DO LARGE LANGUAGE MODELS PERCEIVE ORDERLY NUMBER CONCEPTS AS HUMANS?

Xuanjie Liu* MBZUAI, UAE
Cong Zeng* MBZUAI, UAE
Shengkun Tang MBZUAI, UAE
Zhiqiang Xu MBZUAI, UAE

Ziyu Wang † New York University, USA
Gus Xia† MBZUAI, UAE

ABSTRACT

Large language models (LLMs) have demonstrated powerful abilities in reasoning and mathematics. However, due to their black-box nature, conventional theoretical methods struggle to analyze their internal properties. As a result, researchers have turned to cognitive science perspectives, investigating how LLMs encode concepts that align with human cognition. Prior work has explored constructs such as time and spatial orientation, revealing the alignment between LLM representations and human cognition. Despite this progress, an important concept in human reasoning—**numbers**—remains underexplored. In this paper, we examine numerical concepts by introducing a metric, *orderliness*, to assess how number embeddings are spatially arranged across LLM layers, drawing parallels to the human mental number line. Our experiments reveal that LLMs initially encode numerical order in a structured manner, as evidenced by high orderliness in shallow layers. Using our proposed metric, we observe a two-phase decline in orderliness across layers. Through further analysis of LLaMA 3.1, we identify this decline as being closely linked to contextualization and next-token prediction. Our findings shed light on how LLMs encode numerical concepts, offering a novel perspective on their internal representation of ordered information and its potential alignment with human numerical cognition. Our code and data are released at `https://github.com/cong-zeng/LLM-mental-number-line`.

1 INTRODUCTION

Large Language Models (LLMs) have become one of the most prominent areas of research in artificial intelligence. Their effectiveness in various natural language tasks, such as machine translation, sentiment analysis, and question answering has been well-established (Naveed et al., 2023). Besides, LLMs have also shown remarkable capability in complex mathematics, achieving competitive results in high-level challenges (Frieder et al., 2024). However, despite their demonstrated capabilities, LLMs are often regarded as "black box" models due to a critical lack of interpretability (Wallace et al., 2019a), leaving a significant gap in understanding how they internally process and represent information (Lyu et al., 2024; Madsen et al., 2022; Luo et al., 2024).

One way to bridge this gap is by exploring whether LLMs develop cognitive structures that mirror human models of the real world drawing inspiration from cognitive science. Recent research (Gurnee & Tegmark, 2024) has demonstrated that LLMs, like LLaMA, encode spatial and temporal concepts using "space neurons" and "time neurons" that align with real-world coordinates, suggesting an ability to mirror real-world structures. Similarly, (Huh et al., 2024) indicates that large-scale AI models not only tend to demonstrate the representation convergence between different modalities such as vision and language, but also show similar visual perception as human brains, which implies a "platonic" representation of reality. Based on the findings that deep networks' internal representations tend to align, we start to question whether LLMs develop structured, cognitive-like representations, similar to human numerical cognition.

*These authors contributed equally
†Corresponding authors

One of the most fundamental phenomena in human numerical cognition is the mental number line. First proposed by a mathematician (Wallis, 1685) and now widely studied in psychology, the mental number line represents a spatial-numerical association shared by humans and some animals (Rugani et al., 2015; Vallortigara, 2018). Numbers are typically conceptualized along a left-to-right horizontal axis, with smaller values positioned on the left and larger values on the right (Giurfa et al., 2022). Numerous studies have shown that the development of a mental number line in children is strongly associated with improving numerical knowledge and basic arithmetic skills (Honour, 2020; Lin, 2022). Today, LLMs have far surpassed children in mathematical ability and, in some cases, can even perform on par with humans in mathematical competitions (Yuan et al., 2023; Trinh et al., 2024). This naturally raises the question: *Do LLMs perceive numerical concepts in a way similar to humans?* Specifically, in the hidden states of LLMs, can number representations form a structured "number line" in high-dimensional space, akin to the mental number line observed in human cognition?

To investigate this, we first define what constitutes a number line in high-dimensional space. While the mental number line exhibits many intriguing properties, we argue that its most fundamental characteristic is that numbers are ordered in space, **meaning that sequentially adjacent numbers should also be spatially close**. Clearly, a one-dimensional number line satisfies this property. Based on this principle, we propose *orderliness*. This objective metric quantifies how well a set of elements maintains its intrinsic order in a given space of arbitrary dimensionality, allowing us to assess the structured arrangement of number embeddings within the hidden space of LLMs. However, capturing a complex phenomenon with a single metric inevitably entails a loss of information. Therefore, in addition to computing the orderliness of high-dimensional embeddings corresponding to numerical concepts in LLMs, we further employ t-SNE (van der Maaten & Hinton, 2008) to project these embeddings into a lower-dimensional space for visualization. This complementary approach provides an alternative perspective and offers a more intuitive understanding of how numerical concepts construct the model's mental number line.

