Does Language Stabilize Quantity Representations in Vision Transformers?

Pamela D. Rivière (pdrivier@ucsd.edu)
Department of Cognitive Science
University of California, San Diego

Oisín Parkinson-Coombs (oparkinson@ethz.ch)
History & Philosophy of Mathematical Sciences
ETH Zürich

Cameron R. Jones (cameron@ucsd.edu)
Department of Cognitive Science
University of California, San Diego

Sean Trott (sttrott@ucsd.edu)
Department of Cognitive Science
University of California, San Diego

Abstract

Whether language is essential, sufficient, or a tool for numerical cognition has been hotly debated. Here, we investigate the influence of language on quantity representations by comparing embeddings from vision-only Transformer models (ViTs) and vision-language models (VLMs) exposed to image pairs depicting either the same or different stimulus quantities. If linguistic exposure stabilises quantity representations, VLMs should produce more distinct representations for image pairs with differing numerosity and more similar representations for those with identical numerosity than ViTs. We operationalized this as the variance in Cosine Similarity in response to either categorical (same/different) or continuous differences in stimulus numerosity. We find that VLMs and ViTs are sensitive to the numerosity of visual stimuli, that this sensitivity increases with layer depth, and that VLMs exhibit slightly more sensitivity to image numerosity than ViTs. This work provides initial support for the claim that linguistic exposure can, in principle, stabilise quantity representations.

Keywords: vision language models, vision transformers, numerosity, quantity representations

Introduction

The role of natural language in structuring conceptual representations has long proved controversial. This question is instantiated in debates regarding the necessity or sufficiency of natural language for the acquisition of (exact) number concepts (Carruthers, 2002; Lupyan, 2012; Clarke, 2025). While some theories consider language—in the form of a structured lexical count list—essential to the construction of a number concept (Carey, 2009; Carey and Barner, 2019), other proposals argue for a selective role in representing larger cardinal concepts but not smaller quantities (Margolis & Lawrence, 2008), and still others maintain that language is not required for the development of a wide range of numerical and arithmetical capacities (Dehaene, 1997). Linguistic communities with limited numeral lexica offer conflicting evidence for each of these views, with some work reinforcing the necessity of language for a number concept (Gordon, 2004; Everett & Madora, 2012); arguing for the cognitive-tool position (Frank et al 2008); or suggesting that a number concept is independent from language (Pica et al, 2004). Visually discriminating quantities of objects in the environment is adaptive, enabling, for instance, both linguistic and non-linguistic organisms to accurately model and compare the value of different food caches (Carruthers, 2002). It is unclear, however, whether associating quantities with unique symbols (e.g. a number), confers any additional adaptive *stability* to our mental representations of collections of items. In principle, this question could be informed by comparing the representations of visual quantity in systems with or without exposure to linguistic labels.

Artificial neural networks (ANNs) offer a way to operationalize the notion that linguistic exposure shapes visual representations. Computer vision ANNs can be trained to classify, reconstruct, and transform images. More recently, architectures trained on vision *and* language corpora have become widely available, setting the stage for research programs evaluating the role of a variety of stimulus types in organizing learned representations. While ANNs differ from human neural architectures, their ability to process and represent various input modalities makes them uniquely suited as "model organisms" (McCloskey, 1991; Trott, 2024) with which to explore the in-principle effects of linguistic exposure on visual representations.

To this end, researchers have deployed a wide array of ANN architectures to investigate the computational basis of quantity representations and their developmental trajectory. Sensitivity to the numerosity of visual images has been reported in structured connectionist networks equipped with predefined cognitively-inspired modules (Dehaene & Changeux, 1993), shallow unsupervised networks (Verguts & Fias, 2004), as well as hierarchical generative models (Stoianov & Zorzi, 2012; Testolin et al., 2020b) and the increasingly powerful breed of Transformer-based model architectures (Boccato, Testolin, & Zorzi, 2021; Testolin, Hou, & Zorzi 2024).

