Title
The Representational Geometry of Number

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Hu, Zhimin
Niu, Lanhao
Varma, Sashank

Publication Date
2026

The Representational Geometry of Number

Zhimin Hu[1,2], **Lanhao Niu**[3] **& Sashank Varma**[1,2]
[1]School of Psychological and Brain Science, Georgia Institute of Technology
[2]School of Interactive Computing, Georgia Institute of Technology
[3]School of Informatics, University of Edinburgh

Abstract

A central question in cognitive science is whether conceptual representations converge onto a shared manifold to support generalization, or diverge into orthogonal subspaces to minimize task interference. While prior work has found evidence for both, a mechanistic account of how these properties coexist and transform across tasks remains elusive. We propose that representational sharing lies not in the concepts themselves, but in the *geometric relations* between them. Using number concepts as a target domain and language models as high-dimensional computational testbeds, we show that number representations preserve a stable relational structure across tasks. Task-specific representations are embedded in distinct subspaces, with low-level features like magnitude and parity encoded along near-orthogonal axes. Crucially, we find that these subspaces are largely transformable into one another via linear mappings, indicating that task-specific representations, despite being located in distinct subspaces, share relational structure.

Keywords: Representational Geometry, Conceptual Representation, Numerical Cognition, Large Language Models

Introduction

Representational geometry is a quantitative framework for investigating how the structure of representations supports intelligent behavior. A fundamental challenge in this area is the trade-off between generalization and task interference. On the one hand, Huh et al. (2024) suggests that as neural networks scale, they converge towards a shared ideal representation. On the other hand, empirical evidence suggests that systems often organize distinct tasks into task-specific subspaces to prevent interference (Libby & Buschman, 2021; G. R. Yang et al., 2019). Bernardi et al. (2020) proposes that neural representations resolve this trade-off by expanding representations into a high-dimensional space to make complex information linearly accessible, while ensuring that the underlying abstract categories are compressed into a low-dimensional structure.

Despite this progress, two significant gaps remain. First, existing research lacks a dynamic perspective; representations are often studied in static contexts or within a narrow class of tasks. Second, we lack a comprehensive, mechanistic account of how task subspaces are transformed into each other and how they relate to shared information (if this exists).

In this paper, we argue that the shared nature of neural representations lies not in the concepts themselves, but in the *geometric relations* between them. We hypothesize that task-specific representations can be understood as transformations applied to this common structure, and that they are projected into distinct subspaces to avoid task interference.

To investigate this proposal, we utilize number concepts as a target domain and large language models (LLMs) as computational testbeds. Numbers are ideally suited because (1) they support a number of well-defined tasks and (2) their mental representations, particularly the Mental Number Line (MNL), have been documented in mathematical cognition research (Dehaene, 2001). Furthermore, LLMs realize high-dimensional representational spaces, and their number representations show promising alignment with those of adults and children (Shah et al., 2024).

We analyze a diverse set of LLMs spanning different architectures, training objectives, and scales—from BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) to Qwen2.5-7B and its math-specialized variant (A. Yang et al., 2024, 2025). We extract contextual embeddings of numbers from number-related tasks (e.g., parity, successor) as task-specific representations. To examine their relational structure, we apply multiple representational analysis methods, including Procrustes Analysis (Gower, 1975), Subspace Overlap Analysis, and Canonical Correlation Analysis (Morcos et al., 2018).

To preview our results, we find that across tasks and models, the relative geometric relationships between numbers remain largely consistent, and these relationships largely match the psychophysical organization of the MNL. To minimize interference, models utilize distinct subspaces to represent different tasks. Low-level features such as parity and magnitude are encoded along nearly orthogonal axes. While task subspaces occupy different representational directions, they show similar relational structure: many task-specific representations can be mapped to each other via linear transformations. These findings reveal the representational geometry of number.