Our experiments analyze six open-source LLMs to evaluate embeddings across different layers. In all models, we find a specific layer near the input where **number embeddings exhibit a strikingly ordered structure**, as evidenced by both their high orderliness scores and direct observations in low-dimensional projections. In other words, the shallow layers of LLMs contain a mental number line, remarkably similar to human numerical cognition (section 4.2). Interestingly, as the token embeddings pass to deeper layers, this orderliness undergoes two distinct declines, a pattern evident in all models and most significant in LLaMA 3.1. To further investigate this phenomenon, we conduct further experiments on LLaMA 3.1, exploring potential underlying causes for this progressive loss of numerical structure (section 5.1).

To sum up, our key contributions are as follows:

- We propose a novel metric called orderliness to quantify the spatial arrangement of numerical embeddings, providing an approach to studying numerical cognition in LLMs. This metric can be extended to assess how well the spatial arrangement of any sortable elements preserves their inherent order, irrespective of the space's dimensionality.
- We conduct comprehensive experiments on mainstream LLMs including Mistral, Llama, Qwen, and so on with the orderliness metric and find the embeddings corresponding to numerical concepts in LLMs are arranged in a highly ordered manner within the hidden space of the shallow layers, demonstrating the existence of the Number Line in LLMs. Moreover, we identify a double descent trend in the orderliness metric across layers in the mainstream LLMs, offering a new understanding of the numerical representations in LLMs.

2 Related Work

Numerical Representations in LLMs Recent research has delved into the numerical capabilities and internal representations of Large Language Models (LLMs). (Wallace et al., 2019b) investigated the numeracy embedded in NLP models, examining their ability to understand numerical magnitude and order, emphasizing the need for more robust models of numeracy. (Park et al., 2023) expanded on this by proposing the Linear Representation Hypothesis, suggesting that high-level concepts, including numerical information, are encoded linearly in LLM embeddings. Their findings suggest that numerical representations in LLMs could have geometric properties that parallel cognitive structures, offering a new perspective on model interpretability. Building on this, (Zhang et al., 2024)

extended this line of work by identifying attention mechanisms and specific layers associated with arithmetic reasoning, proposing fine-tuning methods to enhance task performance. While these studies have identified key patterns in numerical representations, they primarily address specific aspects such as arithmetic tasks or contextual adjustments. The most fundamental and universal characteristics of numerical representations—those invariant across tasks, models, and contexts—remain unexplored.

Function Alignment Between LLMs and Human Cognition In cognitive science, the mental number line has been widely studied as a fundamental framework for understanding human numerical cognition. (He et al., 2021) distinguished between symbolic and non-symbolic number representations, providing insights into how humans spatially organize numerical information, laying the groundwork for examining how LLMs encode numerical relationships. (Meng et al., 2022) further explored the interpretability of LLMs, locating neuron activations responsible for numerical and factual knowledge. These studies collectively underscore the growing interest in understanding the internal mechanisms of LLMs and their parallels with human cognition, while a deeper understanding of whether and how these representations inherently reflect universal cognitive principles remains underexplored.

3 METHODOLOGY

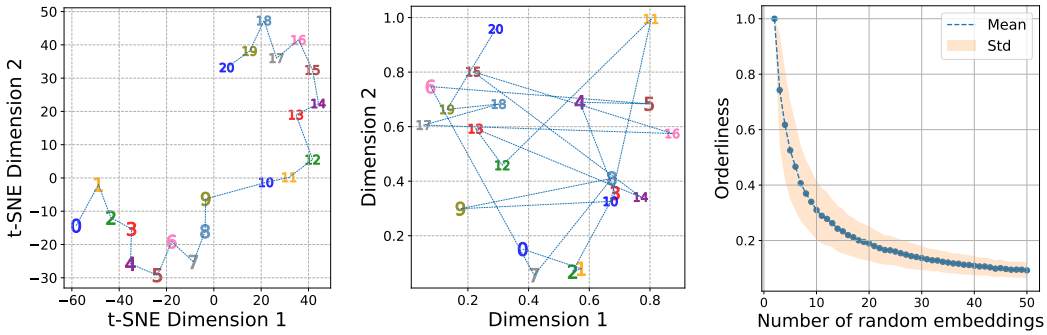
Our methodology consists of three main components. **1) Embedding Extraction** (section 3.1) – We obtain the embeddings corresponding to numerical concepts from the hidden space of LLMs by designing appropriate prompts to extract them effectively. **2) Orderliness Definition and Computation** (section 3.2) – We introduce an objective metric, orderliness, to quantify how well the spatial arrangement of embeddings preserves the intrinsic numerical order of their corresponding values. **3) Dimensionality Reduction and Visualization** (section 3.3) – To analyze the spatial structure of embeddings from another perspective, we apply a dimensionality reduction method that preserves orderliness as much as possible. This allows for a more intuitive understanding of embedding distributions, serves as a cross-validation of the orderliness metric, and reveals additional structural details that orderliness alone may not capture.