With the advent of Transformer models trained on multiple input modalities, it has become possible to explore "number sense" in ANNs exposed to both language and visual inputs. Recently, Testolin et al. (2024) evaluated the ability of four distinct vision-language models (VLMs) to

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

(1) detect the numerosity of image stimuli and (2) *generate* images of a target numerosity via direct (linguistic) prompting. Results were mixed, with relatively smaller models deviating substantially from human behavior: they produced errors for small numerosities, and exhibited a non-Weberian pattern of response variability which was greatly affected by the object types in the visual stimuli. In contrast, the larger, state-of-the-art proprietary models more closely approximated human judgments. The unique effect of linguistic exposure on VLMs' quantity representations is difficult to isolate in this design, however, in the absence of a comparison against vision-only transformers.

Here, we leverage openly available vision-only Transformer models (ViTs) and vision-language models (VLMs) to examine the influence of natural (text, English) language on image representations of numerical quantities. We evaluated models' ability to produce discriminable representations between images containing objects of either the same or different numerosity. We anticipated that linguistic training in the VLMs would result in more similar representations of images with identical numerosity—and more dissimilar representations of images with differing numerosity—than those produced by the ViTs, under the assumption that language crystallizes the number concept, and this in turn organizes the otherwise continuous experience of visual inputs[1].

Specifically, we explored the following research questions (RQs):

RQ1: *Are ViTs and VLMs sensitive to the numerosity of items in images?*

RQ2: *How does sensitivity to image numerosity evolve through the models' architecture (e.g. which layer exhibits highest sensitivity)?*

RQ3: *Are ViTs and VLMs differentially sensitive to image numerosity?*

Materials and Methods

Experimental Stimuli

We sampled images from an existing stimulus set consisting of either white rectangles or white dots against a black background (369x369 pixel size), made publicly available by Testolin et al. (2020a, 2020b)[2] (*Figure 1*). Image stimuli featuring dots were originally developed to examine the effect of non-numerical properties of visual stimuli (e.g. cumulative surface area) on numerosity discriminations in ANNs. As such, the discrete quantities (numerosity), the spacing, and the size of the dots were parametrically manipulated to produce a total of 21970 unique dot images: 1690 per target numerosity for 13 numerosity levels in the set {7, 8, 9, 10, 11, 12, 14, 16, 18, 20, 22, 25, 28} (Testolin et al., 2020a; DeWind et al., 2015). Images containing rectangle stimuli were generated for all numerosities in the interval [1, 32], with cumulative area randomly sampled from a uniform discrete distribution with eight possible area values as described in Stoianov and Zorzi (2012).

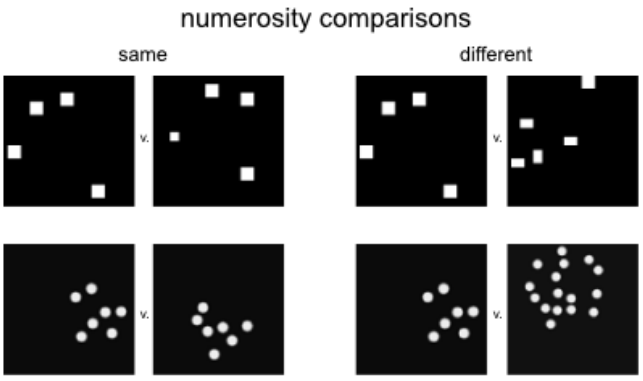

Figure 1. Example stimuli and task schematic. Each image, depicting either rectangles (*top row*) or dots (*bottom row*), is presented individually to each model. We then compute the cosine similarity between model embeddings of image pairs displaying either the same (*left*) or different (*right*) numerosities. Image stimuli from Testolin et al. (2020a, 2020b).

We randomly sampled ten pairs of images per numerosity from a total of 51200 rectangle images (1600 available images per numerosity level), and a total of 21970 dot images (1690 per numerosity level). For each numerosity level, we sampled an additional ten pairs of images which *differed* in numerosity. The resulting image pairs depicted rectangles (or dots) of the *same* or *different* numerosity, and could vary in the spatial distribution and cumulative area spanned by white rectangles (or dots). Our resulting stimulus sets totaled 640 randomly sampled rectangle images, and 260 dot images.

To account for any systematic covariance between numerosity and cumulative surface area in dot image stimuli, we included Area Difference between images in each pair as a factor in all statistical analyses.

[1] We use the words *stabilizing* and *crystallizing* to refer to the *systematic appreciation* or *tracking* of a given property, such as quantity, that describes a collection of items.

