Probing LLM Numeracy: Modeling Low-Dimensional Internal Representations of Numbers of Large Language Models

Chirayu Nimonkar (chirayu@princeton.edu)
Department of Computer Science
Princeton University

Abstract

Numeracy is a critical component of higher-level reasoning, and it forms the basis for many tasks from performing mathematical operations to solving quantitative problems. Large Language Models (LLMs) perform well on most human benchmarks and tasks, but they have consistently failed when it comes to numeracy, even at its most basic level of arithmetic. In this paper, we apply multi-dimensional scaling (MDS) to pairwise similarity judgments of numbers $0-19$, extracting information on how both adult humans and GPT-4 internally represent them. The representations derived from this approach reveal that GPT-4 represents numbers at a very basic level (similar to that of children before they learn mathematics in school) compared to adults who categorize numbers by abstract numerical properties beyond only magnitude. This suggests LLMs may need to form the right numerical inductive biases to better represent numbers if they are to reach the level of adult humans in numerical reasoning.

Keywords: LLM numeracy, similarity judgments, internal representation, multi-dimensional scaling, clustering

Introduction

The problem of determining how people represent numbers has had a long history in psychology and cognitive science due to its centrality to everyday cognition (Cheyette & Piantadosi, 2020; Miller & Gelman, 1983; Nieder & Dehaene, 2009). Recently, the numeracy problem has also drawn attention in the context of large language models (LLMs) (Wu et al., 2023) due to their increasing ubiquity.

While these models seem to perform well on a variety of human benchmarks, they consistently struggle when it comes to arithmetic and numerical tasks—state-of-the-art LLM models like GPT-4 consistently fail at arithmetic and number sense (accounting for 68% of mathematical errors compared to 32% from misunderstanding the problem statement or wrong approach). Given just a 58% accuracy on elementary arithmetic problems (Bubeck et al., 2023), this is a severe limitation in numeracy and mathematical skills.

Unlike natural language tasks, mathematical reasoning (most evidently in arithmetic) has one exact answer, posing a unique challenge to LLMs (Imani, Du, & Shrivastava, 2023). A huge challenge presented by arithmetic problems is a need to "plan ahead" and look multiple steps ahead into a problem. One purported reason for the lack of this ability is that "the model has simply not been trained on enough data that involves arithmetic to develop the inner mechanisms that would allow it to perform successful ahead-planning" (Bubeck et al., 2023). However, it is unclear exactly how much data is needed and if the approach would even lead to success at all.

While there are clever workarounds to the LLM's lack of numeracy such as by delegating arithmetic to other programs (Wu et al., 2023; Imani et al., 2023), this does not address the underlying gap in general quantitative reasoning. This paper aims to present a more direct explanation not from an engineering perspective but rather one from cognitive science. This project investigates the internal representation of numbers in GPT-4 using similarity judgments and agglomerative clustering. This approach sidesteps many of the challenges faced by traditional analysis of numeracy in LLMs by focusing on the key component driving arithmetic error: a potentially faulty internal representation of numbers.

One previous study found significant correlations between human and GPT-4 internal representations such as the color wheel or pitch spiral (Marjieh, Sucholutsky, van Rijn, Jacoby, & Griffiths, 2023), but it is unclear whether GPT-4 has a similar human-like representation for numbers. Identifying the differences between GPT-4 and human internal representations will give more insight into how to better engineer future LLMs to have better numerical skills while utilizing a smaller dataset (more comparable to what an elementary school child may see in school).

Background

We briefly review the numerical internal representations of humans followed by an overview of why the same similarity judgment approach is effective for Large Language Models (LLMs).

Human Numerical Representation

Miller & Gelman previously showed how it is possible to characterize the development of numerical representation in children and adults (Miller & Gelman, 1983). Their technique utilized similarity judgments and multidimensional scaling (MDS), an approach initially proposed by Shepard for "constructing representations of the psychological structure of a set of stimuli on the basis of pairwise measures of similarity" (Shepard, 1980). Specifically, Miller & Gelman reconstructed an internal spatial representation of numbers MDS analysis on similarity judgment scores (rated $0-1$) of pairs of different numbers.

In the same paper, the internal representations of children were shown to mature from kindergarten, third grade, and sixth grade to adulthood as the child gained a more sophisticated understanding of mathematics and numerical reasoning. Spatially, the representations went from a sequential line of numbers increasing in magnitude to a complex web reflecting deeper relationships between numbers. The "basic characteristic" of numbers that participants used to judge numbers changed from magnitude (derived from counting) to features

like odd versus even and powers of two as children matured (Miller & Gelman, 1983; Tenenbaum, 1999).

