* and *fewer*, we found that our models could learn both the meanings of *more (than)* and *fewer (than)*, learning *more (than)* slightly before *fewer (than)*, especially when the training context replicated the relative frequencies of these words from children's input. Additionally, we found that, like with locative prepositions, models favored learning gradient semantic representations over the threshold based definitions they were exposed to in the training data; the larger the difference in number between the two referent categories, the easier it was for models to interpret the meaning of *more* and *fewer*.

How children learn function words Our second set of research questions asked: 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?

We could not fairly compare the relative learning difficulty of *and* and *or* for models given that their contextual distributions in CLEVR were quite different, but we did find that the frequency at which these words appeared in the model interacted with other factors such as the presence or absence of other logical alternatives in determining how well models then performed on our probes. Our results show that both the meaning of *and* and inclusive *or* are learnable using general learning mechanism, without any prior knowledge of logic. Furthermore, we found that models could learn both these words when they were exposed to them at similar relative frequency to what we expect children to be exposed to. Our results support a usage-based theories of the acquisition of logical connectives (Morris, 2008) over proposals like logical nativism (Crain, 2012).

For *behind* and *in front of*, we found evidence from our corpus search using CHILDES that *behind* is much more frequent in children's input than *in front of*. When models saw these words at equal frequency, they also learnt them equally well. There is nothing intrinsically more difficult about learning the meaning of *in front of* then *behind* – at least in the case of models – as some of the theories on their acquisition suggest (Johnston, 1984; Grigoroglou et al., 2019; Windmiller, 1973; Kuczaj & Maratsos, 1975; E. V. Clark, 1977). However, when models were exposed to these words at similar frequency to children, they learnt *behind* but not *in front of*. This asymmetry in terms of exposure may be enough to explain the same asymmetry seen in children's acquisition of the meaning of these locative prepositions.

Finally, for *more* and *fewer*, previous proposals for the discrepancy in age of acquisition between *more* and *fewer* promoted the idea that the acquisition of these comparative quantifiers followed a series of developmental stages. These stages in turn led to the learning asymmetry, where *more* is acquired earlier, either because children first learn to use it in singular referent contexts

like in the additive sense of *more* (Donaldson & Wales, 1970), or because they first learn a simpler generic form of the positive comparative *more* (H. H. Clark, 2018). In our analysis, we could not isolate the different usages of *more* from one another. Thus, it may be difficult to differentiate between these previous proposals and a frequency based explanation, since the existence of other meanings for *more* contribute to it being more frequent in children's input in the first place. However, we found that if we included all instances of *more*, it was almost 100 times more frequent than all instances of *less/fewer* in our corpus of child-directed and child-adjacent language from CHILDES. Given the staggering difference in relative frequency between these words, it seems hasty to eliminate the possibility that frequency accounts for some of the acquisition difficulty asymmetry between *more* and *fewer/less*.

Using neural network models, we have shown that all of these function words are learnable from non-symbolic general learning mechanisms, as opposed to stage-based or symbolic ones. Additionally, our results support frequency over conceptual factors as the main explanation for ordering effects seen during learning. As such, our experiments offer important 'proof of concept' evidence in favor of a usage-based theories for the acquisition of logical, spatial, and numerical reasoning skills and their corresponding function words.

10 Conclusion

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 reccurrent attention-based neural network model exposed to visually grounded language as a testbed to evaluate the learnability of these function words and there respective reasoning skills.

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 as a tangential result to the supervised visual question answering task they were trained on, the CLEVR dataset. Models favored learning gradient semantics for function words requiring spacial and numerical reasoning rather than threshold-based semantics, showing that gradience in meaning may emerge from exposure to language in visually grounded contexts. Models also learnt the meanings of logical connectives *and* and *or* without any prior knowledge of logical reasoning. They showed early evidence of considering alternative possible expressions when inferring the meaning of these words, leading to a rise in exclusive interpretations for *or*.

Second, 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, more frequently seen words generally being easier to learn. 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.

Our results offer the first proof of concept evidence that it is possible to learn these reasoning skills and to map them to novel words without any prior knowledge, supporting more usage-based theories of their acquisition. Congruently, word learning difficulty was found to be mainly affected

by frequency of exposure rather than conceptual factors.

Models like the one in this paper offer a effective way to contribute to existing debates about language learning. They can work in concert with other methods to build a strong picture in favor of one theory over others – in this case, usage-based approaches to word learning. We believe them to be a promising direction for future studies of grounded language learning.

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).

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).

Bloom, P. (2002). *How Children Learn the Meanings of Words*. 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*, 1–16.

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. (1970). Derivational complexity and order of acquisition in child speech. *Cognition and the Development of Language*, 11–53.

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 *Proceedings of the 25th Boston University Conference on Language Development* (pp. 157–168).

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. *Experimental Pragmatics*, 283–300.

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

Clark, E. V. (1977). Strategies and the mapping problem in first language acquisition. *Language Learning and Thought*, 147–168.

Clark, E. V. (1993). The mapping problem. In *The Lexicon in Acquisition* (p. 43–66). 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). Wiley and 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 University Press.

