Pretraining Numerical Frequency and Number-Line in Language Models

Mohammed Ibrahim Awad [1 *] **Ahmed Elshehaby** [1 *] **Hilal AlQuabeh** [1 †] **Velibor Bojkovic** [1 †]

Abstract

Large language models exhibit compressed, non-uniform internal representations of numerical magnitude, but the pretraining factors associated with this geometry remain unclear. We study whether corpus-level integer statistics are related to the learned number-line geometry of these models. For four documented pretraining corpora, we count integers in $[0 : 10{,}000]$ and fit a magnitude-frequency power law, $\text{count}(N) \propto N^{\alpha}$, where more negative α indicates steeper decay and less exposure to large magnitudes. For nine corresponding base models, we extract hidden states for numerical prompts, project them onto a one-dimensional number line with PCA, and estimate a scaling factor β, where smaller β indicates stronger compression. We first show that β is behaviorally meaningful: models with less compressed number-line geometry achieve higher likelihood-based number-comparison accuracy. We then find that flatter integer-frequency distributions, corresponding to less negative α, are associated with larger β, providing correlational evidence that pretraining integer statistics are reflected in LLM number representations.

1. Introduction

Numerical reasoning remains an important challenge for Large Language Models (LLMs), since mathematical tasks often depend on how models internally represent, compare, and manipulate numbers. This makes it important to understand whether numerical values are encoded internally along an implicit number line that may shape reasoning on mathematical tasks.

[*]Equal contribution [†]Equal supervision. [1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. Correspondence to: Hilal AlQuabeh <hilal.alquabeh@mbzuai.ac.ae>, Velibor Bojkovic <velibor.bojkovic@mbzuai.ac.ae>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

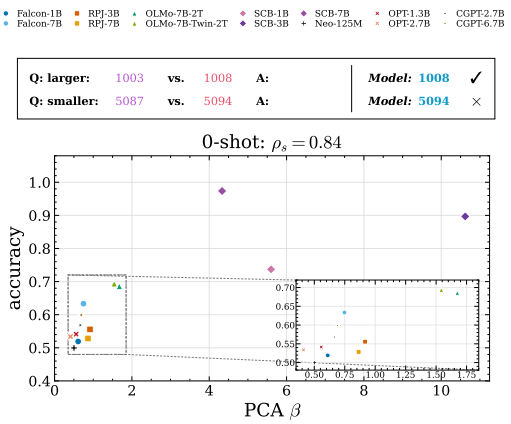

Figure 1. Compression factor β vs. 0-shot numerical comparison accuracy. Each point is a model; larger β means less compressed representations and tends to be associated with higher accuracy, with positive correlation ($\rho_s = 0.84$). Prompt examples show the comparison format; see Section 4 for details.

The linear representation hypothesis (Park et al., 2024) states that high-level concepts are represented internally in LLMs as approximately linear directions or subspaces, motivating the question of whether numerical magnitude is also organized in a simple linear way inside models. Supporting this possibility, Zhu et al. (2025) showed that numerical values can be recovered from LLM hidden states using linear probes. However, being linearly recoverable does not mean that the internal number line is uniformly spaced. Recent findings suggest that LLMs encode numerical magnitude in a compressed, non-linear manner, similar to the logarithmic mental number line observed in human numerical cognition (Shah et al., 2023; AlquBoj et al., 2025). More specifically, AlquBoj et al. (2025) showed that internal number representations in LLMs are not uniformly spaced; rather, the spacing changes as numerical magnitude increases. They quantify this compression using the Scaling Rate Index β. (Smaller β means stronger logarithmic-like compression.)

While prior work identifies compressed number-line geometry in LLMs, two questions remain open: **1- What drives different compression rates across models. 2- Whether this geometry has practical relevance for numerical reasoning**. We address the latter by relating β to model performance on simple multiple-choice numerical reasoning questions. We find that models with larger β (less compressed and more evenly spaced internal number line), tend

to achieve higher accuracy, providing evidence that number-line geometry is related to numerical reasoning ability.

