Can Generative Multimodal Models Count to Ten?

Sunayana Rane
Department of Computer Science
Princeton University
srane@princeton.edu

Alexander Ku
Google DeepMind
alexku@google.com

Jason Baldridge
Google DeepMind
jasonbaldridge@google.com

Ian Tenney
Google Research
iftenney@google.com

Thomas L. Griffiths
Departments of Psychology and Computer Science
Princeton University
tomg@princeton.edu

Been Kim
Google DeepMind
beenkim@google.com

Abstract

The creation of sophisticated AI systems that are able to process and produce images and text creates new challenges in assessing the capabilities of those systems. We adapt a behavioral paradigm from developmental psychology to characterize the counting ability of a model that generates images from text. We show that three model scales of the Parti model (350m, 3B, and 20B parameters respectively) each have *some* counting ability, with a significant jump in performance between the 350m and 3B model scales. We also demonstrate that it is possible to interfere with these models' counting ability simply by incorporating unusual descriptive adjectives for the objects being counted into the text prompt. We analyze our results in the context of the knower-level theory of child number learning. Our results show that we can gain experimental intuition for how to probe model behavior by drawing from a rich literature of behavioral experiments on humans, and, perhaps most importantly, by adapting human developmental benchmarking paradigms to AI models, we can characterize and understand their behavior with respect to our own.

Introduction

Modern AI systems are capable of performing sophisticated tasks using images and text, such as generating a picture based on a description. With such text-to-image multimodal models gaining widespread use, it is more important than ever to characterize and methodically study their behaviors. Recent research has focused on studying whether these models demonstrate compositionality, appropriately producing the right combination of abstract concepts (Thrush et al., 2022). Here we focus on an even simpler form of abstraction: understanding number. For example, we might wonder whether a model can reliably count to ten, and whether its internal understanding of number concepts matches what we would expect of a human. Many of the questions now being asked about a model's "understanding" of concepts are the same questions previously asked about human children (Frank, 2023a, 2023b). Developmental psychologists have devised tests and measures that probe many aspects of a child's understanding of number. For example, researchers discovered that children often produce (speak) number words in order ("one, two, three") before they understand how to use them or what they mean (before they can produce 3 objects, or even correctly count the 3 objects placed in front of them; Frye et al., 1989; Fuson, 2012; Sarnecka and Carey, 2008). Here we adapt these procedures to study multimodal models.

Standard evaluations of multimodal models focusing on a broad set of capabilities also often include some measure of counting ability among a larger set of metrics (Cho, Hu, et al., 2023; Cho, Zala, & Bansal, 2023; Hu et al., 2023; Lee et al., 2023). In contrast to these broader metrics designed to provide a standard measure of a wide range of model abilities, here we provide a deeper behavior-based analysis which systematically and specifically varies counting along controlled lines, supporting direct comparison with human children.

Number Concepts in Children

While number concepts may seem simple to learn, the mechanisms by which children acquire them are nuanced and still debated (Condry & Spelke, 2008; Le Corre & Carey, 2007). Although children recite an ordered list of number words early on, usually between 2-3 years old (Geary et al., 2018), it takes them surprisingly longer to ground these number words and thoroughly understand number concepts (Condry & Spelke, 2008). We can take cues from these studies of children's acquisition of number words and concepts to develop in-depth studies for models that seem to exhibit some knowledge of number words.

Several studies support the idea that there are at least two distinct systems related to number. Up to counts of about 3 to 4, children can represent objects exactly (Condry & Spelke, 2008; Feigenson & Carey, 2005) and often understand each number as a separate concept representing a set of precisely that many objects. Beyond 3 to 4 objects, their mechanism for learning larger numbers seems to involve an understanding of the idea that adding 1 to a previous number produces each subsequently larger number (Condry & Spelke, 2008; Sarnecka & Carey, 2008). We discuss this idea and the related "knower-level" framework of child number learning in the analysis of our results.