3.1 CONTEXT-FREE EMBEDDING EXTRACTION

We use Arabic numerals as prompts to obtain the corresponding numerical embeddings. When a number is input as text, the model's tokenizer converts it into tokens, which are then mapped to high-dimensional embeddings in the hidden space. These embeddings correspond to individual tokens and encode a vast amount of information, reflecting the model's semantic understanding of the tokenized text. To extract embeddings that represent numerical concepts as purely as possible—minimizing interference from linguistic patterns, contextual associations, or model-specific prompt biases—we input each integer from 0 to 200 into the models individually (e.g., "42") without any additional context or prompts. If a model's tokenizer splits a number into multiple tokens, we use the embedding of the last token to represent the number. This choice is motivated by the self-attention mechanism, where the final token attends to all preceding tokens, allowing it to integrate global information. Alternative strategies for handling multi-token numbers, such as averaging token embeddings, are discussed in the Appendix A.1.

3.2 THE ORDERLINESS OF NUMERICAL REPRESENTATION

We design the algorithm based on an intuitive principle of arrangement orderliness: elements adjacent in their intrinsic order should also remain spatially close. Formally, for a set of n vectors $E = \{e_i \mid i \in \{1, 2, 3, ..., n\}, e_i \in \mathbf{R}^d\}$, where d is the dimension of vector e_i, i represents the ordinal index of these vectors after being sorted in a certain manner. Given any distance function $f_d(e_i, e_j)$, we check if e_{i+1} is the closest vector among all vectors greater than e_i, and if e_{i-1} is the



(a) 21 Embs in LLaMA 3.1, Orderliness = 0.95

(b) 21 random Embs, Orderliness = 0.25

(c) Orderliness statistics for 2-50 random embeddings.

Figure 1: An illustration of embedding orderliness. The distance function f_d is the Euclidean distance.

closest among all vectors less than e_i. The complete calculation process is as follows:

$$Orderliness(E) = \frac{1}{2(n-1)} \left(\sum_{i=1}^{n-1} O_{\text{right}}(i) + \sum_{i=2}^{n} O_{\text{left}}(i) \right) \in (0, 1] \tag{1}$$

$$O_{\text{right}}(i) = \begin{cases} 1, & \text{if } f_d(e_i, e_{i+1}) = \min\{f_d(e_i, e_j) \mid j \in \{i+1, \ldots, n\}\} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$O_{\text{left}}(i) = \begin{cases} 1, & \text{if } f_d(e_i, e_{i-1}) = \min\{f_d(e_i, e_j) \mid j \in \{1, \ldots, i-1\}\} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

This metric ranges from 0 to 1. The closer it is to 1, the more orderly the arrangement of these vectors. A set of elements with an orderliness score of 1 implies that, for any given element e_i, the next element in the sequence e_{i+1} is always the closest among all larger elements $e_j : j > i$, and the previous element e_{i-1} is always the closest among all smaller elements $e_j : j < i$. This structure aligns with human intuition about numerical ordering – the mental number line.

To illustrate the concept of orderliness more intuitively, Figure 1 presents two examples: 1a shows 21 elements with an orderliness of 0.95, while 1b depicts 21 elements with an orderliness of 0.25. As observed, 1a exhibits a stronger sense of order compared to 1b. Furthermore, Figure 1c shows the mean and standard deviation of orderliness scores for randomly arranged sets of 2 to 50 elements, computed separately for each set size over 1,000 random initializations, providing a baseline for unordered distributions. Crucially, this metric is applicable to spaces of any dimensionality, which is essential since numerical concepts in LLMs are represented as embeddings in hundreds or thousands of dimensions.

3.3 Embedding Visualization through Dimensionality Reduction

To visualize the orderliness of numerical embeddings, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008) for dimensionality reduction. As a nonlinear technique, t-SNE excels at preserving local structure and revealing clusters in high-dimensional data when projected into lower dimensions. Although we experimented with various alternatives—including PCA (F.R.S., 1901), Truncated SVD (Hansen, 1990), and other nonlinear methods (Mehrbani & Kahaei, 2021; Tenenbaum et al., 2000)—only t-SNE consistently maintains the orderliness of the embeddings during the dimensionality reduction process. Formally:

$$E_{\text{low-dim}} = tSNE(E_{\text{high-dim}}, perplexity) \tag{4}$$

We adjust the parameter $perplexity$ such that

$$\Delta O = |O(E_{\text{high-dim}}) - O(E_{\text{low-dim}})| < \epsilon \tag{5}$$

where $O(E_{\text{high-dim}})$ is the orderliness metric computed on embeddings in their original high-dimensional space, $O(E_{\text{low-dim}})$ denotes the metric computed on those dimension-reduced by t-SNE. ϵ is a predefined threshold indicating an acceptable level of change in orderliness. This criterion

ensures that t-SNE reliably retains the sequential numerical relationships inherent in the original embeddings, preserving the spatial order crucial for visualization and subsequent analysis. In our experiments, we set ϵ to 0.02.