To answer what factors shape β across models, we investigate pre-training number frequency as one possible factor. This motivation is inspired by human numerical development, where exposure and formal training affect number-line judgments, and by recent evidence that pre-training frequency influences representation geometry in LLMs (Merullo et al., 2025). To test this hypothesis, we connect corpus-level number statistics to model-level representation geometry: we estimate the frequency decay term α from integer counts in pre-training data using $f_{\text{count}}(N) \propto N^{\alpha}$, and compare it with the number-line compression factor β, estimated from one-dimensional projections of hidden-state number representations. Across nine LLMs trained on known datasets, we find a positive correlation between α and β: sharper frequency decay in the data, i.e. more negative α, is associated with stronger internal number-line compression, i.e. smaller β.

In short, we summarize our contributions as follows:

- We identify an empirical trend showing that LLM accuracy on numerical reasoning tasks increases as the number-line compression factor β increases.
- We extract empirical frequency distributions of integers $x \in [0, 10000]$ from four distinct open-source pretraining datasets and fit a power-law model based on magnitude, introducing the decay term α, where more negative values indicate that the integer frequency decreases more rapidly with numerical value.
- We study 9 LLMs trained only on their pretraining corpus and find a positive correlation between β and α, providing evidence that the integer-frequency distribution is related to the geometry of internal representation of numbers.

2. Related Work

Many previous studies have explored whether language models can understand and reason about numbers. Early work on numeracy in LLMs showed that while models can capture some numerical information, they still struggle with arithmetic reasoning (Spithourakis & Riedel, 2018; Wallace et al., 2019). More recent work has shown that LLMs exhibit human-like numerical-comparison effects, suggesting that they may capture behavioral patterns of numerical magnitude similar to human numerical cognition (Shah et al., 2023). Closely related to our motivation, Razeghi et al. (2022) showed that LLMs perform better on numerical reasoning tasks when relevant terms occur more frequently in pretraining data, suggesting that numerical ability may be influenced by the distribution of numbers during pretraining.

Beyond task-based performance, other works have studied how numerical values are represented in the hidden layers of LLMs. Some studies show that internally encoded number values can be recovered from hidden states using linear probes (Zhu et al., 2025), but this does not mean that the internal number line is uniformly spaced. Recent studies also show that LLMs may represent numbers using base-10 patterns and string-like information, rather than treating all numbers as pure numerical magnitude (Levy & Geva, 2025; Marjieh et al., 2025). Most closely related to our work, AlquBoj et al. (2025) showed that LLMs encode numerical representations in a compressed, non-uniform manner, quantified using the Scaling Rate Index β. In parallel, work on pre-training frequency suggests that corpus statistics can shape LLMs' internal representations (Merullo et al., 2025).

Inspired by these directions, we study whether integer distributions in pretraining datasets are related to compressed number-line geometry. Since natural language datasets are highly non-uniform and often follow Zipf-like patterns (Zipf, 1949; Newman, 2005), it is natural to expect number frequencies to follow a similar heavy-tailed structure. However, classical Zipf's law studies frequency as a function of rank, while our question requires studying frequency as a function of the numerical magnitude. Therefore, instead of fitting the classical Zipf rank-frequency law, we fit a magnitude-based power law and estimate the decay term α, which captures how rapidly number frequency decreases as numerical magnitude increases (Clauset et al., 2009). We then investigate the relationship between α and the compression factor β.

3. Background

Previous work has introduced approaches to evaluate compression rate and analyze numerical representations in LLMs using dimensionality reduction techniques and linear probing (AlquBoj et al., 2025). The broader idea is investigating whether LLMs encode the number-line in an intuition similar to humans–logarithmic, sublogarithmic, and super-logarithmic.

The investigation starts by analyzing hidden representations across model layers. Each input number x is associated with an internal representation $f(x) \in \mathbb{R}^d$. To examine whether numerical values align along an implicit number-line, we project these hidden representations onto a one-dimensional space using PCA, $T : \mathbb{R}^d \to \mathbb{R}$, and denote the projected representation as

$$f_{\text{LLM}}(x) := T(f(x)). \tag{1}$$

This one-dimensional projection allows us to analyze the geometric structure of numerical magnitudes, including whether the representation preserves numerical order and whether the spacing is uniform or compressed. Distances between projected representations can be measured using

Euclidean distance, while order preservation is measured using the absolute Spearman rank correlation $|\rho|$, treating both increasing and decreasing monotonic directions interchangeably.

