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

We adapt a developmental psychology paradigm to characterize the counting ability of the foundation model Parti. 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.

1 Background and Motivation

With 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 criterion for 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 we are now asking of a model's "understanding" of concepts are the same questions we've asked previously about human children (Frank, 2023a;b). Developmental cognitive 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) Sarnecka & Carey (2008); Frye et al. (1989); Fuson (2012). 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 et al., 2023a;b; 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.

2 Task and Data

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; Schaeffer et al., 1974; Wynn, 1990; 1992; Marchand et al., 2022). 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), the Give-N task is understood to provide a very 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 used in developmental psychology (Wynn, 1992; Connor et al., 2024). We gave the Parti model text prompts based on the Give-N task and engaged a pool of human raters through a crowdsourcing contractor to count objects in the generated images.

2.1 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 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.

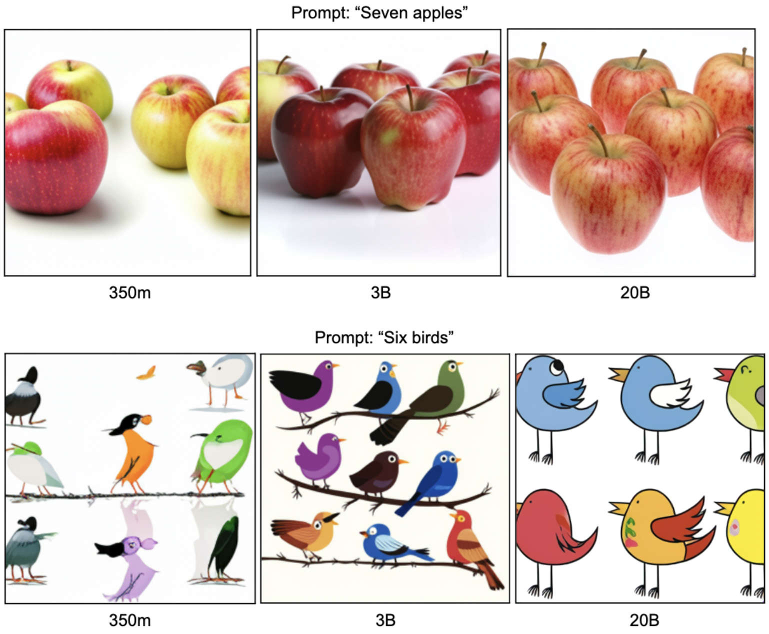
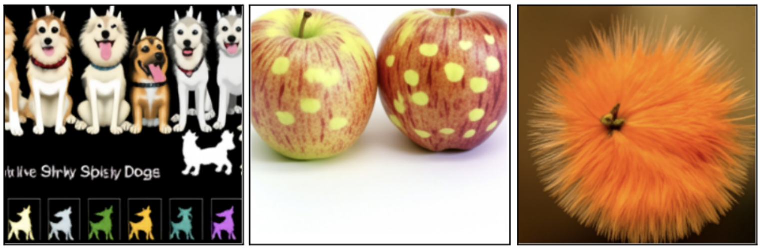


Figure 1: Images generated by each of the three scales of Parti models for the input text prompts "seven apples" and "six birds".

2.2 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.



“twelve spiky dogs” “two spotted apples” “one hairy orange”

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

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”

3 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.

4 Results

For each of the prompt categories (standard, common, and uncommon 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 the model was given (Table 2). In addition, figure 3 gives a detailed breakdown of each model’s count-by-count performance on the standard prompts.

Table 2: Correlations between model-generated image count and true input count from text prompt

	Model	Pearson	p-value	Spearman	p-value
Standard prompts	350m	0.4071	2.5601e-25	0.4735	8.4204e-35
	3B	0.6647	1.6888e-77	0.7159	4.6031e-95
	20B	0.6781	6.9107e-82	0.7099	5.7467e-93
Common prompts	350m	0.6334	1.5229e-12	0.6873	2.8497e-15
	3B	0.7768	2.1413e-21	0.7810	9.5063e-22
	20B	0.6458	4.0101e-13	0.6648	4.6304e-14
Uncommon prompts	350m	0.2882	3.6466e-03	0.1438	1.5356e-01
	3B	0.4862	2.9217e-07	0.5062	7.7494e-08
	20B	0.4734	6.5363e-07	0.4508	2.5144e-06

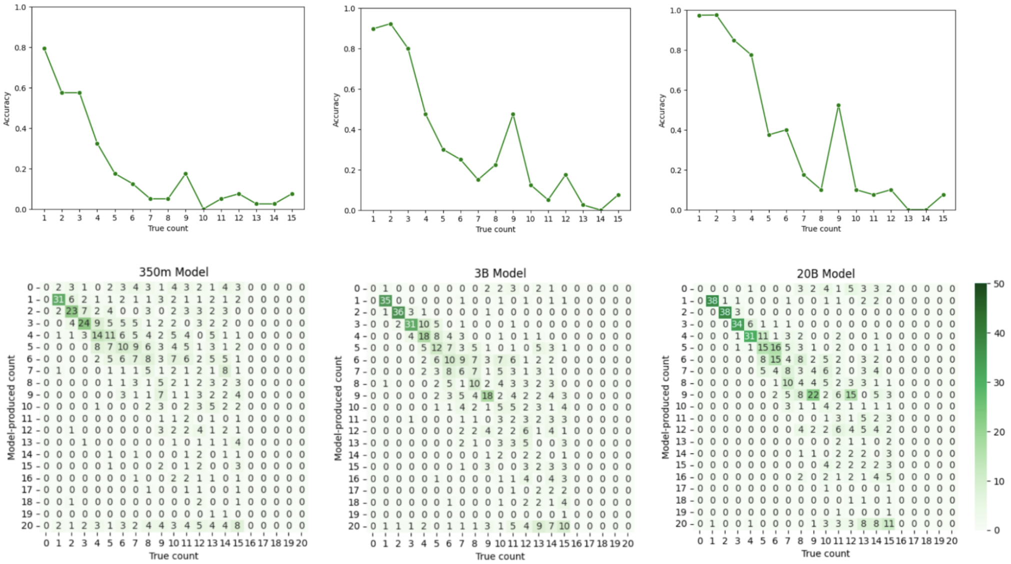


Figure 3: Standard prompt results in greater detail: counts of objects in images generated by each of the three scales of Parti model, compared to the true count as specified in the text prompt.

5 Discussion

We show results of the Give-N task from developmental psychology, as adapted to evaluate the large multimodal model Parti at three different model scales. Our results show that all three scales of the 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 (Wynn, 1992; Sarnecka & Carey, 2008). 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 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 significant gap in performance across all 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. There are also areas of strength: notably, the 3B model demonstrates particularly strong performance for common adjective prompts, surpassing even the 20B model and painting a more nuanced picture of how model scale relates to performance.

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 foundation models.

References

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023a.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023b.

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

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

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

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694, 2017.

Douglas Frye, Nicholas Braisby, John Lowe, Celine Maroudas, and Jon Nicholls. Young children's understanding of counting and cardinality. *Child Development*, pp. 1158–1171, 1989.

Karen C Fuson. *Children's counting and concepts of number*. Springer, 2012.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Elisabeth Marchand, Jarrett T Lovelett, Kelly Kendro, and David Barner. Assessing the knower-level framework: How reliable is the give-a-number task? *Cognition*, 222:104998, 2022.

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

Benson Schaeffer, Valeria H Eggleston, and Judy L Scott. Number development in young children. *Cognitive Psychology*, 6(3):357–379, 1974.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.