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). Wiley and sons.

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

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.*

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).

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).

Hill, F., Clark, S., Blunsom, P., & Hermann, K. M. (2020). Simulating early word learning in situated connectionist agents. In *Proceedings of CogSci.*

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

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).

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations.*

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).

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

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.*

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).

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., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kraljević, J. K., Hrzica, G., … Noveck, I. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, *113*, 9244–9249.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Procceedings of the 3rd International Conference for Learning Representations.*

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.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. In (Vol. 123, pp. 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 the Association for Computational Linguistics Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 46–55).

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

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition*. Erlbaum.

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the Conference on Neural Information Processing Systems.*

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. In *Proceedings of the International Conference on Learning Representations.*

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).

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

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

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

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).

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

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).

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

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. In *Proceedings of the Conference on Neural Information Processing Systems.*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

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*.

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 University Press.

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*. Harvard university press.

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

Wang, R., Mao, J., Gershman, S. J., & Wu, J. (2021). Language-mediated, object-centric representation learning. In *Findings of the Association for Computational Linguistics.*

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.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., … Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning* (pp. 2048–2057).

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.*

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* (p. 5014-5022).

A MAC model hyperparameters

Here are the hyperparameters of the models we used for the experiments reported in the main paper. All models were trained on an Nvidia RTX 3080 GPU, taking about 15-20 hours including the probe evaluations every epoch. Like in the original paper (Hudson & Manning, 2018), models are trained using an Adam optimizer (Kingma & Ba, 2015). Word vectors were initialized randomly using a standard uniform distribution based on a random seed. The models used a variational dropout of 0.15 across the network. We also used ELU as our non-linearity type.
number of epochs: 25
learning rate: 0.0001, if performance on the validation set didn't improve from an epoch to the next, then the learning rate got reduced using a decay rate of 0.5.
batch size: 64
number of MAC cell layers: 4
hidden size: 512 throughout network
embedding size: 300
image features output size: 1024
random seed: [0:4]

B On model's learning SAME

In initial versions of these experiments, we also included probes for the relational adjective *same*. Though the word *same* is part of the CLEVR vocabulary, the word *different* is not. So instead of having a pair of opposing words for this reasoning skill, we consider how well models learn the meaning of *same* in contexts where objects do or do not share attributes. We tried two probe versions, one which tested the comparison of object sets and a second simpler one which required one-to-one object comparison. We chose not to include these results in the main paper as these probes proved to be either too hard or too easy for the models to answer, leading to overall uninformative results as to the learning dynamics of these models.

B.1 The set comparison SAME probe results

This version of the SAME probe requires models to reason about sets of objects or varied size rather than to compare two specific referents. Questions were of the form 'Are the *X*s the same *property*', X being all possible referring expressions for objects within the CLEVR universe and *property* being either *color*, *shape*, *size* or *material*. We matched possible properties to referents such that the value of these properties was not explicitly given by the referent terms (e.g. If the referents were *brown sphere* and *metal cube*, then the only possible value for property was *size*, since the referent terms already mention the color, material, or shape of one of the objects). This first template presupposes that there are at least two *X*s. After finding all images for each question where this presuppositions was satisfied and then sampling 10 images if there were more than 10 available, this probe had a total of 1,170 image-question pairs.

In order for the answer to a question to be 'yes' in the case of the first template all Xs in the image had to have the same value for the *property* listed in the question, otherwise the answer was

'no'. We used the 'scene' metadata file associated with each image to verify whether this condition was met.

We ran this version with random seeds [0-2] before deciding to change to a different SAME question template, so the results presented here are based on 3 random runs rather than 5 like the rest of the experiments in the paper. We plot the mean accuracy and standard deviation across runs.

Figure 30 shows the overall performance of the models on this version of the SAME probe. The models' average performance does not surpass the 50% mark which we can consider chance since models quickly learn that existential questions only have two possible answers, 'yes' and 'no'.

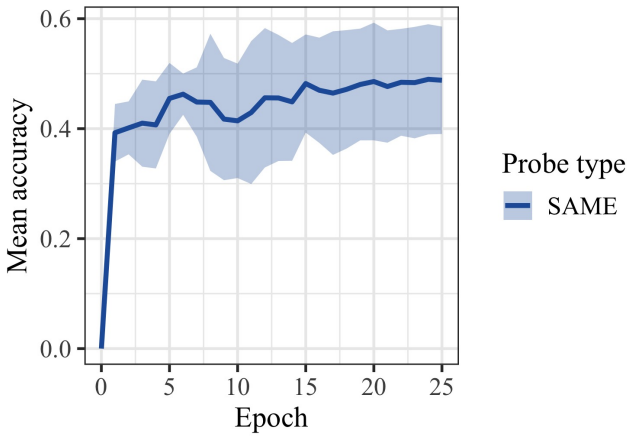


Figure 30: Mean accuracy on previous SAME probes overall, shading represents standard deviation across 3 models.