4 EXPERIMENTS

4.1 SETTINGS

Models Different open-source large language models typically belong to a series, with multiple variants differing in parameter sizes. To balance computational efficiency and inference capability, we select representative mid-sized versions from each model series for evaluation. Specifically, we select six lightweight open-source large language models for evaluation: LLaMA 3.1-8B (Dubey et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen 2.5-7B (Yang et al., 2024), Gemma 2-9B (Team et al., 2024), LLaMA 2-7B (Touvron et al., 2023), and Phi 3.5-4B (Abdin et al., 2024). This selection ensures coverage of diverse architectures while maintaining feasible computational requirements. Additionally, We access these models and their corresponding tokenizers via the Hugging Face Transformer Packages[1] to ensure a standardized evaluation framework.

Layer-wise Orderliness Computation The representations learned by different layers of LLMs encode varying levels of semantic information. Given a numerical prompt such as "13", some layers may focus on its numerical properties—such as its relationship to other numbers—while others may emphasize non-numerical associations, including cultural connotations (e.g., "13" being considered unlucky). Since orderliness captures only the sequential structure of numbers, its value is expected to vary across layers, reflecting the transition from raw numerical representations to more context-dependent meanings. To systematically analyze this phenomenon, we extract embeddings for 201 numbers (0 to 200) from each layer and compute their orderliness scores at every depth. This allows us to quantify how well numerical order is preserved throughout the model. By examining these trends, we can identify where in the network numerical structure is most clearly maintained and where it begins to dissolve into broader contextual associations. Understanding these transitions provides insight into how LLMs process numerical concepts across different levels of abstraction.

4.2 MAIN RESULTS

Orderliness Our results shown in Figure 2 reveal that all LLMs exhibit high orderliness in their shallow layers, confirming that the orderly mental number line is a universal phenomenon in LLMs. Notably, in most models, orderliness is nearly absent in layer 0 but rises sharply between layers 1 to 6, reaching its peak. This suggests that the earliest layers primarily handle raw token representations, while the initial transformation of numerical embeddings into a structured order occurs within the first few layers of processing. At their peak, orderliness scores across models range from 0.7 to 0.9, which represents a remarkably high level of spatial organization. For comparison, a baseline computed from 201 randomly arranged embeddings yields an orderliness score of 0.029 ± 0.008, demonstrating that LLMs encode numerical structure far beyond random chance. As layers deepen, orderliness gradually declines, likely due to the increasing influence of contextualization and next-token prediction, with some models exhibiting a two-phase decline that suggests distinct processing stages affecting numerical representations.

Visualization For the top-one-orderliness layer of each model, we apply t-SNE to reduce the dimensionality of their number embeddings to 2D and visualize the results in Figure 3. This visualization provides an intuitive view of the relative spatial arrangement of number embeddings and reveals structural details that the orderliness metric alone may not fully capture. For instance, while all models exhibit some degree of numerical organization, the way this order manifests varies. In LLaMA 3.1-8B, the number line appears smooth, with numbers distributed evenly along a continuous trajectory. In contrast, models such as Qwen 2.5-7B, Gemma 2-9B, LLaMA 2-7B, and Phi 3.5-4B display a different pattern—multiples of 10 tend to cluster together. Mistral-7B exhibits an intermediate behavior between these two extremes. This clustering of multiples of 10 introduces folds in the number line, disrupting its smoothness. Although the sequence of numbers between adjacent

[1] https://github.com/huggingface/transformers

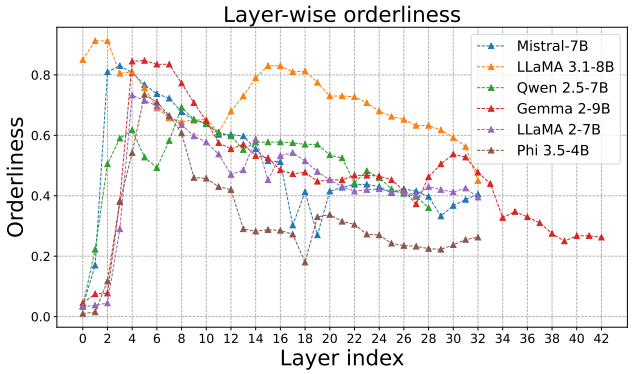


Figure 2: Layer-wise orderliness of number embeddings across six LLMs. For most models, orderliness is initially low in layer 0, which corresponds to the embedding layer before any transformer computations. However, it rapidly increases in the first few layers before reaching its peak. LLaMA 3.1 is an exception, exhibiting high orderliness from layer 0. Scores then decline as layer depth increases. A baseline of 0.029 ± 0.008 from randomly arranged embeddings is shown for reference.

multiples of 10 (e.g., 21 to 29) remains well-ordered, numbers immediately adjacent to a multiple of 10 are often positioned closer to numbers differing by ±10 rather than ±1. This phenomenon lowers the orderliness metric but reveals an alternative form of numerical organization—a hierarchical structuring of number representations.