There is also evidence that children learn "one" differently and earlier than larger numbers. Wynn (1992) found that a group of 2.5-year-old children produced a single object when asked for "one" and multiple objects when asked for a number greater than one, but they were not able to accurately produce the *correct* number of the multiple objects. This indicates that they understood "one" fairly early on, and distinguished it from larger numbers, but did not fully understand the larger numbers until much later. Further studies in Wynn (1992) suggest that children learn to successfully produce "two" many months after learning "one," and that the learning of these early number words can be context-dependent.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Qualitative analyses have complemented these quantitative results to provide a richer, more detailed picture of number learning in children. In a qualitative observational study Mix (2002) noted that child understanding of number was highly context-dependent, saying that "numerical competence can appear full-blown in one context but nonexistent in another." In the study, a 26-month-old child gave one dog treat to each of his dogs, but could not (or would not) give one freeze-dried pea to each of his toy trains. Our experiments take inspiration from these studies, investigating whether models' understanding of number is also context-dependent.

Behavioral studies of number understanding in children provide ample evidence that there is nuance and complexity involved in learning number concepts – there is much more diversity of behavior than meets the eye. Contradictions and inconsistencies in child behavior are reminiscent of some of the more unpredictable behavior anecdotally observed in models. Three-year-old children have been shown to skip and double count numbers in a verbal counting list, invent their own number words, and use the same numeral twice in a count (Baroody & Price, 1983; Briars & Siegler, 1984; Frye et al., 1989; Fuson, 2012; Sarnecka & Carey, 2008; Wagner & Walters, 1982). Furthermore, Condry and Spelke (2008) showed that three-year-old children who recited numbers up to "ten" but could not produce a correct number of objects beyond "two" also failed to understand that a set of eight objects is better labeled by "eight" than "four." However, once they were told that the set was "eight" objects they understood that it was not *also* "four." They also understood that "eight" is more than "two," but not that "eight" is more than "four." These incredibly subtle nuances in children's acquisition of number concepts illustrates the challenge of characterizing their developmental trajectory for this task, and demonstrates the need for both depth and nuance in such behavioral studies. The developmental psychology methods that have illuminated how counting works in children can help us understand similar behavior that models are now beginning to exhibit.

A key feature of many developmental psychology tests for counting ability in children is the combination of number words with visual and sensorimotor modalities (pointing to pictures on cards, giving toys to a puppet) to provide evidence of understanding of the words. Reciting a list of number words usually comes much earlier than understanding them and demonstrating that understanding in the physical world (Sarnecka & Carey, 2008). Now that AI systems have generative capabilities in multiple modalities including language and vision (Radford et al., 2021; Ramesh et al., 2021), we can probe them in a way that mirrors the tests created by developmental psychologists for characterizing these very same skills in human children.

Methods

Models

One way to characterize a developmental trajectory in models is to explore several different scales of one model type. The Pathways Autoregressive Text-to-Image (Parti) model (Yu et al., 2022) provides just such an opportunity, because it presents a common model architecture at multiple scales. For our experiments we use three different scales of the Parti model: 350M, 3B, and 20B parameters.

The model architecture follows a Transformer-based (Vaswani et al., 2017) encoder-decoder framework, with the decoder receiving the major share of the increase in each increasingly large model size. The Parti models take text input such as "five lemons" and output generated images. Transformer-based models like Parti use the same self-attention-based architecture that has led to recent advances in large language models (LLMs). One feature of this architecture is that it can be easily scaled, allowing for studies like this one with a common model architecture at different scales.

Parti treats text-to-image generation as a sequence-to-sequence modeling task (Sutskever et al., 2014), a type of task initially developed for language translation. In order to do so, Parti treats the input text prompt as a sequence of text tokens, and also uses an image tokenizer to break the images in the training data into a sequence of smaller visual "patch" tokens whose vector representations are also learned during training. A contrastive loss function is used to align the image representations with the text representations, as is standard for text-image alignment in multimodal vision and language models (Radford et al., 2021; Ramesh et al., 2021).