To quantify number-line compression, following AlquBoj et al. (2025), we consider inputs of the form $x_i = 10^i$ and define $y_i = f_{\text{LLM}}(x_i)$. We then examine how the differences between consecutive representations evolve:

$$y_{i+1} - y_i = A \cdot \beta^i. \tag{2}$$

The scaling parameter β is fitted to the observed differences by minimizing the least-squares objective:

$$\min_{A,\beta} \sum_{i=1}^{n} \left((y_{i+1} - y_i) - A\beta^i\right)^2. \tag{3}$$

This objective models how the spacing between numerical representations changes across orders of magnitude. Since $x_i = 10^i$, the index i corresponds to the logarithmic scale of the input. Therefore,

$$\beta^i = \beta^{\log_{10}(x_i)} = x_i^{\log_{10}(\beta)}. \tag{4}$$

This gives an intuitive interpretation of different scaling regimes. If $\beta = 1$, the differences remain approximately constant across powers of ten, producing logarithmic behavior. If $\beta < 1$, the differences decrease with magnitude, indicating sub-logarithmic compression where larger numbers become increasingly packed. If $1 < \beta < 10$, the mapping grows faster than logarithmic but slower than linear, corresponding to super-logarithmic but sublinear behavior. If $\beta = 10$, then

$$\beta^i = 10^i = x_i,$$

which gives a linear mapping. Finally, if $\beta > 10$, the growth becomes faster than linear, indicating superlinear behavior where distances between larger numbers expand at an increasing rate. Thus, smaller β means stronger compression, while larger β means a less compressed or more expanded internal number-line.

4. LLMs Numerical Reasoning Ability and β

Following up on LLMs conceiving of the number-line, we evaluate whether the compression factor β is related to their ability to compare numerical magnitudes. The prompts are comparison-based: given two numbers $\{a, b\}$, the model selects either $\min\{a, b\}$ or $\max\{a, b\}$, depending on whether the query asks for the smaller or larger value. This directly evaluates whether number-line encoding impacts numerical reasoning. Intuitively, if the internal number-line is more compressed, comparing larger numbers should become harder as their representations become closer.

To test this, we construct a synthetic pairwise number-comparison dataset designed to control for numerical magnitude, absolute gap size, answer position, and query direction. Firstly, we partition comparison pairs into four magnitude regimes corresponding to powers of ten: $\mathcal{G}_1 = [10, 99]$, $\mathcal{G}_2 = [100, 999]$, $\mathcal{G}_3 = [1000, 9999]$, and $\mathcal{G}_4 = [10000, 99999]$. Within each group, we generate fixed-gap pairs $(a, b) = (n, n + d)$ with two gap regimes: $SG_{small\_gap} : d \in \{1, 2, 3, 4, 5\}$ and $MG_{medium\_gap} : d \in \{6, 7, 8, 9, 10\}$.

For each group and gap value, we sample 32 starting values using a fixed random seed, include both input orderings (a, b) and (b, a), and use both prompt directions ("larger" and "smaller") to prevent positional or heuristic biases. We evaluate each model in 0 – 4 shot settings using multiple exemplar sets, giving 15,360 comparisons per model-shot condition. We report the main 0-shot results in Figure 1, and provide the 1–4 shot results in Section E.3.

The evaluation was carried out on the 9 open-source language models summarized in Table 1. For the few-shot comparison analysis, we also include GPT-Neo-125M (Black et al., 2021), OPT-1.3B, OPT-2.7B (Zhang et al., 2022), Cerebras-GPT-2.7B, and Cerebras-GPT-6.7B (Dey et al., 2023); these models are not included in the pretraining corpus analysis because their exact pretraining datasets are not available in the same directly matched form.