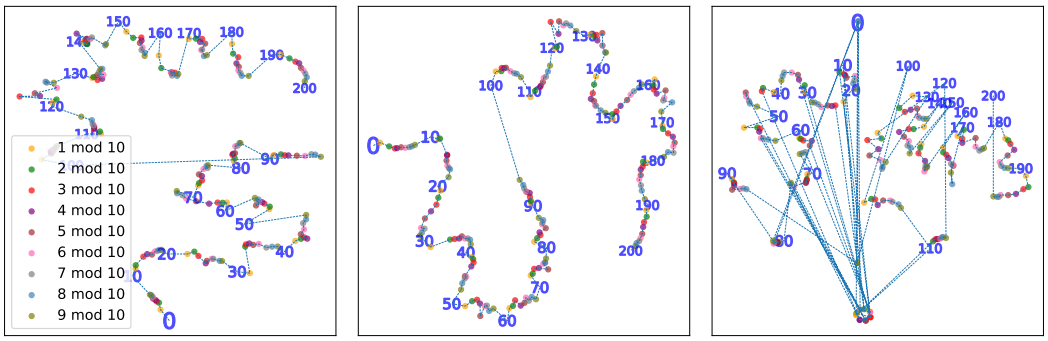
Overall, the visualization provides a complementary perspective on how LLMs encode numerical concepts in their hidden space. At a local level, the embeddings exhibit sequential order akin to a number line, while at a global level, they reveal a hierarchical structure. This analysis not only validates the effectiveness of the orderliness metric but also highlights its limitations. Moreover, in the context of representation alignment, it suggests that **LLMs may internalize numerical concepts in a way that is more intricate and cognitively aligned than a simple mental number line**.

5 Analysis

During the layer-wise orderliness computation, we observe a common phenomenon across all tested models: near the input layer, large language models exhibit a surprisingly high degree of numerical orderliness. However, as embeddings propagate through deeper layers, this orderliness undergoes a two-phase descent. What are the potential causes of these two declines? Given that LLaMA 3.1-8B is a widely used model and exhibits the most pronounced two-phase descent, we use it for further analysis. Specifically, we propose two hypotheses and conduct context-based experiments on LLaMA 3.1-8B to investigate the underlying mechanisms driving this phenomenon.

5.1 The First Decline in Orderliness

The first valley of orderliness for LLaMA 3.1-8B is observed within the interval from layer 5 to layer 15, as shown in Figure 4. We assume that this descent of orderliness is related to **contextualization**. In those layers, the model encodes more non-numerical information into embeddings of numbers, causing the first descent of orderliness. The following experiment supports this assumption. In Figure 4, we examine the orderliness of numbers in different contexts. Compared with context-free numbers, all context-based numbers have a lower orderliness from layer 5 to layer 20. After that, their average orderliness remains the same as the context-free numbers. This phenomenon indicates that any context brings more non-numerical information to the number embeddings in the middle layers than in any other layers, which implies that those layers focus on more non-numerical information about numbers. Please see Appendix A.1 for contexts used during the experiment.



(a) Mistral-7B, Layer 3, Orderliness = 0.83

(b) LLaMA 3.1-8B, Layer 1, Orderliness = 0.91

(c) Qwen 2.5-7B, Layer 8, Orderliness = 0.69

(d) Gemma 2-9B, Layer 5, Orderliness = 0.85

(e) LLaMA 2-7B, Layer 4, Orderliness = 0.73

(f) Phi 3.5-4B, Layer 5, Orderliness = 0.74

Figure 3: Top-1 orderly number embeddings for each model visualized in 2-D plane by t-SNE. The x-axis and y-axis of each subplot represent the two dimensions of t-SNE. Each circle in the plot corresponds to a number from 0 to 200, with its embedding projected into the t-SNE space. We connect the points sequentially from 0 to 200 using a blue dashed line. Since most points naturally form clusters of ten, we label only the multiples of ten in the plot. As indicated in the legend of (a), numbers that do not end in zero are grouped by their last digit and assigned different colors. For example, orange points correspond to numbers ending in 1 (e.g., 1, 11, 21, ..., 191).

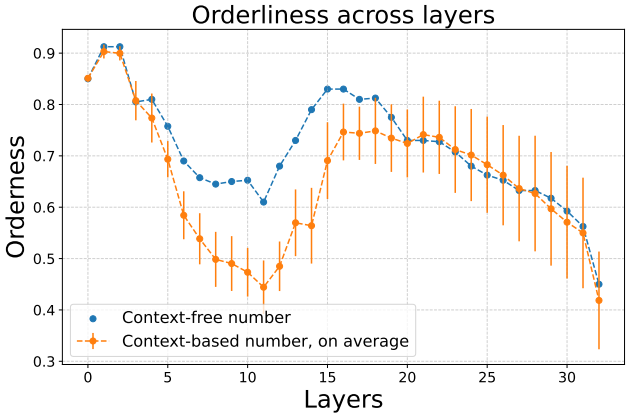


Figure 4: Orderliness affected by contexts. To see how context affects orderliness, we use context-based number prompts (orange) compared with the context-free number prompts (blue). We prepend fixed context texts to the numbers from 0 to 200 and calculate the layer-wise orderliness of these 201 number embeddings. By applying different context texts, we obtain orderliness for numbers in different contexts, and take the average on them for each layer, resulting in the orange points (mean) and bars (standard deviation).