Analyzing Representation in LLMs

The benefit of using Shepard's MDS approach is that it is possible to reconstruct a consistent internal representation simply from pairwise comparisons of numbers, meaning complex architectures from the human brain to LLMs can be modeled easily. In fact, due to the strength of LLMs in natural language, it is possible to prompt LLM "participants" directly using the same questions we can ask human participants. As pointed out in a paper by Tversky & Hutchinson, additional nearest-neighbor analysis using MDS methods can "help diagnose the nature of the data and shed light on the adequacy of the representation" (Tversky & Hutchinson, 1986). This kind of analysis was critical decades ago when cognitive psychologists were first discovering critical aspects of human thinking, and it can be a useful exploratory tool when interpreting LLMs.

Recently, Marjieh et al. (2023) applied the similarity judgment approach to LLMs in the context of perceptual domains. It was shown that judgments from GPT-4 produced internal representations significantly correlated with those of humans, including known representations such as the "color wheel" and "pitch spiral" (Marjieh et al., 2023). These representations also serve as inductive biases when reasoning beyond a learned set, giving ways of interpolating and extrapolating from learned data. It remains to be seen how LLMs internally represent numbers and—if they do so in a human-like way—if the representation is closer to that of a child or an adult.

Approach

To reconstruct the internal representation of numbers in GPT-4, there are two key steps: collecting pairwise similarity judgments for numbers $0-19$ and then applying non-metric multidimensional scaling (MDS) to create a spatial representation. By running the same suite of tests on existing data from adults, it is possible to compare internal representations as well as individual similarity scores. An agglomerative clustering step (plotted both on the MDS representation and visualized in a dendrogram) helps better quantify the groups in the spatial representation, giving insight into how humans and LLMs differently categorize numbers.

The strength of this approach is that it is effective at mapping out the internal representations just using similarity judgments, which has already been done in previous work (Marjieh et al., 2023). It is a natural approach to prompt an LLM with natural language to get similarity scores (identically as done with humans). Thus, by running the same experiment on humans and LLMs, it is possible to find the exact differences between human and LLM representation.

Methods

Similarity Judgments

This paper used a similar method described in Marjieh et al. for getting similarity judgments. The prompt consisted of one sentence describing the dataset and one describing the similarity rating scale and task (Marjieh et al., 2023). The similarity on a scale from $[0,1]$ was queried for every unique pairwise comparison of numbers in $[0,19]$ from GPT-4. Note that it was assumed that the matrix is symmetric, meaning that a comparison between a,b is the same as a comparison between b,a. The similarity matrix was then plotted for both GPT-4 and the human dataset.

Multi-Dimensional Scaling (MDS)

Using the similarity matrix from the previous step, non-metric MDS analysis was performed to construct a low-dimensional spatial representation of each number in space. Note that the actual input to the MDS was the dissimilarity ($1-$ *similarity*) because MDS analysis uses distances rather than similarity. The same analysis was conducted on the similarity data of the same experiment conducted on humans to reconstruct a similar spatial representation. From here, agglomerative clustering was used to extract groups and visual hierarchies. The MDS plots (with and without clusters) were then plotted.

Adult Human Similarity Data

Data from Griffiths & Kalish (2002) was used as a human benchmark for the LLM similarity comparisons. This data was collected from "twenty undergraduate psychology students from the University of Western Australia," where each student was asked to rate the similarity of various numbers in a set $0-99$ (Griffiths & Kalish, 2002). A normalization was applied to get a pairwise similarity rating in the range $0-1$ to construct a similarity matrix. In this paper, only the portion of the matrix corresponding with numbers $0-19$ was considered. From here, the same MDS step for the LLMs was applied to the human similarity data.

Results

We first plot the similarity matrix for LLM and human similarity judgments in Figure 1:

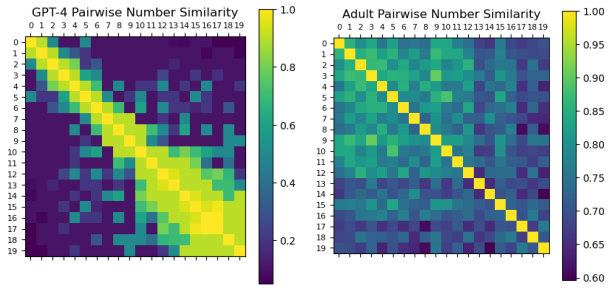

Figure 1: Pairwise Similarity Matrix of GPT-4 and Adult

The GPT-4 matrix shows a strong diagonal structure and consistently groups numbers closer together in magnitude as seen from the spread over the diagonal. This is in contrast to the adult humans, who did not always rate numbers closer together on the number line as more similar. GPT-4 also was more variable with similarity judgments than humans, rating with $\mu = 0.3752, \sigma = 0.3272$ compared to humans ($\mu = 0.7467, \sigma = 0.06435$) when not including the diagonal. The Pearson product-moment correlation coefficient between the two matrices is $r = 0.01789$, which is too low to consider correlated. Thus, GPT-4's internal representation is significantly different from that of adults.