Related Work

Representational Geometry

Early theories of representation emphasized decomposing concepts into independent parts. Higgins et al. (2018) formalized disentanglement via group theory, proposing that ideal representations mirror the world's symmetries (e.g., object identity remains invariant under rotation). Similarly, Park et al. (2024) argued that concepts in LLMs correspond to linear directions in latent space, with semantic opposites (e.g., man vs. woman) lying in opposite directions.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Building on this work, there are two major lines of current research. The first explores how models mitigate task inference. G. R. Yang et al. (2019) found that neural networks trained on multiple cognitive tasks spontaneously cluster representations into functionally specialized regions. Libby and Buschman (2021) showed that neural circuits reformat sensory inputs via rotational dynamics into orthogonal memory subspaces over time. The second line focuses on generalization. Bernardi et al. (2020) argued that neural representations adopt geometries that support parallel coding directions across conditions in PFC and hippocampus, which supports cross-condition generalization. Most recently, Huh et al. (2024) posit that as models scale, they tend to converge on a shared, ideal representation of reality, invariant of model architecture or modality.

These works characterize the static constraints on representational geometries. What is missing is a mechanistic account of how they are dynamically transformed across tasks.

Numerical Cognition

Number representations in the mind are often conceptualized as belonging to two systems (Ansari, 2016; Cantlon, 2012). The Approximate Number System (ANS) is an evolutionarily ancient module, shared across species, for estimating and comparing magnitudes. The Exact Number System is a symbolic system that enables exact, rule-based calculation. This conceptualization is consistent with the earlier, influential triple-code model, which proposes that numerical information is represented in three distinct formats — Arabic (i.e., digits), verbal (i.e., words), and analog magnitude (i.e., the MNL) — each associated with different neural correlates (Dehaene, 2001). These systems interact dynamically when performing numerical tasks, mapping between codes as needed.

The ANS, which is critical code for estimating and comparing magnitudes, is often viewed as the key player. It gives rise to three behavioral effects which collectively imply a core representation for number: as a logarithmically compressed MNL. The Distance Effect is that the time to compare which of two numbers is greater increases with the numerical distance between them (e.g., 1 vs. 9 is compared faster than 1 vs. 2) (Moyer & Landauer, 1967). The Size Effect is that, for a fixed distance, smaller numbers are compared more quickly than larger ones (e.g., 1 vs. 2 is faster than 8 vs. 9) (Parkman, 1971). Finally, the Ratio Effect is that comparison time decreases as the ratio between the numbers increases (Halberda et al., 2008). Shah et al. (2023) showed that LLMs capture these effects, and thus the MNL structure, in their non-contextual embeddings. However, they did not address how number representations transform across different contexts (i.e., tasks), which is one goal of the current study.

Methods

Task Design and Material

To examine the representational geometry of numbers across tasks, we chose a set of numerical tasks from the mathematical cognition literature applied to the numbers 1–9. Numbers were presented in two formats: Digit (e.g., “3”) and English words (e.g., “three”). For each task, we crafted five sentences that elicited a specific cognitive operation. These sentences were inputted into LLMs, and we extracted internal activation vectors (embeddings) corresponding to the number token (shown in **bold** below). The tasks were the following:

- **Tangible Quantity**: Access the quantity of objects.
  *e.g., “I have a total of **3** apples.”*
- **Comparison**: Compare a number’s size to others.
  *Greater: e.g., ”A number bigger than two is **three**.”*
  *Smaller: e.g., ”A number lower than two is **one**.”*
- **Arithmetic**: Simple operations involving addition and multiplication, in two variants: *pre-equal* (target as operand) and *post-equal* (target as result).
  *Pre-equal: e.g., “The product of 1 and **3** is 3.”*
  *Post-equal: e.g., “Zero added to three gives **three**.”*
- **Properties**: Access the parity or primality of numbers
  *Primality: e.g., “Example of prime value is **5**.”*
  *Parity: e.g., “The set of odd numbers includes **three**”*
- **Ordinal**: Identify what comes before or after a number.
  *Successor: e.g., “The number after two is **three**.”*
  *Predecessor: e.g., “The number before four is **three**.”*

In addition to these task-specific sentences, to evaluate number representations in naturalistic settings, we constructed two sentence corpora. **Pseudo Sentences** contains 900 seven-word segments generated by chunking Wikipedia text, with the numbers randomly inserted 100 times each (e.g.,*“When **one** she was put off service and...”*) **Real Sentences** comprises 450 naturally occurring Wikipedia sentences containing the numbers, similarly chunked (e.g.*“After a **2** year long engagement, Hilda...”*)