5.2 The Second Decline in Orderliness

The second decline in orderliness occurs between layer 18 and the output layer, which we attribute to the influence of **next-token prediction**. We hypothesize that in these deeper layers, as number embeddings encode more information about their potential next tokens, their numerical relationships become less distinct, leading to a decrease in orderliness. To test this hypothesis, we design an experiment where we construct a sequence of 201 inputs such that the expected next-token predictions form an orderly number sequence (from 1 to 201). Instead of evaluating the orderliness of number tokens, we measure the orderliness of the last token in each input, assessing whether next-token prediction influences numerical structure in deep layers. For example, we construct inputs in the form of "%d + 1 [=]", where "%d" is the placeholder for the numbers from 0 to 200. We then measure the orderliness of the last token, denoted by "[ ]", which in this case is the "=" sign. Since large language models are expected to predict the next token after the equal sign as "%d + 1," the orderliness of the "=" token should increase in the final layers compared to context-free numerical inputs ("[%d]").

The results presented in Figure 5 confirm our hypothesis. Specifically, the green and red lines represent inputs of the form "%d + 1 [=]" and "%d + 1 is equal [to]," respectively. We observe a sudden increase in embedding orderliness of "=" and "to" at layer 23, suggesting that the orderliness of the last token embedding in deep layers is influenced by the expected next token. As a comparison, the orange line, representing inputs of the form "%d + [1]", does not exhibit an increase in orderliness at deep layers. This experiment provides a compelling explanation for the decline in orderliness observed in context-free number prompts at deep layers. Since isolated numbers lack contextual support, their next-token predictions are inherently uncertain. Meanwhile, deeper layers of the model prioritize next-token prediction, which reduces the orderliness of individual number embeddings. However, this decline does not indicate a loss of numerical concepts. Instead, it suggests that the model internalizes numerical representations through forward computation and utilizes them for next-token prediction.

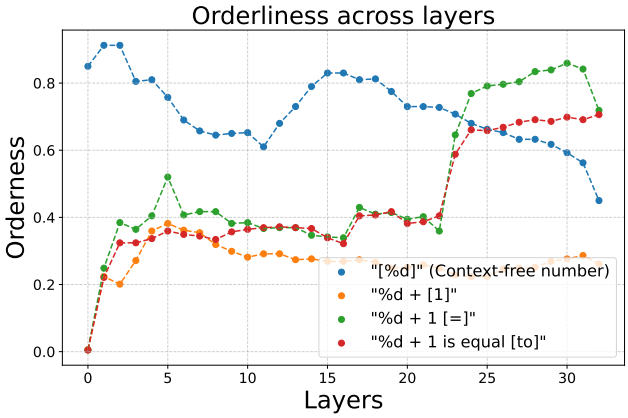


Figure 5: Orderliness affected by next-token prediction.

6 Conclusion

In this paper, we propose a new metric orderliness to quantify the structured arrangement of elements in high-dimensional space, validating that LLMs perceive numerical concepts in an orderly manner analogous to human cognition. By leveraging t-SNE visualization, we show that the high-dimensional embeddings representing numerical concepts exhibit a highly ordered spatial structure in the shallow layers of LLMs, akin to the human mental number line. We further find a two-phase decline in orderliness through the analysis of LLaMA 3.1 and identify that these two declines are closely associated with contextualization and next-token prediction. Our work establishes an alignment between numerical representations in LLMs and human numerical cognition, while also shedding light on how LLMs internalize and manipulate numerical concepts.

7 LIMITATION

Despite these insights, our study mainly focuses on medium-sized open-source models, leaving the exploration of larger open-source models and frontier black-box models unaddressed. Besides, our analysis only focuses on integer representations; further research can explore how LLMs encode more complex numerical concepts such as decimals, fractions, and rational numbers. Furthermore, our investigation of the two-phase descent phenomenon is conducted solely on LLaMA 3.1-8B, and its presence and underlying causes in other models remain to be explored. Moreover, we discover numerical concepts are highly ordered, but a deeper understanding of why this feature exists in the shallow layers of LLMs and how this structure contributes to numerical computation and reasoning remains an open question.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, 36, 2024.

Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

Martin Giurfa, Claire Marcout, Peter Hilpert, Catherine Thevenot, and Rosa Rugani. An insect brain organizes numbers on a left-to-right mental number line. *Proceedings of the National Academy of Sciences*, 119(44):e2203584119, 2022.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Per Christian Hansen. Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. *SIAM J. Sci. Stat. Comput.*, 11(3):503–518, May 1990. ISSN 0196-5204.