Task

The task often used as the gold standard for measuring a child's understanding of number concepts is known as the Give-N task (Frye et al., 1989; Fuson, 2012; Marchand et al., 2022; Wynn, 1990, 1992). The idea is simple: prompted with an instruction like "give five lemons," the child must physically count out and give five toy lemons to a puppet. Instead of other tasks that ask children to count sets of objects they are given (as in the How-Many task; Schaeffer et al., 1974), the Give-N task is understood to provide a rigorous standard for number understanding; many children who can verbally count up to 5 cannot successfully produce sets of 5 items, so their performance on the Give-N task indicates that they *thoroughly* understand a number (Wynn, 1990, 1992).

Generative vision and language models can now be probed using something similar to the Give-N task, prompted with text like "five lemons" and asked to generate an image from scratch. This is in contrast to classification or captioning models, in which you can only be asked to count the number of objects in an input image, corresponding to the more lax How-Many task (Connor et al., 2024; Wynn, 1992). We gave the Parti models our text prompts based on the Give-N task. We engaged a pool of human raters through a crowdsourcing contractor to count objects in the generated images.

Standard prompts

To approximate the Give-N task in a way that makes sense with the modalities of our models, we start with child word learning data. The "object" words used to create the prompts

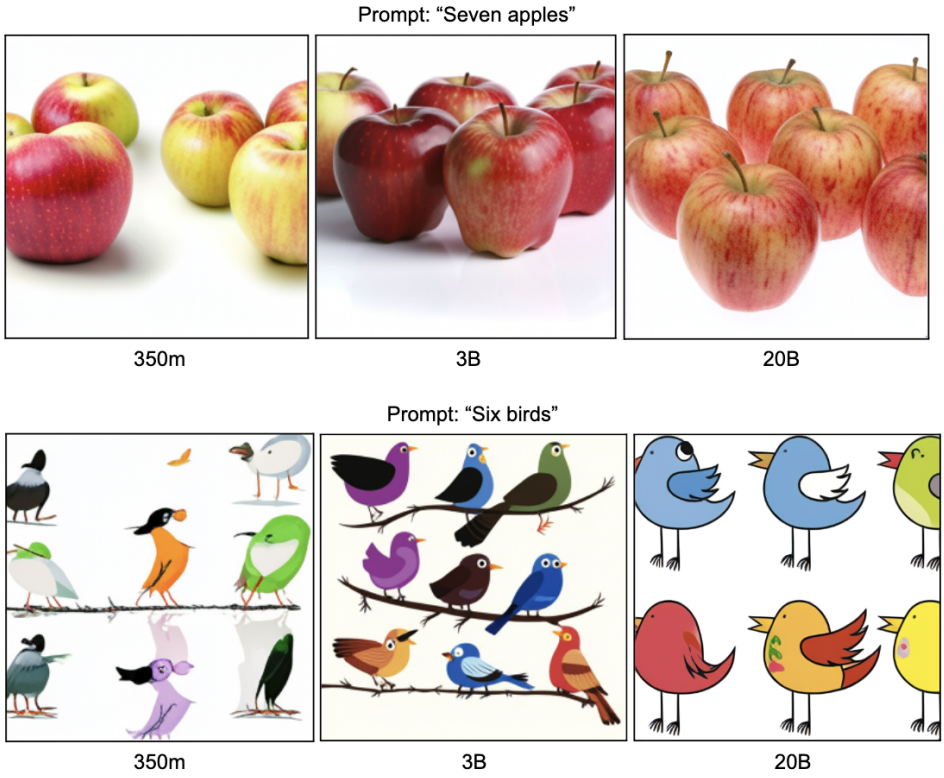

Figure 1: Images generated by each of the three scales of Parti models for the input text prompts “seven apples” and “six birds”.

of our version of the Give-N task were the 40 most easily learned food and animal words (20 of each) of all the words in the WordBank database of child vocabulary development (Frank et al., 2017). These are words most children learn prior to 36 months, and are close analogues to the food and animal toys in the original Give-N experiments. We give each model prompts of these 40 objects, with counts from 1-15.