Measures, Metrics, and Analysis Strategies

To quantitatively assess the geometric structure of numerical representations, we first examine the three effects that signal recruitment of magnitude representations (i.e., an MNL). Following Shah et al. (2023), we linked (model) cosine similarity to (human) reaction time. For each number pair (i, j), the distance effect was quantified by regressing similarity on $|i - j|$; the size effect by regressing normalized similarity on $\min(i, j)$; and the ratio effect by fitting a negative exponential function to similarity as a function of $x = max(i, j)/min(i, j)$.

We also apply Procrustes Analysis to estimate the degree to which the geometric relations between numbers are shared. Given two embeddings X and Y, it seeks a rigid transformation (translation, rotation, and uniform scaling) to minimize the: $M^2 = \min_{s,R,\vec{t}} \|sXR + \vec{t} - Y\|_F^2$, where s is a scalar, R is an orthogonal rotation matrix, and $\vec{t}$ is a translation vector. M^2 (disparity) represents the residual geometric distortion.

To assess whether one task spans the same subspace as another, we apply asymmetric Subspace Overlap Analysis. This evaluates how well task A’s top-k principal components (Abdi & Williams, 2010) explain variance in task B. This is

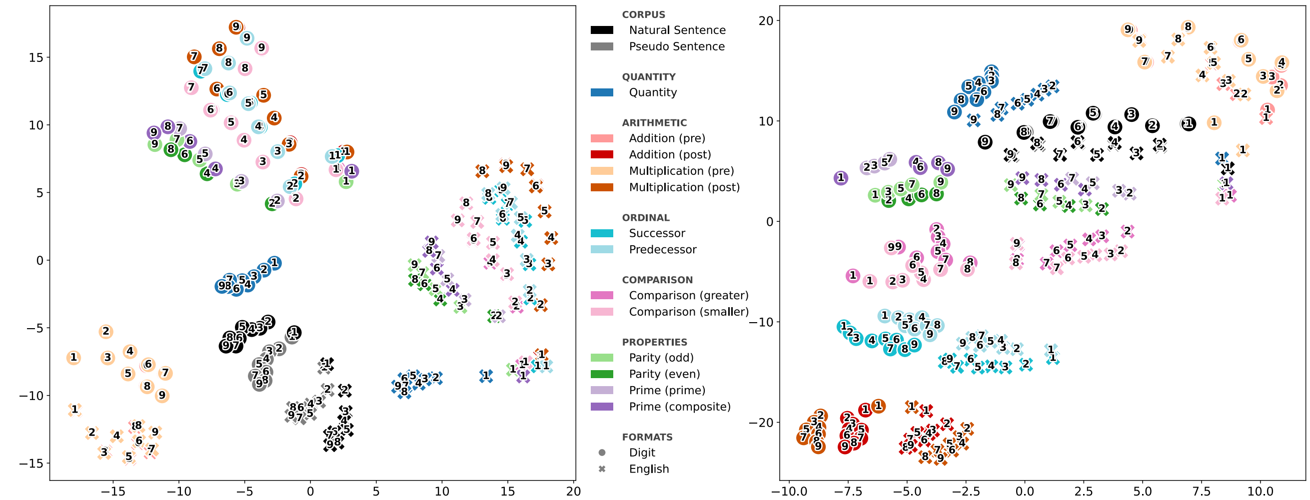

Figure 1: t-SNE projections of number representations extracted from BERT (left) and Qwen2.5-Math (right) across numerical tasks. Each point corresponds to a representation of a number (1–9). Colors indicate task categories, while marker shapes represent formats (digit vs. English word).

formalized as: $Overlap(A \leftarrow B) = \mathrm{Var}(\mathrm{Proj}_A(B))/\mathrm{Var}(B)$. $\mathrm{Proj}_A(B)$ denotes the projection of B onto the subspace spanned by A's top-k principal components.