X. He, P. Guo, S. Li, et al. Non-symbolic and symbolic number lines are dissociated. *Cognitive Processing*, 22:475–486, 2021. doi: 10.1007/s10339-021-01019-4. URL https://doi.org/10.1007/s10339-021-01019-4.

Lesley Anne Honour. *Children's mental representation of number, their number line estimations and maths achievement: exploring the role of 3D mental rotation skills*. PhD thesis, University of Southampton, 2020.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Chien-Heng Lin. Developing mental number line games to improve young children's number knowledge and basic arithmetic skills. *Journal of Experimental Child Psychology*, 222:105479, 2022. ISSN 0022-0965. doi: https://doi.org/10.1016/j.jecp.2022.105479. URL https://www.sciencedirect.com/science/article/pii/S0022096522001084.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. Local interpretations for explainable natural language processing: A survey. *ACM Comput. Surv.*, 56(9), April 2024. ISSN 0360-0300. doi: 10.1145/3649450. URL https://doi.org/10.1145/3649450.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey, 2024. URL https://arxiv.org/abs/2209.11326.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8), December 2022. ISSN 0360-0300. doi: 10.1145/3546577. URL https://doi.org/10.1145/3546577.

Eysan Mehrbani and Mohammad Hossein Kahaei. Low-rank isomap algorithm, 2021. URL https://arxiv.org/abs/2103.04060.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Rosa Rugani, Giorgio Vallortigara, Konstantinos Priftis, and Lucia Regolin. Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, 347(6221):534–536, 2015. doi: 10.1126/science.aaa1379. URL https://www.science.org/doi/abs/10.1126/science.aaa1379.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. URL https://www.science.org/doi/abs/10.1126/science.290.5500.2319.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Giorgio Vallortigara. Comparative cognition of number and space: the case of geometry and of the mental number line. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373 (1740):20170120, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. Allennlp interpret: A framework for explaining predictions of nlp models. In *Conference on Empirical Methods in Natural Language Processing*, 2019a. URL https://api.semanticscholar.org/CorpusID:202712654.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*, 2019b.

John Wallis. A treatise of algebra, both historical and practical. *Philosophical Transactions of the Royal Society of London*, 15(173):1095–1106, 1685. doi: 10.1098/rstl.1685.0053. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1685.0053.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. Interpreting and improving large language models in arithmetic calculation. *arXiv preprint arXiv:2409.01659*, 2024.

A APPENDIX

A.1 CONTEXTS

In our study on the first decline in orderliness (section 5.1), we compare how numerical orderliness changes under different contextual conditions. Specifically, we examine the impact of various context settings on numerical orderliness. The specific contexts used in our experiments are listed in the table below.

Table 1: Prompts used in different contexts. [%d] is the placeholder for numbers from 0 to 200.

Context	**Prompt**
No Context	"[%d]"
Card Number	"The digit of this card number starts with [%d]"
Share Money	"How much money does each person get on average if three people share [%d]"
Total Donation	"Calculate the total amount of funds raised. A total of N individuals participated in the donation, with each person contributing an average of [%d]"
Left Apple	"How many apples are left if 3 apples are stolen, assuming we originally had [%d]"
1 plus	"1 + [%d]"
Focus Difference	"Ignore all properties of numbers; focus only on how they differ from each other. Begin counting: [%d]"
Forget Difference	"Imagine numbers as equal entities, no size difference, only pure existence: [%d]"
Meditation	"Close your eyes, clear your mind, breathe deeply, immerse yourself in stillness, and count: [%d]"

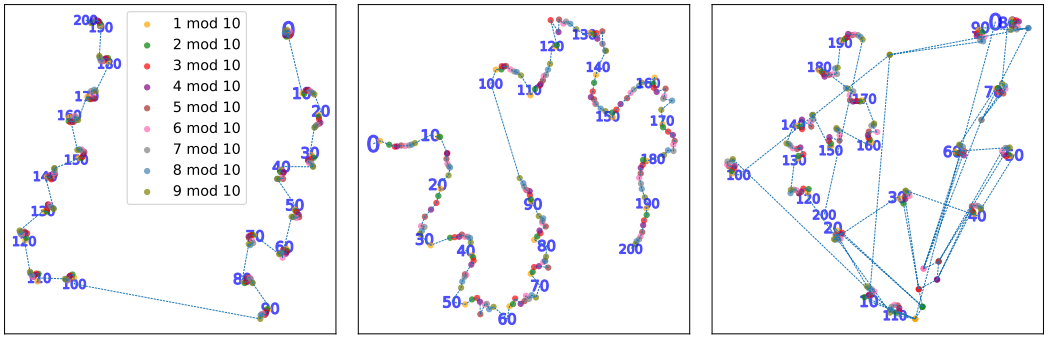
A.2 DETAILS OF EMBEDDINGS EXTRACTION

LLMs typically process input as a sequence of characters, which first pass through a tokenizer that segments them into subword tokens based on the model's vocabulary. When we input numbers from 0 to 200 into LLMs, they are tokenized into different tokens. The embeddings corresponding to each token at specific positions can be precisely located, and their positions remain unchanged throughout the layer-wise transformations of the model. This positional stability forms the foundation of our embedding extraction method.