Common and uncommon adjective prompts

In addition to the standard prompts, we also constructed a smaller set of prompts to probe how dependent counting performance is on the familiarity of the objects in the prompt – in other words, can we interfere with counting ability by modulating other things about the prompt? We refer to these as the common and uncommon prompts, fully listed in Table 1. We use a subset of the objects used in the standard prompts: apples, oranges, bananas, cats, and dogs. For each of these object types, we modulate an adjective: we use one that is common for the object, such as “fluffy” for dogs, and one that is uncommon, such as “spiky” for dogs.

Different media prompts

Although we use child development to ground our behavioral analysis, the Parti models also possess abilities we would not normally expect in children – for example, producing correct counts in different artistic styles.

When creating our standard prompts, in order to create simple prompts as faithful to the analogous child development studies as possible, we did not specify any type of media. However, in order to separately explore this ability, we designed a smaller subset of prompts specifically around different types of media. Once again, as with the common and uncommon adjective prompts, we used a subset of five objects: apples, oranges, bananas, cats, and dogs. We then used five different types of media: graffiti, painting, sculpture, DLSR high-resolution photo, and screen. The types of prompts used are listed in Table 2.

Table 1: Common and uncommon prompts

Common prompts	Uncommon prompts
“six red apples”	“six spotted apples”
“two black cats”	“two green cats”
“four yellow bananas”	“four blue bananas”
“three fluffy dogs”	“three spiky dogs”
“eight shiny oranges”	“eight hairy oranges”

Results

For each of the prompt categories (standard, common and uncommon, different media prompts) we provide a detailed breakdown of the correlation between the number of objects in the model-generated images (as recorded by human evaluators) and the true count from the text prompt given to the

Figure 2: Examples of images generated for the common and uncommon adjective prompts. These outputs were generated by the Parti 3B model.

Figure 3: Examples of images generated by the different media prompts. These outputs were generated by the Parti 20B model.

Table 2: Different media prompts

Different media prompts
“graffiti of two apples”
“a painting of five dogs”
“a sculpture of two bananas”
“a dslr high-resolution photo of three cats”
“a screen showing four apples”

model (Tables 3, 4, 5). In addition, Figure 4 gives a detailed breakdown of each model's count-by-count performance on the standard prompts.

In Table 3, all three scales of model (350m, 3B, and 20B) have *some* statistically significant counting ability, as evidenced by the model outputs' correlation with the true number in the input text prompt. However, there is a substantial increase in counting ability between the 350m and 3B scales; there is a statistically significant difference in correlation between the 350m model and both the 3B ($z = 6.37, p < .0001$ for Pearson, $z = 6.63, p < .0001$ for Spearman) and 20B ($z = 6.79, p < .0001$ for Pearson, $z = 6.42, p < .0001$ for Spearman) models. All significance tests use the Fisher r-to-z transformation for comparing independent correlations.

For the common and uncommon adjective prompts, we compare each model scale pairwise according to correlations with true number as shown in Table 4. We observe that common vs. uncommon adjectives result in a statistically significant difference in counting ability for both the 350m ($z = 3.14, p < .01$ for Pearson, $z = 4.86, p < .0001$ for Spearman) and 3B ($z = 3.53, p < .001$ for Pearson, $z = 3.41, p < .001$ for Spearman) models, but only for one correlation metric for the 20B model (Spearman, $z = 2.20, p < .05$).

Finally, for the different media prompts in Table 5, we observe the same pattern as we did for the standard prompts: there is once again a substantial increase in counting ability between the 350m and 3B scales; there is a statistically significant difference in correlation between the 350m model and both the 3B ($z = 3.91, p < .001$ for Pearson, $z = 5.04, p < .0001$ for Spearman) and 20B ($z = 4.18, p < .0001$ for Pearson, $z = 4.65, p < .0001$ for Spearman) models.