[2] We accessed stimuli at the following Open Science Framework repositories: https://osf.io/j7dvc/, https://osf.io/h5pfm/

Vision Transformer (ViT) Models

All ANNs were accessed in Python using the HuggingFace Transformers package (version 4.45.1, Wolf et al., 2020)[3]. For vision-only ANNs, we selected a suite of models trained to classify images using the Transformer architecture, originally developed for language models (Vaswani et al., 2017) and adapted to process visual inputs by Dosovitskiy et al. (2020). Transformers form the backbone of the powerful foundation models that have proved revolutionary in our understanding of the performance standards that self-supervised ANNs can achieve. Foundation models remain proprietary and closed-source, however, precluding study of their representational space. Here we leverage smaller, openly-accessible Transformer architectures, which offer a window into the kinds of representations that Transformers can learn.

To process images, ViTs segment images into fixed-size patches. Notably, ViT architectures differ in the degree of image resolution they are trained with, reflected in their patch size hyperparameter (*Table 1*). Generally, models with a larger number of parameters tend to be trained with smaller patch sizes (e.g. enhanced image resolution). To ensure that sensitivity to numerosity is not solely a function of the resolution with which a model was trained, or of the model size, we opt for exploring a variety of publicly accessible ViTs.

Vision-Language Models (VLMs)

A number of different VLM architectures have been developed, but we focus here on *dual-encoder* models such as CLIP (Radford et al., 2021), which we investigated through the open-source OpenCLIP implementation (Cherti et al., 2022). CLIP models are trained using *contrastive learning,* which adjusts both ViT and text encoder parameters to maximize the similarity between embeddings for target image-text pairs. Because the ViT's parameters are altered with the same error signal as the text encoder, we can ask whether a vision encoder's participation in a multimodal architecture and training procedure systematically influences the vision encoder's representational space.

We selected ViT-equipped, CLIP-based VLMs to maximize the architectural parity of their vision encoder with available ViT-only models. We restricted our analysis to pretrained models listed in Table 1, and we explore the limitations inherent to their use in the Discussion.

[3] All code is available at the following GitHub repository: https://github.com/pdrivier/vlm-vit-num/

Table 1: Summary of ViT and VLM specifications.

Model Type	Model Name	Pretraining Dataset Size	Patch Size	# Params
ViT	vit-base-patch16	~14 M images[4]	16x16	~86 M
	vit-large-patch16	~14 M	16x16	~304M
	vit-large-patch32	~14 M	32x32	~306 M
	vit-huge-patch14	~14 M	14x14	~632 M
VLM	clip-base-patch32	~2 B[5] image-text pairs	32x32	~151 M
	clip-large-patch14	~2 B	14x14	~427 M
	clip-huge-patch14	~2 B	14x14	~986 M
	clip-giant-patch14	~2 B	14x14	~1.3 B
	clip-bg-patch14	~2 B	14x14	~2.5 B

Procedure

To extract model representations of images in the stimulus set, we present images in a given pair individually and obtain each model layer's [CLS] token embedding, considered representative of the entire image—rather than individual patches—at a given model layer (Dosovitsky et al., 2020). We then compute the cosine similarity between embeddings for each of the images in the pair, at each model layer, resulting in as many cosine similarities (per image pair) as there were model layers.

[4] Drawn from ImageNet-21k, containing 14 million images and 21843 classes, at 224x224 resolution (Ridnik et al., 2021).

[5] Drawn from the English-text subset of the LAION-5B dataset (Schuhmann et al., 2022).

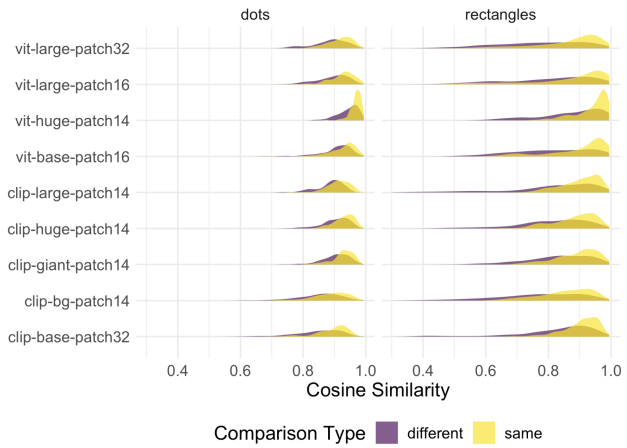

Figure 2. Distribution of cosine similarities between model embeddings of image pairs depicting either the same (*yellow*) or different (*purple*) quantities of dots (*left*) or rectangles (*right*). Cosine similarities were calculated using representations from the final layer of each model.