Rather than relying on free-form generation, we adopt a likelihood evaluation. Given a prompt q and two candidate outputs c_a and c_b, we compute the length-normalized log-likelihood and select the candidate with the highest score:

$$\begin{aligned} s(c|q) &= \frac{1}{|c|} \sum_{t=1}^{|c|} \log p(c_t|q, c_{<t}) \\ \hat{c} &= \arg \max_{c \in \{c_a, c_b\}} s(c|q). \end{aligned} \tag{5}$$

This ensures that the evaluation reflects the model's internal preference between numerical magnitudes and not decoding artifacts. We observe a strong positive correlation between model performance and β across shot settings: models with larger β consistently achieve higher comparison accuracy, indicating that less compressed numerical representations correspond to better discrimination between numerical magnitudes. This motivates us to study what factors give rise to different β values, and whether pre-training number frequency is one such factor.

5. Power-Law Structure of Pretraining Data

To understand the underlying influence of the training dataset on the compression rate of number-lines, access to the pretraining datasets of the models was crucial. Therefore, we selected 9 open-source base models trained on 4

distinct publicly accessible pretraining datasets, summarized in Table 1. These models were trained only on their respective datasets and were not finetuned, instruction-tuned, or enhanced with additional reasoning methods, reducing the possibility that β was altered by unexplainable factors while studying its correlation with the power-law exponent α.

Datasets	Models	Desc
Falcon-RefinedWeb (Penedo et al., 2023)	falcon-rw-7b falcon-rw-1b	Filtered web pages from CommonCrawl
RedPajama-Data-1T (Weber et al., 2024)	RedPajama-INCITE-Base-3B-v1 RedPajama-INCITE-7B-Base	Mixed text corpus with web, books, Wikipedia, and code
Dolma (Soldaini et al., 2024)	OLMo-7B OLMo-7B-Twin-2T (Groeneveld et al., 2024)	Large open corpus with web, books, papers, wiki, and code
The Stack (Kocetkov et al., 2023)	starcoderbase-1b starcoderbase-3b starcoderbase-7b (Li et al., 2023)	Programming code dataset from public repositories

Table 1. Pre-training datasets and model families. All models are open-source and trained exclusively on their respective datasets.

5.1. Integer-Frequency Analysis of Pretraining Corpora

To test whether corpus statistics may explain the geometry of learned number representations, we measure how often integers occur in each pretraining corpus. For each dataset, we perform a full pass over the text, count integer mentions in the range $0 \leq N \leq 10000$, extract digit strings corresponding to non-negative integers, normalize each matched integer to its numeric value, e.g., 001 becomes 1, and aggregate the counts into $\mathrm{count}(N)$. This range matches the numerical scale used in our representation analysis.

We model this empirical integer distribution as an approximately Zipfian power law with numerical magnitude, $\mathrm{count}(N) \propto N^{\alpha}$. Here, N denotes the integer value itself, not the frequency rank. Thus, α is a magnitude-frequency exponent, where smaller values of α (more negative values) indicate that frequency decays more rapidly as numerical magnitude increases. Equivalently, we estimate α by fitting $\log \mathrm{count}(N) = c + \alpha \log N + \epsilon_N$ for integers with nonzero counts. This exponent is used in our corpus-frequency hypothesis because both α and the representation compression factor β are defined with respect to numerical magnitude.

Dataset	Count				Magnitude Fit	
	$[0, 10)$	$[10, 10^2)$	$[10^2, 10^3)$	$[10^3, 10^4]$	α	R^2
Stack v1.2	51.530	23.656	13.420	10.050	-1.18	0.91
Dolma v1.5 sample	18.879	11.297	4.701	5.725	-1.30	0.81
RedPajama-Data-1T	9.849	8.402	3.059	6.741	-1.40	0.68
Falcon RefinedWeb	4.633	4.017	1.426	2.222	-1.42	0.79

Table 2. Integer-frequency mass by magnitude bin, reported in billions. Final columns show the fitted exponent α from $\mathrm{count}(N) \propto N^{\alpha}$ and log-log fit R^2; smaller α means steeper decay with numerical magnitude.