Discussion

We used the Give-N task from developmental psychology to evaluate the large multimodal model Parti at three different model scales. Our results show that all three scales of the

Table 3: Correlations between model-generated image count and true input count from text prompt

	Model	Pearson	p	Spearman	p
Standard prompts	350m	0.4071	$< .0001$	0.4735	$< .0001$
	3B	0.6647	$< .0001$	0.7159	$< .0001$
	20B	0.6781	$< .0001$	0.7099	$< .0001$

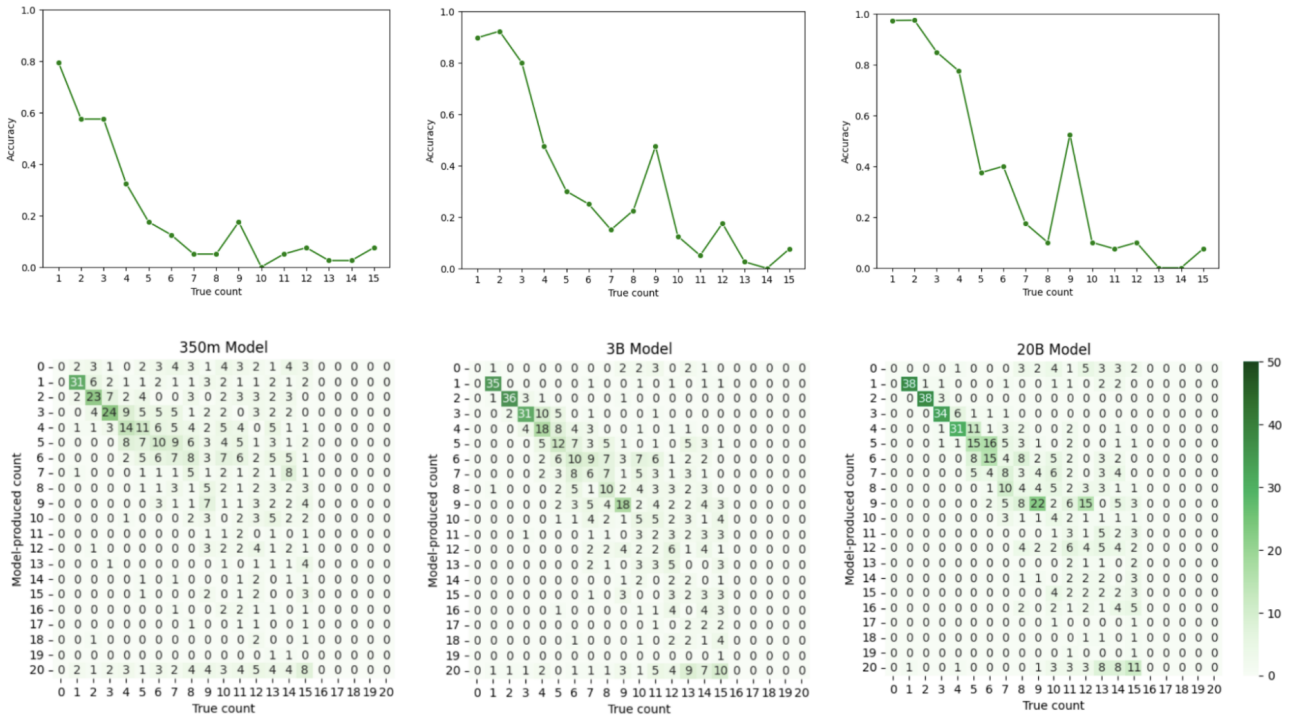

Figure 4: Standard prompt results in greater detail: counts of objects in images generated by each of the three scales of Parti model, compared to the true count as specified in the text prompt. Model accuracy plots show the counting accuracy demonstrated by each model scale for each number.