It is worth noting that different models do not always use the same vocabulary, meaning that the same number may be tokenized differently across models. Some models may split a single number into multiple tokens, resulting in embeddings distributed across multiple positions. In such cases, it is necessary to apply specific strategies to aggregate these embeddings, commonly referred to as pooling strategies. A common approach in previous studies is to use pooling strategies to obtain a single representation for a word composed of multiple tokens. The most frequently used methods include average pooling, where the embeddings of all tokens are averaged, and last-token pooling, where only the embedding of the last token is used as the word representation.

In the experiments presented in this study, we primarily use last-token pooling for embedding extraction. Additionally, we provide results using the average pooling method in Figure 7, which also exhibits the double descent phenomenon on orderliness.

Notably, different embedding pooling methods have a significant impact on the visualization. Figure 6 presents the visualization of the highest-orderliness layer using average pooling for embeddings. Compared to the last-token pooling method, numbers appear more tightly clustered around multiples of 10. The only exception is LLaMA 3.1-8B, which remains unaffected due to its tokenizer represents all numbers from 0 to 200 as single tokens, making average pooling and last-token pooling yield identical results.



(a) Mistral-7B, Layer 2, Orderliness = 0.86

(b) LLaMA 3.1-8B, Layer 1, Orderliness = 0.91

(c) Qwen 2.5-7B, Layer 3, Orderliness = 0.73

(d) Gemma 2-9B, Layer 5, Orderliness = 0.86

(e) LLaMA 2-7B, Layer 4, Orderliness = 0.82

(f) Phi 3.5-4B, Layer 3, Orderliness = 0.89

Figure 6: Top-1 orderliness number embeddings for each model visualized in a 2D plane using t-SNE. All settings are identical to Figure 3, except that embeddings are pooled using the **average** method

A.3 Results of Different Distance Function

In section 3.2, we introduced our algorithm for calculating orderliness, where one of the key components is the distance function f_d. The experiments presented in the main text are all based on Euclidean distance. To further validate the robustness of our findings, we also provide results using an alternative distance function: cosine similarity in Figure 8 and Figure 9.

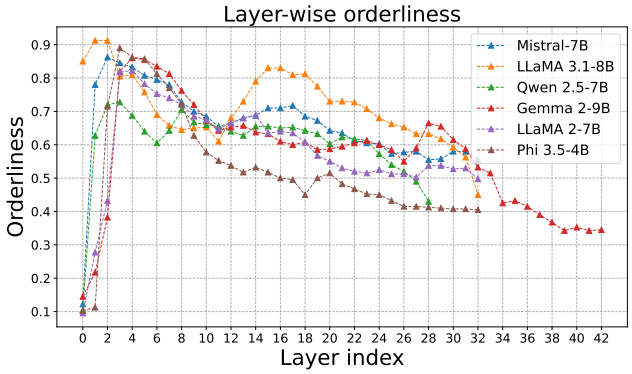


Figure 7: Average pooling $\times$ Euclidean distance

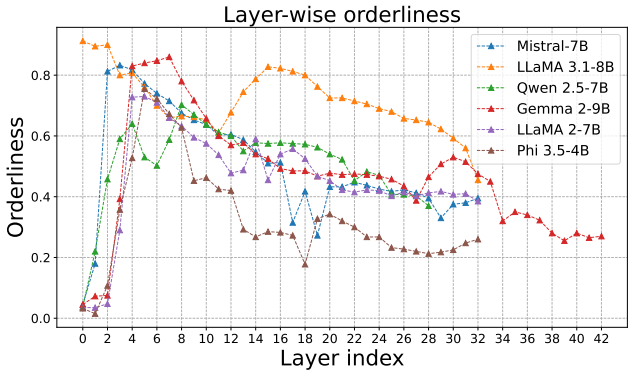


Figure 8: Last token pooling $\times$ Cosine similarity

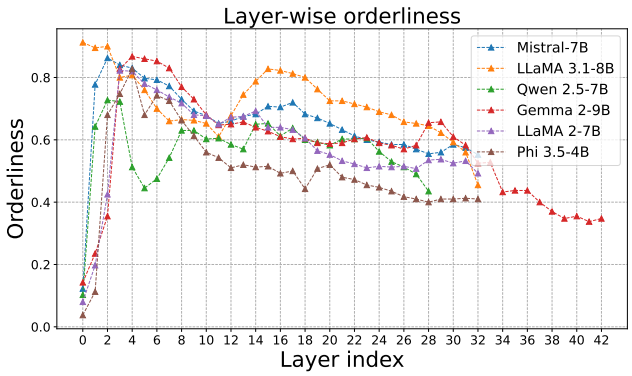


Figure 9: Average pooling $\times$ Cosine similarity