5.2. Number-Line Geometry and Integer-Frequency

We next ask whether corpus-level integer statistics are reflected in model representations. For each model, we extract hidden states for numerical inputs and compute PCA-based number-line statistics, illustrated in Figure 2. Most models exhibit a clear monotonic number-line structure: Falcon, RedPajama, and OLMo models reach $\rho \approx 0.92$–0.93, while StarCoderBase models have slightly lower monotonicity but much larger β (expanded spacing at larger magnitudes).

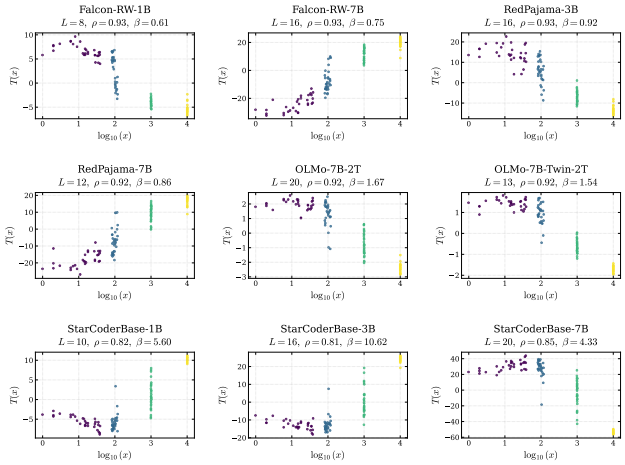

Figure 2. 1-D PCA projections at the layer with highest explained variance. $T(x)$ is the projected hidden representation and the x-axis is $\log_{10}(x)$. The monotonic pattern shows an internal number-line: ρ measures how well the projected order matches the true numerical order, while β measures scaling, where smaller β means stronger compression and larger β means more expanded spacing.

Comparing Table 2 and Figure 2, we observe a monotonic corpus–model trend. Stack has the flattest integer-frequency decay ($\alpha = -1.18$) and its StarCoderBase models have the largest mean β (6.85). Dolma is intermediate ($\alpha = -1.30$, $\beta \approx 1.6$), while RedPajama and Falcon RefinedWeb have steeper decay ($\alpha = -1.40, -1.42$) and smaller β. Thus, less negative α values correspond to larger β, supporting the hypothesis that flatter numerical frequency distributions are associated with less compressed number-line representations; The aggregate visualization is provided in Figure 7.

6. Conclusion & Limitations

We studied how LLM number-line geometry relates to integer-frequency structure in pretraining data. Across model families, numerical representations form a measurable internal number line whose spacing varies systematically: flatter integer-frequency decay corresponds to larger β, while steeper decay corresponds to stronger compression. The exponent α from $\mathrm{count}(N) \propto N^{\alpha}$ predicts variation in learned number-line geometry, and our number-comparison probe shows that less compressed representations are associated with higher accuracy. At the same time, our analysis depends on access to documented pretraining data; for example, Falcon RefinedWeb serves as a proxy for Falcon's full training distribution.

References

AlquBoj, H. V., AlQuabeh, H., Bojkovic, V., Hiraoka, T., El-Shangiti, A. O., Nwadike, M., and Inui, K. Number representations in llms: A computational parallel to human perception, 2025. URL https://arxiv.org/abs/2502.16147.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review*, 51 (4):661–703, 2009. doi: 10.1137/070710111. URL https://doi.org/10.1137/070710111.

Davies, A. O., Nzoyem, R., Ajmeri, N., and Silva Filho, T. M. Language models do not embed numbers continuously. *arXiv preprint arXiv:2510.08009*, 2025. doi: 10.48550/arXiv.2510.08009. URL https://arxiv.org/abs/2510.08009.

Dehaene, S. The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, 7(4):145–147, 2003. doi: 10.1016/S1364-6613(03)00055-X. URL https://doi.org/10.1016/S1364-6613(03)00055-X.

Dehaene, S., Izard, V., Spelke, E., and Pica, P. Log or linear? distinct intuitions of the number scale in western and amazonian indigene cultures. *Science*, 320(5880): 1217–1220, 2008. doi: 10.1126/science.1156540. URL https://doi.org/10.1126/science.1156540.