We can also look at results by answer type in Figure 31. The initial peak in accuracy for 'yes' questions is due to the fact that models answers 'yes' indiscriminately to the large majority of SAME probe questions as an early strategy. They then learn to also answer 'no', though the models still seem to really struggle in contexts where a 'no' answer is expected, while it seems to do a little better in contexts where all *X*s share the same *property* and a 'yes' answer is expected.

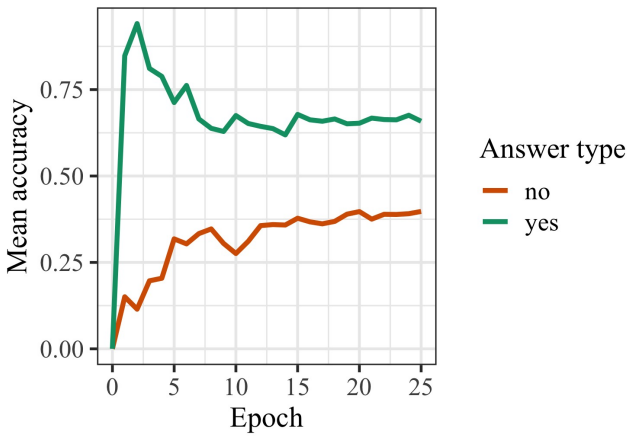


Figure 31: Mean accuracy on previous SAME probes by answer type.

This initial version of the SAME probe may have been too different of a use of the word *same* from those the model was exposed to during training, which could explain the models poor performance. Interestingly, there does seem to be a noticeable difference in accuracy between contexts where the *property* was in fact the same – 'yes' answered questions – and those where it was not – 'no' answered questions. There is no such difference with the newer version of the SAME probe.

B.2 The one-to-one comparison SAME probe results

SAME probes Given that models struggled with probe questions of the form 'Are the *X*s the **same** *property*?', we opted to try 'Are the *X* and the *Y* the **same** *property*?', where *property* is one of the following: *shape*, *color*, *material*, *size*. This second template presupposes that there are exactly one *X* and one Y. This first template presupposes that there are at least two *X*s. After finding all images for each question where this presuppositions was satisfied and then sampling 10 images if there were more than 10 available, the second SAME probe had 38,580 question-image pairs.

In order for the answer to a question to be 'yes', *X* and *Y* had to share the same value for the given *property* for a 'yes' answer, otherwise the answer was 'no'.

We ran this version with random seeds [0-4], so the results presented here are based on 5 random runs like the rest of the experiments in the paper. We plot the mean accuracy and standard deviation across runs.

Figure 32 shows the overall mean accuracy of models on this SAME probe using the newer two referent comparison template instead. The models clearly learn to answer these questions with almost perfect accuracy.

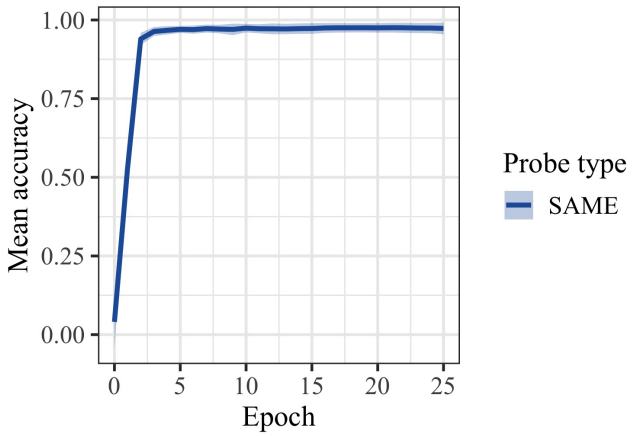


Figure 32: Experiment 1: Mean accuracy on SAME probe overall, shading represents standard deviation across 5 models.

When we considered model accuracy on the probe as a function of the answer type expected, we found that there was no visible difference in performance between contexts which expected 'yes' or 'no' answers, Figure 33.

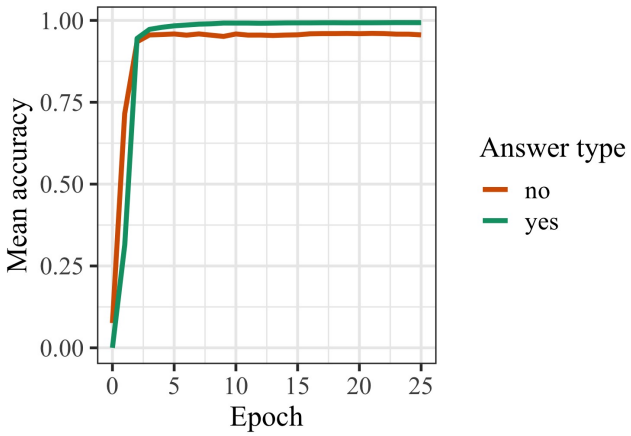


Figure 33: Experiment 1: Accuracy on SAME probe by answer type.

Thus, though the models may struggle to generalize using *same* to reason about items in a set, they learn to quickly use it to compare two unique referents. Additionally, they show no difference between correctly answering questions in contexts where these two referents share the *same property* versus contexts where they do not.