Table 4: Correlations between model-generated image count and true input count from text prompt for the common and uncommon adjective prompts

	Model	Pearson	p	Spearman	p
Common prompts	350m	0.6334	$< .0001$	0.6873	$< .0001$
	3B	0.7768	$< .0001$	0.7810	$< .0001$
	20B	0.6458	$< .0001$	0.6648	$< .0001$
Uncommon prompts	350m	0.2882	$< .0001$	0.1438	$< .0001$
	3B	0.4862	$< .0001$	0.5062	$< .0001$
	20B	0.4734	$< .0001$	0.4508	$< .0001$

Parti model have *some* counting ability, with a steep increase in performance between the 350m model and the 3B model. Both the 3B model and 20B model have similar performance in most categories, indicating that counting skills may be "unlocked" at the 3B model scale.

In analyzing the results, we also draw from the knower-levels framework often used to understand Give-N task results in child psychology studies (Sarnecka & Carey, 2008; Wynn, 1992). At the "one-knower" level, which most children reach by 2.5-3 years, they understand only the concept of 1. A few months later, a child becomes a "two-knower," when they reliably give 1 and 2, but not 3, 4, 5. Then slowly

Table 5: Correlations between model-generated image count and true input count from text prompt for different media prompts

	Model	Pearson	p	Spearman	p
Different media prompts	350m	0.3598	$< .0001$	0.2431	$< .0001$
	3B	0.5220	$< .0001$	0.4692	$< .0001$
	20B	0.5320	$< .0001$	0.4533	$< .0001$

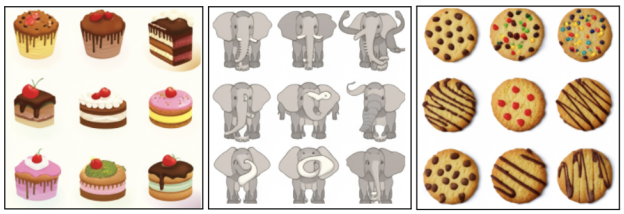

Figure 5: Many of the models' generated output images for prompts starting with "nine" produced 3×3 object grids, which might relate to the increase in counting accuracy for "nine" in Figure 4. We observed this increased tendency from both the Parti 3B and Parti 20B models.

comes the "three-knower" and, some studies report, the "four-knower" level before the child learns the subsequent numbers not slowly, as before, but all at once and through induction (they have learned that adding one to a prior number results in the next number).

Interestingly, our results suggest that this inductive step is missing from all three scales of the models' behavior. The 20B model seems to be inching along in this direction, getting fairly reliable results up to 4 and improved results on 5 and 6 compared to both the 350m and 3B models. However, this behavior obviously has not scaled past 5-6, and from 7 onward it is quite difficult for the models, in contrast to children who learn 5 onward quickly and inductively. Our approach illuminates this gap, and shows behavioral similarities between these models and children of approximately 3-4 years of age.

Furthermore, the results for the common and uncommon adjective prompts demonstrate a gap in performance across all models and particularly in the 350m and 3B models, indicating that unusual adjective and noun combinations do indeed interfere with the model's counting ability. This highlights an area for further training, investigation, and improvement in these models.

The results on the common, uncommon, and different media prompts indicate that for models, like for child learners, context does matter (Mix, 2002). In this case, the context is provided by the additional (often confusing) information given in the text prompt. In the case of the common adjectives, this seems to help generative counting abilities, and in the case of uncommon adjectives, it seems to hinder counting abilities. All three model scales seem very susceptible to effects from this added textual context.