To quantify the functional equivalence between two representations under linear transformations, we employed Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al., 2017). This first applies Principal Component Analysis (PCA) to both tasks and retains their top n components. Canonical Correlation Analysis (CCA) (Morcos et al., 2018) is then used to find pairs of projection vectors (a, b) that maximize: $\rho = \max_{a,b} \mathrm{corr}(a^T X_A^{PC_i}, b^T X_B^{PC_i})$, yielding a sequence of $\rho_1, \rho_2, \ldots, \rho_n$. Their average reflects the overall similarity between task representations.

Models

To examine how conceptual representations differ across architectures, training objectives, and domains, we selected four representative LLMs in their base (pre-trained only) forms. We included BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) to isolate the effect of training objectives. BERT is trained to predict masked tokens utilizing bidirectional context (left and right). In contrast, GPT-2 learns representations by predicting only the next token. Qwen2.5 (A. Yang et al., 2025), a modern high-capacity model, offers insight into number concepts in higher-dimensional space. Qwen2.5-Math (A. Yang et al., 2024), built on Qwen2.5, is further pre-trained on a large amount of mathematical corpora, allowing assessment of how domain-specific training reshapes the geometry of number representations.

Results

We primarily use representations from the 75% depth of the layers for all experiments because middle-to-late layers are generally found to exhibit stable brain alignment and semantic structure (Caucheteux & King, 2022). We present results for BERT and Qwen2.5-Math, as they are representative of the broader trends observed across the four models. Embeddings are scaled by L2 normalization. Unless otherwise specified, each task-specific number representation is defined as the centroid of its embeddings within that task.

Shared Geometric Structure

We first examine whether numerical concepts are embedded in a consistent relational geometric structure across tasks, number formats, and models, and whether such a structure reflects the three effects of MNL.

To visualize the global organization of embeddings across tasks, we projected them into a 2D space using t-SNE. As shown in Figure 1, both models exhibit a highly organized layout where numbers 1–9 show a consistent linear order within their respective task clusters. In BERT, the Comparison, Ordinal, and Arithmetic tasks show significant overlap in their geometric layout. Another point of interest is that the two number formats (i.e., digits vs. words) occupy distinct subspaces of the latent space, suggesting that the semantic identity of a number does not fully converge. This is in contrast to the Platonic hypothesis (Huh et al., 2024). Interestingly, their structures are reflected across a diagonal axis, implying that while the model distinguishes number formats, it uses a shared relational structure for the underlying number representations. Qwen2.5-Math shows a quite different segregation of the tasks, locating number representations for the same task across both formats in the same subspace. In addition, Addition and Multiplication are separated in post-equal condition and form seemingly symmetric clusters.

We turn next to the question of the MNL in these models. We find that the distance, size, and ratio effects are largely