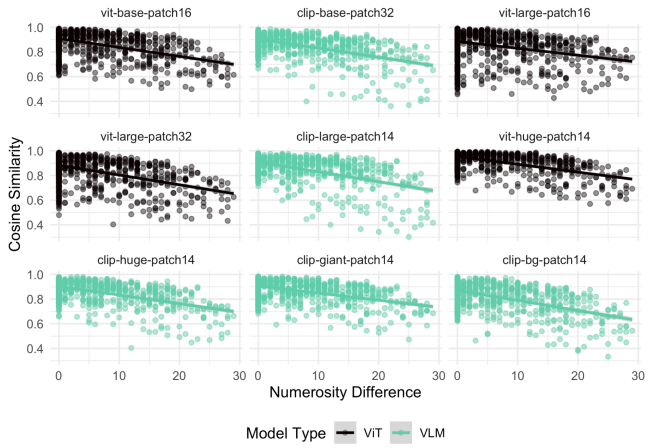

Figure 3. Relationship between Numerosity Difference (difference in quantity across members of each image pair) and Cosine Similarity, faceted by model type: ViT vs. VLM. In all cases, the relationship was significantly negative: representations were more distant for more different quantities. Here, cosine similarities were calculated using representations from the final layer of each model.

Results

All statistical analyses were performed in R using the *lme4* package (Bates et al., 2015). Random effects structure was identical across models unless otherwise specified, including random intercepts for model, the images being compared, and image type (i.e., dots vs. rectangles). Finally, for inferential tests we report the value and significance of model coefficients; key theoretical inferences were robust to whether the original measures or standardized values were used.

Sensitivity to Quantity

We began by asking whether the models tested produced representations that were *sensitive* to differences in quantity (RQ1). Sensitivity was operationalized as systematic variance in Cosine Similarity in response to either categorical (same/different) or continuous differences in the quantity depicted across the members of each image pair.

First, using all observations from all models (i.e., across all layers), we constructed a linear mixed effects model predicting Cosine Similarity, with fixed effects of Numerosity Comparison Type (Same vs. Different) and Area Difference (as a control measure). Cosine Similarity was higher for image pairs with the same quantity, as indicated by a positive coefficient: [B = .016, SE = .001, $p < .001$]. Note that there was also a significant, negative relationship with Area Difference—but importantly, the effect of Numerosity Comparison Type was preserved even when controlling for this factor. We also replicated this finding using Numerosity Difference, the absolute value of the difference in quantity between images. Here, Numerosity Difference displayed a negative relationship with Cosine Similarity [B = -.002, SE = .00006, $p < .001$], i.e., image pairs with larger differences in their quantity elicited representations that were more distant in vector-space.

Finally, we replicated both sets of results considering the final layer of each model only (see also *Figure 2* and *Figure 3*); findings are also robust to whether the original Cosine Similarity was used, or a standardized (*z*-scored) measure.

Tracking the Evolution of Quantity Sensitivity

We then asked whether models showed stereotyped patterns of quantity sensitivity. That is, were representations from certain layers more or less sensitive than representations from other layers (RQ2)? To test this, we built on the mixed models constructed in the section above, and asked whether the relevant categorical (Numerosity Comparison Type) or continuous (Numerosity Difference) factors exhibited a significant interaction with model Layer; we also controlled for Area Difference and its interaction with Layer. We found increased sensitivity to quantity at later layers for both Numerosity Comparison Type [B = .002, SE = .000015, $p < .001$] and Numerosity Difference [B = -.0002, SE = .000001, $p < .001$]. See *Figure 4* for a depiction of the average difference in Cosine Similarity across