The different media prompts demonstrate the models' skills in an area where we would not expect corresponding skills in human children. As the results in Table 5 show, all three model scales demonstrate *some* ability to produce the correct number of objects despite the demanding changes in media. Producing a *painting* of five bananas is not something we ordinarily expect children to do, and it illustrates an area where generative text-to-image models have developed considerable and unexpected skill. The gap between models' ability to successfully transform media and their very child-like inability to thoroughly and reliably count to ten (even given the simplest standard prompts) highlights one way that models are *not* human-like in their developmental trajectories. These divergences are just as important to study as the commonalities. One way to characterize these differences is through hybrid tasks like the different media prompts, which we designed based on model creators' knowledge and intuition of the models' advanced capabilties with different media. We combined this intuition with experimental paradigms from developmental psychology to produce these hybrid prompts designed specifically for models. In future work, we hope to continue combining these two powerful sources of knowledge and intuition to probe model behavior both within and outside the scope of behavior we would expect from human children and adults.

There are also interesting qualitative observations that relate to the quantitative results in Figure 4. One puzzling artifact is the uptick in accuracy for all three model scales for counting the number "nine." While the reasons for this uptick remain obscure, in investigating this tendency we made an interesting qualitative observation: given a prompt of the form "nine x," models often produced images that showed the nine objects in a 3×3 grid (Figure 5). Our qualitative observation was that this occurred more for the number "nine" than it did for any other number. This may be a peculiarity resulting from an over-representation of such image-text pairs in the training data, or some other attribute that makes such 3×3 grids easier for the model to reliably produce. Whatever the reasons, both the 3×3 grids and the accuracy increase are particularly noticeable for the Parti 3B and 20B models. Such artifacts illustrate the need for further in-depth behavioral studies of model behavior, complete with qualitative components.

We hope this approach empowers model designers to address developmental gaps in knowledge and performance, and that the practice of using developmental psychology paradigms to probe model behavior continues to help us develop more reliable, responsible AI systems.

References

Baroody, A. J., & Price, J. (1983). Brief report: The development of the number-word sequence in the counting of three-year-olds. *Journal for Research in Mathematics Education*, *14*(5), 361–368.

Briars, D., & Siegler, R. S. (1984). A featural analysis of preschoolers' counting knowledge. *Developmental Psychology*, *20*(4), 607.

Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., & Wang, S. (2023). Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*.

Cho, J., Zala, A., & Bansal, M. (2023). Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*.

Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General*, *137*(1), 22.

Connor, D., Kirkland, P. K., & Purpura, D. J. (2024). The how many and give-n tasks: Conceptually distinct measures of the cardinality principle. *Early Childhood Research Quarterly*, *66*, 61–74.

Feigenson, L., & Carey, S. (2005). On the limits of infants' quantification of small object arrays. *Cognition*, *97*(3), 295–313.

Frank, M. C. (2023a). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*(8), 451–452.

Frank, M. C. (2023b). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Frye, D., Braisby, N., Lowe, J., Maroudas, C., & Nicholls, J. (1989). Young children's understanding of counting and cardinality. *Child Development*, 1158–1171.

Fuson, K. C. (2012). *Children's counting and concepts of number*. Springer Science & Business Media.

Geary, D. C., et al. (2018). Growth of symbolic number knowledge accelerates after children understand cardinality. *Cognition*, *177*, 69–78.

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., & Smith, N. A. (2023). TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395–438.

Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H. B., Bellagente, M., et al. (2023). Holistic evaluation of text-to-image models. *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Marchand, E., Lovelett, J. T., Kendro, K., & Barner, D. (2022). Assessing the knower-level framework: How reliable is the give-a-number task? *Cognition*, *222*, 104998.

Mix, K. S. (2002). The construction of number concepts. *Cognitive Development*, *17*(3-4), 1345–1363.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *International Conference on Machine Learning*, 8821–8831.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662–674.

Schaeffer, B., Eggleston, V. H., & Scott, J. L. (1974). Number development in young children. *Cognitive Psychology*, *6*(3), 357–379.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wagner, S., & Walters, J. (1982). A longitudinal analysis of early number concepts: From numbers to number. *Action and thought: From sensorimotor schemes to symbolic operations*, 137–161.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, *24*(2), 220–251.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. (2022). Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, *2*(3), 5.