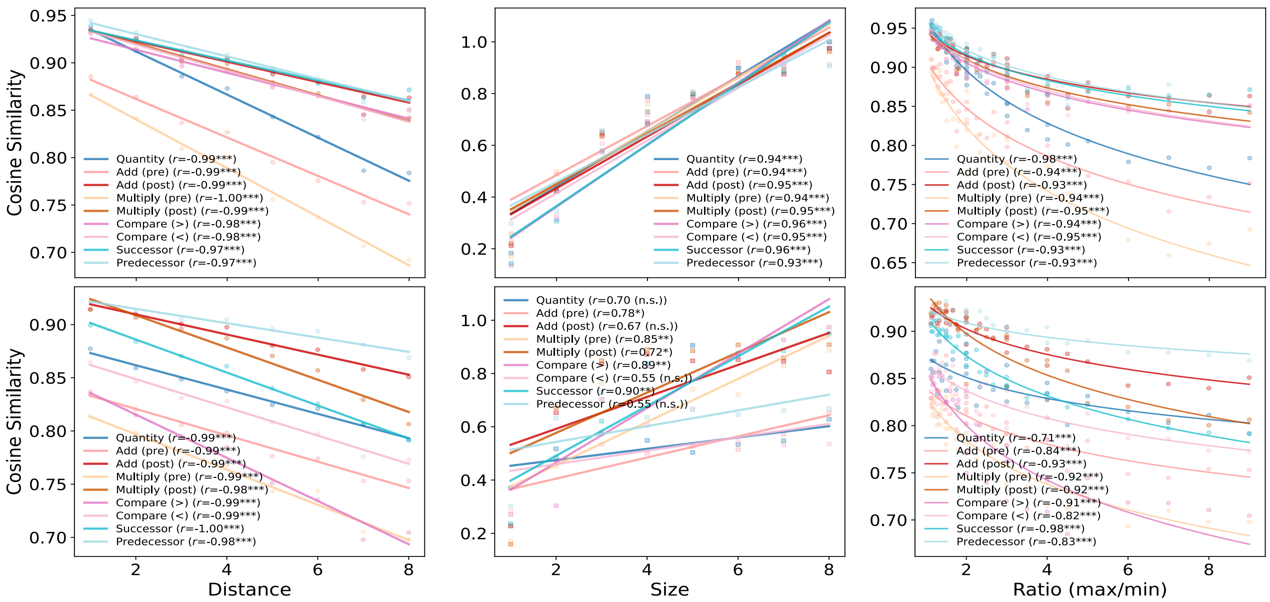

Figure 2: Three effects fitted across models. Cosine similarity between number embeddings (Digit) is plotted against numerical distance (left), size (middle), and ratio (right) for each task. Top row: BERT; bottom row: Qwen2.5-Math. Each line represents a linear (left/middle) or logarithmic (right) fit for a given task (color-coded), with absolute Pearson's r and significance level indicated in legend ($p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***).

present across tasks.[1] As shown in Figure 2, both models exhibit a near-perfect fit for the distance and ratio effects, indicating that the relative positioning of numerical concepts across tasks is consistent with an MNL representation. For the size effect, BERT shows a strong fit across all tasks, whereas Qwen2.5-Math exhibits greater variance.

To evaluate whether the geometric structure of number representations is a shared property across models, we applied Procrustes Analysis (Gower, 1975). Across all pairs of models, we observed a remarkably low average disparity between task-specific subspaces ($mean \approx 0.010$). To establish the significance of this alignment, we generated a permutation baseline by randomly shuffling numerical identities. Doing so produced significantly higher disparities (ranging from 0.077 to 0.273). This substantial gap confirms that the observed geometric alignment is highly non-random, suggesting that a consistent relations structure of numbers emerges across fundamentally different model families.

These findings reveal a structured spatial organization of number representations and also clear clustering based on tasks. The spatial organization is that of the MNL documented for humans. In particular, the tasks shown in Figure 2 largely show the distance, size, and ratio effects, with some model-specific variation. Further, Procrustes Analysis suggests that this relational structure among number representations is shared across tasks.

[1]Note that we excluded the parity and primality tasks from this analysis as they exhibit a different pattern; see the next section.

Orthogonality and Linear Equivariance

Although the prior section confirmed the shared relational structure, it remains unclear how the representational geometry supports task flexibility. As Qwen2.5-math demonstrated task-specific subspaces, we further explored how it maintains functional properties of numbers without interference.

To explore the encoding of low-level features like magnitude and parity, we visualize the subspace of the parity and primality tasks by defining a 2D projection on the raw embeddings (instead of the mean). In this projection, one axis captures numerical magnitude (via linear regression) and the other reflects categorical distinctions, parity (odd vs. even) or primality (prime vs. composite), derived from the mean difference between group centroids. As shown in Figure 3, the parity axis is nearly orthogonal to the magnitude axis (88.8°), while the primality axis shows a noticeable tilt (68.8°). For parity, numbers cluster symmetrically into odd groups (below) and even groups (above) regardless of input format (digit or word), indicating that parity is encoded as a categorical feature. In contrast, the representation of primality shows partial coupling with that of magnitude, consistent with the fact that prime numbers become sparser with increasing magnitude. This analysis reveals how the model encodes distinct numerical properties along near-orthogonal axes.

Moving from low-level features to functional relationships, we further analyzed the relationships between task-specific subspaces using Subspace Overlap Analysis and SVCCA. For both methods, we used raw embeddings and set the number of principal components to the averaged number of components needed to explain 95% of the variance across tasks.