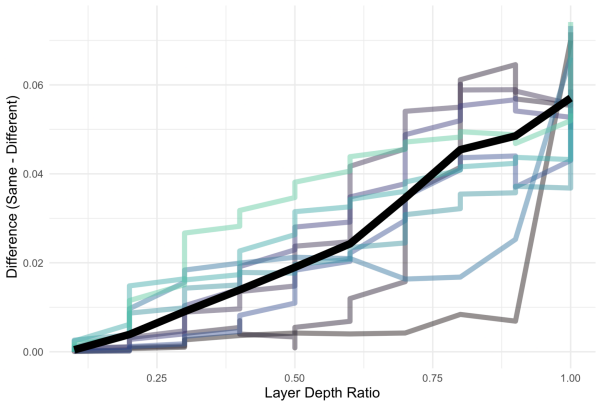

Figure 4. Average difference in Cosine Similarity for same-quantity vs. different-quantity image pairs, plotted as a function of layer depth (layer divided by total layers) for each model. Solid black line depicts the grand average. In general, later layers of all models produced representations that were more sensitive to differences in quantity.

same-quantity vs. different-quantity pairs as a function of layer depth.

Comparing Model Types

Our crucial question was whether models exposed to language input and equipped with a text encoder (VLMs) displayed greater sensitivity than models exposed only to images (ViT) (RQ3). To examine this question, we construct a linear mixed effects model to evaluate the interaction between categorical variables Model Type (with levels: VLM and ViT) and Numerosity Comparison Type—also controlling for Area Difference and its interaction with Model Type, as well as model Patch Size, and the log Number of Parameters, and each of their interactions with Numerosity Comparison Type. We found that VLM representations exhibited higher Cosine Similarities for same-numerosity images than their ViT counterparts [B = .006 , SE = .003, p = .03]. In a separate analysis, the interaction between Model Type and the continuous Numerosity Difference measure was also significant (and negative, as predicted) [B = -.0008, SE = .0002, p < .001]; that is, image pairs with larger differences in their numerosity produced more distant representations in VLMs on average than in ViT models. In a multiverse analysis (Steegen et al., 2016), we find that this latter interaction remains significant regardless of which covariates were included or excluded; the former was also robust to other model specifications. Note that this difference in sensitivity,

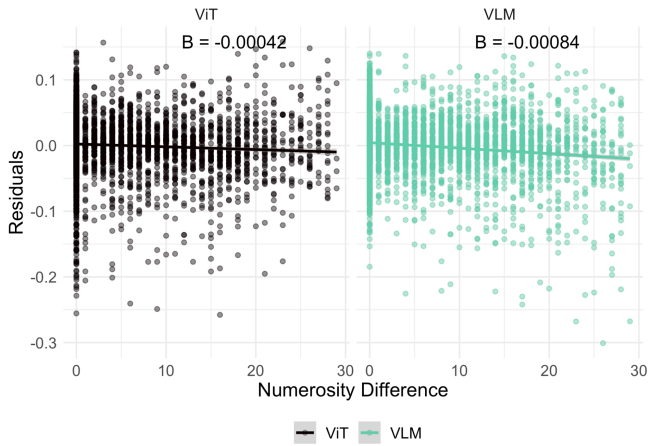

Figure 5. Relationship between Numerosity Difference and the residuals of a statistical model controlling for covariates such as Patch Size and Area Difference. Once these covariates are statistically controlled for, the residuals exhibit a slightly (and significantly) more negative relationship with Numerosity Difference for VLMs than ViT. The remaining variance is slightly more correlated with numerosity differences for VLMs than for ViTs.

while significant, was not particularly large. To visualize this effect, we plotted the residuals of a statistical model predicting Cosine Similarity as a function of Area Difference and Patch Size against Numerosity Difference for both VLMs and ViTs (*Figure 5*). The slope of the line is slightly (and significantly) more negative for VLMs, suggesting that leftover variance in Cosine Similarity is more correlated with numerosity differences for VLMs than ViTs—though again, this effect is quite small.

Crucially, in the same analysis, we do not find a significant interaction between Area Difference and Model Type, which we would expect to observe if differences in training data size (*Table 1*) accounted for VLMs' enhanced sensitivity to numerosity. To ensure that the absence of an Area Difference effect held across different image types, we carried out the same analysis, but subsetting by image type. While dot images did not yield a significant interaction between Area Difference and Model Type, rectangle images did, with ViTs exhibiting *increased* sensitivity to Area Difference relative to VLMs [B = 4e-04, SE = 2e-5, p < 0.001], where VLMs evince a flatter slope describing the relationship between Cosine Similarity and Area Difference: the opposite of the interaction observed for Numerosity Difference. The effects of Numerosity Difference and Area Difference are dissociable, indicating that differences in

training dataset size (*Table 1*) are unlikely to be exclusively accounting for VLM sensitivity to numerosity.