The results can also be visualized as a spatial representation after applying nonmetric MDS on the similarity matrix as shown in Figure 2:

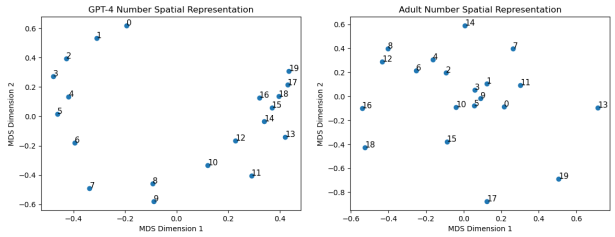

Figure 2: MDS Spatial Representation of GPT-4 and Adult

GPT-4 represents the numbers on a curved number line of increasing magnitude in contrast to the human representation, which has a more complex structure. The spatial representation from GPT-4 is very similar to the one found in children in kindergarten and third grade, who also placed numbers in a sequential increasing order (Miller & Gelman, 1983).

Using agglomerative clustering on these MDS spatial embeddings, it is possible to quantitatively extract the categories that both GPT-4 and adult humans grouped numbers into when making similarity judgments. The number of clusters $k = 8$ was selected as Tenenbaum also used that many clusters in their analysis, allowing for better comparison with previous studies (Tenenbaum, 1999).

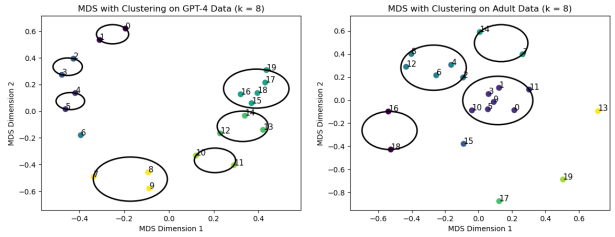

Figure 3: MDS Spatial Representation (with Clustering) of GPT-4 and Adult

As seen from figures 3 and 4, the larger groups GPT-4 uses are strictly by ranges in magnitude (e.g. $0-1$, $7-9$, $15-19$), while the adult data shows more sophisticated numerical groupings. Figure 4 shows the complex groupings

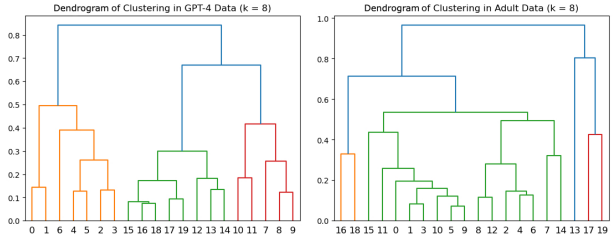

Figure 4: Dendrogram of GPT-4 and Adult Clustering over MDS

adults make in a dendrogram, compared with the groupings from GPT-4. Some examples of groups from the adult data set include big primes, numbers divisible by 7, big composites, evens, and more. In the GPT-4 dendrogram, there are only two examples ($16, 18$ and $17, 19$) where GPT-4 does not group numbers consecutively next to each other on the number line.

Discussion

Despite best efforts to improve Large Language Model (LLM) numeracy, even cutting-edge tools like GPT-4 struggle to match human ability on simple arithmetic tasks and numerical reasoning. The problem is quite difficult given the sheer scale and complexity of LLMs, and failures in numeracy are often attributed to insufficient training in arithmetic. However, the goal of this paper was to identify differences in a key component driving arithmetic error: a faulty internal representation of numbers.

A Child-Like Representation of Numbers

Spatial (and categorical) representations are one way of peeking into complex internal cognitive processes to derive insight into how humans (or LLMs) process information. The results from this paper show GPT-4 spatial and categorical representation was almost identical to that of a kindergartner-age child who has not yet learned basic mathematics (Miller & Gelman, 1983), in far contrast to the nuanced internal representations of adult humans. This suggests GPT-4 has a basic internal representation that does not show the same sophistication that adults gain after they learn "addition," "oddness and evenness of stimuli," "other multiplicative relations," and other mathematical concepts (Miller & Gelman, 1983), suggesting that GPT-4 has not fully learned basic numeric ability on a representational level.