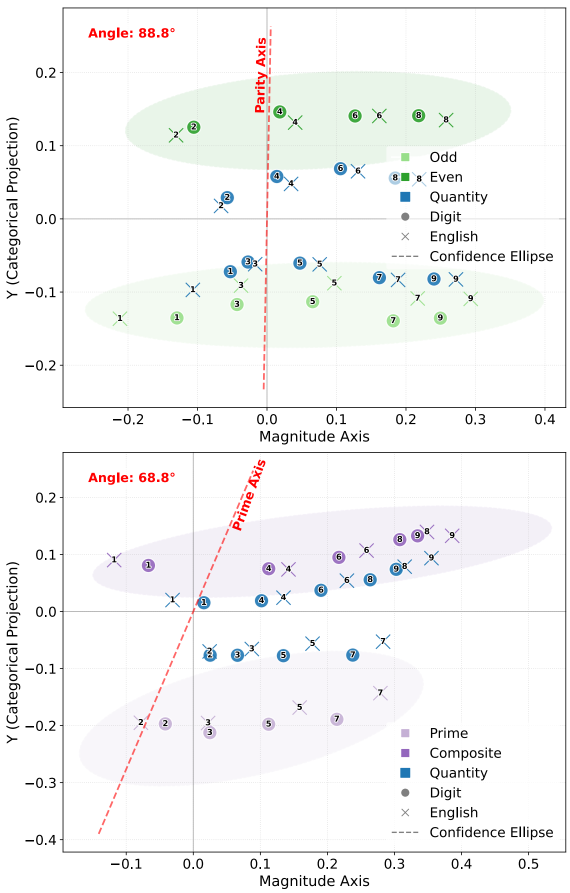

Figure 3: Disentanglement of magnitude and parity/primality in Qwen2.5-Math. The x-axis represents the magnitude axis, and the red dashed line represents the categorical axis (odd vs. even above, prime vs. composite below). The y-axis shows orthogonal projection for visualization only. Circles and crosses denote formats (the means of embeddings); shaded ellipses mark 95% confidence regions of raw embeddings.

As shown in Figure 4, most task pairs exhibit minimal subspace overlap (typically < 0.2), with notable exceptions only among functionally related tasks such as Successor and Predecessor (0.36–0.42). This indicates that the model encodes tasks in highly decoupled subspaces. Despite this low physical overlap, SVCCA reveals consistently high canonical correlations (typically 0.80–0.90) across task pairs, suggesting that although subspaces are separated to reduce interference, they retain a shared representational structure. Notably, the representations for parity and primality show lower SVCCA alignment compared to other tasks. This reduction in correlation implies that these task subspaces are more structurally distinct and cannot be fully aligned via simple linear transformations. This is because of the non-linear structure of these task-specific number representations, each of which includes a nearly-orthogonal second dimension to encode relevant information, as shown in Figure 3.

Taken together, the low subspace overlap and high (but variable) SVCCA scores support our central claim: numerical representations in LLMs are not fixed coordinates but stable, transformable geometries. To achieve efficient task-switching while minimizing task interference, such geometric structures are then transformed largely through linear mapping to distinct subspaces.

Distribution of Number Concepts

To close, we provide a global distribution of QWEN2.5-MATH's numerical concept space under more natural semantics. We project raw embeddings (instead of mean) of numbers from both the corpus and handcrafted task sentences using PCA and visualize their concentration with kernel density estimation (Rosenblatt, 1956). This is a nonparametric method that yields a smooth estimate of where samples cluster in the projected space. Despite the limited variance captured in 2D, the projected distributions exhibit a stable ordinal organization across numbers, consistent with the shared relational structure discussed; see Figure 5. Crucially, highlighted trajectories (e.g., Quantity vs. Multiplication) occupy distinct regions within this distribution, echoing our subspace-level results.

Furthermore, the near-parallel layouts for Digit and English inputs further indicate that the same relational structure is preserved despite being processed through separate representational channels. This separation is achieved without altering the distributional properties of the representations. We quantified this using representation sparseness (Rolls & Tovee, 1995), which measures how neural activity is distributed across the available dimensions. We found that sparseness remains largely similar across formats and numbers (Digit: 0.137 ± 0.007; English: 0.156 ± 0.012). It suggests that the model maintains a consistent level of feature concentration, which echoes the consistent topological density shown in Figure 5. Together with the previous results, it implies that the model does not merely reuse the shared relational structure but preserves a consistent coding resolution.