Discussion

Our primary research question was whether exposure to language input augments representations of quantity. For example, does learning to associate visual depictions of quantity with linguistic *labels* (e.g., "two dots") make quantity a more salient feature—and thus draw a sharper contrast with depictions of other quantities (e.g., "three dots")? We compared sensitivity to depicted quantities in artificial neural networks that were either trained on *both* text and image data (vision-language models, or VLMs) vs. those trained only on image data (vision transformers, or ViTs). We found: first, that both types of models were sensitive to quantity differences (*Figures 2 & 3*); second, that sensitivity was higher in later layers of both types of models (*Figure 4*); and third, that sensitivity was slightly (but significantly) higher for VLMs than ViTs (*Figure 5*). While this approach has clear limitations (see below), the results serve as a proof-of-concept that comparisons between ViTs and VLMs may prove valuable in isolating the effect of language on perceptual experience.

One crucial limitation of this approach is the lack of experimental control over the models under consideration. We used pre-trained ViTs and VLMs (*Table 1*), which varied in their patch size, number of parameters, and training data. We controlled for some of these features (e.g., patch size and number of parameters) in the selection of stimuli and our statistical analyses, but we could not control for the fact that VLMs were simply trained on a larger *volume* of data, which could in principle account for enhanced sensitivity to numerosity. That said, in a follow-up analysis, we did *not* find enhanced sensitivity in VLMs to non-lexicalized features such as surface area, which one would expect to find if the difference was purely due to training data volume. Future work could address outstanding concerns about the lack of experimental control in two ways: first, through controlled pre-training that selectively includes or excludes examples of labeled quantities; and second, through a larger range of stimulus probes that vary in realism and depicted content. Recent work training state-of-the-art models like Molmo (Deitke et al., 2024) made use of custom counting datasets (e.g., PixMo-Count) that could be useful to explore here.

Second, our operationalization of sensitivity to quantity—the cosine distance between the embeddings for image pairs—is imperfect. The geometry of transformer embedding spaces may introduce additional confounds, such as anisotropy, which poses challenges when interpreting cosine similarity metrics (Ethayarajh, 2019; Godey et al., 2024). Future work could explore the use of other metrics or approaches, e.g., image-generation tasks, training a decoder on the embeddings, among others (Boccato, Testolin, & Zorzi, 2021; Testolin, Hou, & Zorzi, 2024).

Relatedly, future work on quantity representations in Transformer models could adapt interpretability methods (Wang et al., 2022) to identify circuits that may encode information like quantity—as well as potential confounds like surface area, item spacing, or size (DeWind et al., 2015). In principle, the emergence of these circuits, and the various magnitudes they track, could also be identified throughout pre-training (Saphra & Lopez, 2019). Additionally, state-of-the-art VLMs such as Molmo (Deitke et al., 2024) could be leveraged to investigate the effect of alternative training objectives and model behaviors, such as "pointing" and chain-of-thought "reasoning"—both of which could plausibly affect a system's ability to enumerate objects in a scene.

The question of whether and how language input influences conceptual representations is a longstanding one in Cognitive Science (Lupyan, 2012; Lupyan, 2016; Lupyan et al., 2020; Dubova & Goldstone, 2023). Regarding concepts of quantity and numerosity specifically, there is considerable debate about the extent to which mental representations (and the capacity to form those representations) are shaped by symbolic reference (Núñez, 2017). Yet isolating the contribution of linguistic input specifically is methodologically challenging at best. In the current work, we leveraged advances in engineering to inform the question of whether exposure to language shapes conceptual representations. We found some evidence to support this hypothesis, though as noted previously, the lack of experimental control over the models and training data rules out strong conclusions at this time. Moving forward, the study of language and cognition may benefit from further methodological refinement over the use of systems like VLMs, LLMs, and related tools (Luo et al., 2023).

Acknowledgements

We'd like to thank all anonymous reviewers, as well as Alberto Testolin, who took the time to offer helpful comments on the manuscript. This work was funded in part by the UCSD Chancellor's Postdoctoral Fellowship Program, the Diverse Intelligences Summer Institute, an Open Philanthropy Fellowship, and the QUANTA ERC Synergy Grant No. 951388.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., Tanaka, E., Jagan, M., & Bolker, M. B. (2015). Package 'lme4'. *Convergence*, *12*(1), 2

Boccato, T., Testolin, A., & Zorzi, M. (2021). Learning numerosity representations with transformers: number generation tasks and out-of-distribution generalization. *Entropy*, *23*(7), 857.