As Bubeck et al., point out, GPT-4 may not be trained on sufficient data involving arithmetic to form a complete internal representation (Bubeck et al., 2023). Now, the similarity judgment approach from this paper provides a rough benchmark in evaluating progress as future versions of LLMs train on more data or use intelligent inductive biases.

Limitations & Next Steps

With access to only the average similarity rating instead of the individual human participants' responses to the similar-

ity task, it was not possible to run additional statistical tests such as forming a 95% confidence interval for ratings. With more human data points available, more tests can be run on GPT-4 to test variance in ratings, error, and other statistical measures.

Future work could also look at other basic operations such as multiplication (for which human data is already available) and identify where in a complex operation the model fails to "plan ahead." Looking at more complex tasks like proofs that combine natural language is also crucial as even workarounds for arithmetic do not work in this domain. To improve numeracy, methods such as prompting the model to have a better representation, training on select representative data, and explicitly teaching mathematical concepts may provide the inductive bias needed for LLMs to perform more accurately.

Conclusion

While it still may be insufficient to have a sophisticated internal representation of numbers to accurately reason numerically, a lack of one signals the absence of knowledge and skill that children can pick up with far more limited data than LLMs. As symbolic reasoning provides for an infinite space of questions, LLMs may soon face the same issues as humans do but on a larger scale. Especially since basic numeracy serves as a building block for higher-level logical reasoning, LLMs must form the right numerical inductive biases if they are to reason to the level of humans.

Acknowledgments

Thanks to Raja Marjieh for feedback on the paper and pointers for directing the project, Ilia Sucholutksy for support in setting up the GPT-4 experiments, and Tom Griffiths for guidance in the initial stages of the project and introduction to the topics in the paper.

References

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... Zhang, Y. (2023, April). *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* arXiv. Retrieved 2023-11-30, from http://arxiv.org/abs/2303.12712 (arXiv:2303.12712 [cs])

Cheyette, S. J., & Piantadosi, S. T. (2020, September). A unified account of numerosity perception. *Nature Human Behaviour*, *4*(12), 1265–1272. Retrieved 2023-12-12, from https://www.nature.com/articles/s41562-020-00946-0 doi: 10.1038/s41562-020-00946-0

Griffiths, T. L., & Kalish, M. L. (2002, January). A multidimensional scaling approach to mental multiplication. *Memory & Cognition*, *30*(1), 97–106. Retrieved 2023-10-24, from http://link.springer.com/10.3758/BF03195269 doi: 10.3758/BF03195269

Imani, S., Du, L., & Shrivastava, H. (2023, March). *MathPrompter: Mathematical Reasoning using Large Language Models.* arXiv. Retrieved 2023-11-30, from http://arxiv.org/abs/2303.05398 (arXiv:2303.05398 [cs])

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2023, June). *Large language models predict human sensory judgments across six modalities.* arXiv. Retrieved 2023-10-24, from http://arxiv.org/abs/2302.01308 (arXiv:2302.01308 [cs, stat])

Miller, K., & Gelman, R. (1983, December). The Child's Representation of Number: A Multidimensional Scaling Analysis. *Child Development*, *54*(6), 1470. Retrieved 2023-10-24, from https://www.jstor.org/stable/1129809?origin=crossref doi: 10.2307/1129809

Nieder, A., & Dehaene, S. (2009, June). Representation of Number in the Brain. *Annual Review of Neuroscience*, *32*(1), 185–208. Retrieved 2023-12-12, from https://www.annualreviews.org/doi/10.1146/annurev.neuro.051508.135550 doi: 10.1146/annurev.neuro.051508.135550

Shepard, R. N. (1980, October). Multidimensional Scaling, Tree-Fitting, and Clustering. *Science*, *210*(4468), 390–398. Retrieved 2023-12-12, from https://www.science.org/doi/10.1126/science.210.4468.390 doi: 10.1126/science.210.4468.390

Tenenbaum, J. B. (1999). A Bayesian Framework for Concept Learning.

Tversky, A., & Hutchinson, J. W. (1986). Nearest Neighbor Analysis of Psychological Spaces. *Psychological Review*.

Wu, Y., Jia, F., Zhang, S., Li, H., Zhu, E., Wang, Y., ... Wang, C. (2023, June). *An Empirical Study on Challenging Math Problem Solving with GPT-4.* arXiv. Retrieved 2023-11-30, from http://arxiv.org/abs/2306.01337 (arXiv:2306.01337 [cs, stat])