Discussion

In this study, we show that while individual numbers occupy distinct vector locations across tasks, what remains invariant is their relational structure. [2] This suggests that what it means to be a given number is not encoded in any single vector, but rather emerges from its position within a relational geometry. Number concepts, in this view, are not points but relations. This provides support for Gardenfors (2004)'s conceptual space theory. Meanwhile, it moves beyond the idea of concepts solely as vectors (Piantadosi et al., 2024), suggesting that the relational structure is also an essential property of concepts.

Our findings connect to other theories of representational geometry, including the Platonic representation hypothesis (Huh et al., 2024), the linear representation hypothesis (Park

[2] These results are not confined to a specific layer of models. We observe that both middle and late layers consistently preserve the geometric organization.

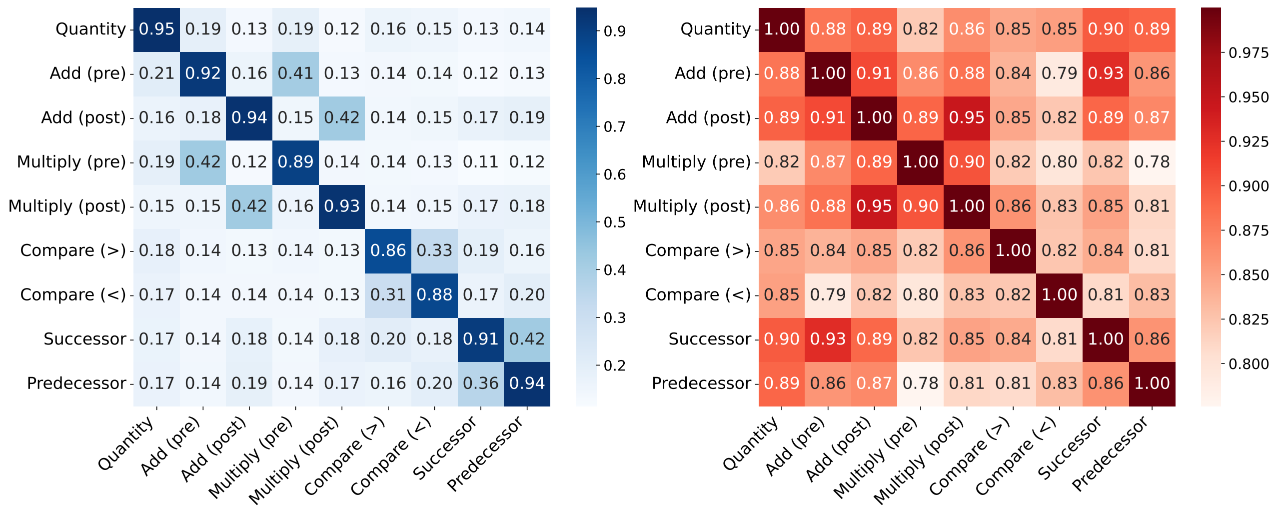

Figure 4: Task-specific representation (Digit) alignment in Qwen2.5-Math. The left panel shows an asymmetric Subspace Overlap Ratio, while the right panel shows symmetric SVCCA similarity.

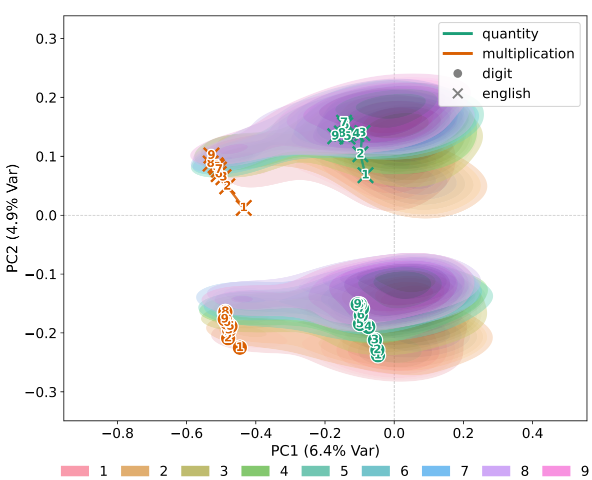

Figure 5: Distribution of number concepts in Qwen2.5-Math. Raw embeddings are projected using PCA (6.4% and 4.9% variance) and visualized using KDE that encloses the top 80% highest-density region.