Carey, S. (2009). The Origin of Concepts. *Oxford University Press*.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in Cognitive Sciences*, 23(10), 823–835.

Carruthers, P. (2002) The cognitive functions of language. *Behavioral and Brain Sciences, 25,* 657-726.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., & Jitsev, J. (2022). Reproducible Scaling Laws for Contrastive Language-Image Learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818-2829.

Clarke, S. (2025). Number nativism. *Philosophy and Phenomenological Research*, *110*(1), 226-252.

Dehaene, S. (1997). The Number Sense: How the Mind Creates Mathematics. *Oxford University Press*.

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A dynamical model. *Cognitive Development*, 8(2), 95–112.

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., Salehi, M., Muenninghoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Branson, E., Ehsani, K., Ngo, H., Chen Y, Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng., K-H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N. A., Hajishirzi, H., Girshick, R., Farhadi, A., & Kembhavi, A. (2024). Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247-265.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv, abs/2010.11929*.

Dubova, M., & Goldstone, R. L. (2023). Carving joints into nature: reengineering scientific concepts in light of concept-laden evidence. *Trends in Cognitive Sciences*, *27*(7), 656-670.

Ethayarajh, K. (2019, November). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 55-65).

Everett, D., & Madora, M. (2012). The role of language in number concept development among Amazonian tribes. Cognition, 124(2), 152–167.

Frank, M.C., Tenenbaum, J.B., & Fernald, A. (2008). The role of language as a cognitive tool in number concept development. Cognitive Science, 32(5), 889–912.

Godey, N., de La Clergerie, E. V., & Sagot, B. (2024, March). Anisotropy Is Inherent to Self-Attention in Transformers. In *EACL 2024-18th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 35-48).

Gordon, R. (2004). Numerical cognition in Amazonian indigenous groups. *Science*, 306(5695), 496–499.

Luo, X., Sexton, N. J., & Love, B. C. (2023). A deep learning account of how language affects thought. *Language, Cognition and Neuroscience*, *38*(4), 499-508.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label‐feedback hypothesis. *Frontiers in Psychology*, 3, 54.

Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning*, *66*(3), 516-553.

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in cognitive sciences*, *24*(11), 930-944.

Margolis, E., & Lawrence, S. (2008). Concepts. *The Blackwell Guide to Philosophy of Mind*, 190.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological science*, *2*(6), 387-395.

Nieder, A. (2017). Number faculty is rooted in our biological heritage. *Trends in cognitive sciences*, *21*(6), 403-404.

Núñez, R. E. (2017). Is there really an evolved capacity for number? *Trends in cognitive sciences*, *21*(6), 409-424.

Pica, P., Lemer, C., Izard, V., Spelke, E., & Dehaene, S. (2004). Exact arithmetic in an Amazonian indigene group. Science, 306(5695), 499–503.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*.

Ridnik, T., Ben-Baruch, E., Noy, A., & Zelnik-Manor, L. (2021). ImageNet-21K Pretraining for the Masses. *ArXiv, abs/2104.10972*.

Saphra, N., & Lopez, A. (2019, June). Understanding Learning Dynamics Of Language Models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3257-3267).

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *ArXiv, abs/2210.08402*.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712.

Stoianov, I., & Zorzi, M. (2012). Emergence of a 'visual number sense' in hierarchical generative models. *Nature neuroscience*, *15*(2), 194-196.

Testolin, A., Dolfi, S., Rochus, M., & Zorzi, M. (2020a). Visual sense of number vs. sense of magnitude in humans and machines. *Scientific reports*, *10*(1), 10045.

Testolin, A., Zou, W. Y., & McClelland, J. L. (2020b). Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental science*, *23*(5), e12940.

Testolin, A., Hou, K., & Zorzi, M. (2024). Visual enumeration is challenging for large scale generative AI. *arXiv preprint arXiv:2402.03328*

Trott, S. (2024). Large language models and the wisdom of small crowds. *Open Mind*, *8*, 723-738.

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

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience*, 16(12), 1493–1504.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.