et al., 2024), and the task-disentangled subspace perspective (G. R. Yang et al., 2019). In our study, each task occupies a distinct subspace, yet they are largely related through linear transformations, preserving a similar underlying relational structure. In addition, number features such as magnitude and parity are represented along nearly orthogonal axes in the representational space. While we differ in emphasis, our account unifies them under a common relational perspective and has the potential to extend beyond number concepts.

These findings also bear on a foundational question in neural systems and representation: if the same behavior or structure can arise from multiple implementations, what constrains representational convergence? One answer is the interplay between system capacity and environmental complexity. Our results suggest that, unlike the diverse shortcut representations observed in toy neural networks (Hu et al., 2025), all large-scale models, varying in size, training data, and architecture, share the same underlying geometry of numerical concepts. Hu et al. (2025) has shown that in small systems, representational redundancy often impedes the emergence of stable geometry. However, when systems face increasingly complex learning environments, as our findings demonstrate at scale, they are pressured to converge to a more structured coding scheme. This suggests that knowledge representation is not necessarily innate (Spelke & Kinzler, 2007), but is a learned structural adaptation to the demands of the data.

Unlike the human brain, LLMs are not constrained by biological wiring or evolutionary contingencies. Rather, they are an unconstrained substrate on which representations can potentially organize more freely. As such, they offer a powerful tool for probing what kinds of representational structures are possible for supporting thought, and also how and why such structures may succeed or fail to emerge in the human. For example, numerical cognition in humans is implemented over anatomically distributed systems (Dehaene, 2001). If one system lacks sufficient scale or connectivity, it may fail to converge onto the expected geometric abstraction. This may help explain why individuals with dyscalculia, whose intraparietal sulcus (IPS) often shows reduced gray matter volume or atypical connectivity (Butterworth et al., 2011), struggle with higher-order number representations.

Acknowledgments

Portions of the code used to analyze experimental results were developed with the assistance of AI tools, which provided suggestions for Python implementation, data visualization, and debugging.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Ansari, D. (2016). Number symbols in the brain. In *Development of mathematical cognition* (pp. 27–50). Elsevier.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, *183*(4), 954–967.

Butterworth, B., Varma, S., & Laurillard, D. (2011). Dyscalculia: From brain to education. *science*, *332*(6033), 1049–1053.

Cantlon, J. (2012). Math, monkeys, and the developing brain. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 10725–10732.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Dehaene, S. (2001). Précis of the number sense. *Mind & language*, *16*(1), 16–36.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186.

Gardenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A. (2018). Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Hu, Z., Cingiler, O., Bohm, C., & Albantakis, L. (2025). From function to implementation: Exploring degeneracy in evolved artificial agents. *Neural Computation*, *37*(9), 1677–1708.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The platonic representation hypothesis. *International Conference on Machine Learning*, 20617–20642.

Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature neuroscience*, *24*(5), 715–726.

Morcos, A., Raghu, M., & Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, *31*.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520.

Park, K., Choe, Y. J., & Veitch, V. (2024). The linear representation hypothesis and the geometry of large language models. *International Conference on Machine Learning*, 39643–39666.

Parkman, J. M. (1971). Temporal aspects of digit and letter inequality judgments. *Journal of experimental psychology*, *91*(2), 191.

Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, *28*(9), 844–856.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raghu, M., Gilmer, J., Yosinski, J., & Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, *30*.

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of neurophysiology*, *73*(2), 713–726.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), 832–837.

Shah, R., Bhardwaj, K., & Varma, S. (2024). Development of cognitive intelligence in pre-trained language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9632–9657.

Shah, R., Marupudi, V., Koenen, R., Bhardwaj, K., & Varma, S. (2023). Numeric magnitude comparison effects in large language models. *Findings of the Association for Computational Linguistics: ACL 2023*, 6147–6161.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., & others, H. W. (2025). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., et al. (2024). Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, *22*(2), 297–306.