Dey, N., Gosal, G., Zhiming, Chen, Khachane, H., Marshall, W., Pathria, R., Tom, M., and Hestness, J. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster, 2023. URL https://arxiv.org/abs/2304.03208.

El-Shangiti, A. O., Hiraoka, T., AlQuabeh, H., Heinzerling, B., and Inui, K. The geometry of numerical reasoning: Language models compare numeric properties in linear subspaces. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–561, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.47. URL https://aclanthology.org/2025.naacl-short.47/.

Fechner, G. T. *Elemente der Psychophysik*. Breitkopf und Härtel, Leipzig, 1860. doi: 10.3931/e-rara-10879. URL https://doi.org/10.3931/e-rara-10879.

Fritz, A., Ehlert, A., and Balzer, L. Development of mathematical concepts as basis for an elaborated mathematical understanding. *South African Journal of Childhood Education*, 3(1):38–67, 2013.

Groeneveld, D., Beltagy, I., Walsh, E., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N., and Hajishirzi, H. OLMo: Accelerating the science of language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Gurnee, W., Ameisen, E., Kauvar, I., Tarng, J., Pearce, A., Olah, C., and Batson, J. When models manipulate manifolds: The geometry of a counting task, 2026. URL https://arxiv.org/abs/2601.04480.

Hasani, H., Banayeeanzade, M., Nafisi, A., Mohammadian, S., Askari, F., Bagherian, M., Izadi, A., and Baghshah, M. S. Mechanistic interpretability of large-scale counting in llms through a system-2 strategy, 2026. URL https://arxiv.org/abs/2601.02989.

Heinzerling, B. and Inui, K. Monotonic representation of numeric attributes in language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 175–195, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.18. URL https://aclanthology.org/2024.acl-short.18/.

Kocetkov, D., Li, R., allal, L. B., LI, J., Mou, C., Jernite, Y., Mitchell, M., Ferrandis, C. M., Hughes, S., Wolf, T., Bahdanau, D., Werra, L. V., and de Vries, H. The stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=pxpbTdUEpD.

Levy, A. A. and Geva, M. Language models encode numbers using digit representations in base 10. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 385–395, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.33. URL https://aclanthology.org/2025.naacl-short.33/.

Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J., Liu, Q., Zheltonozhskii, E., Zhuo, T. Y., Wang, T., Dehaene, O., Lamy-Poirier, J., Monteiro, J., Gontier, N., Yee, M.-H., Umapathi, L. K., Zhu, J., Lipkin, B., Oblokulov, M., Wang, Z., Murthy, R., Stillerman, J. T., Patel, S. S., Abulkhanov, D., Zocca, M., Dey, M., Zhang, Z., Bhattacharyya, U., Yu, W., Luccioni, S., Villegas, P., Zhdanov, F., Lee, T., Timor, N., Ding, J., Schlesinger, C. S., Schoelkopf, H., Ebert, J., Dao, T., Mishra, M., Gu, A., Anderson, C. J., Dolan-Gavitt, B., Contractor, D., Reddy, S., Fried, D., Bahdanau, D., Jernite, Y., Ferrandis, C. M., Hughes, S., Wolf, T., Guha, A., Von Werra, L., and de Vries, H. StarCoder: May the Source Be with You! *Transactions on Machine Learning Research*, 2023. URL https://mlanthology.org/tmlr/2023/li2023tmlr-starcoder/.

Marjieh, R., Veselovsky, V., Griffiths, T. L., and Sucholutsky, I. What is a number, that a large language model may know it?, 2025. URL https://arxiv.org/abs/2502.01540.

Merullo, J., Smith, N. A., Wiegreffe, S., and Elazar, Y. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EDoD3DgivF.

Newman, M. E. J. Power laws, pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005. doi: 10.1080/00107510500052444. URL https://doi.org/10.1080/00107510500052444.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Park, S., Ryu, S., and Choi, E. Do language models understand measurements? In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1782–1792, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.128. URL https://aclanthology.org/2022.findings-emnlp.128/.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79155–79172. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fa3ed726cc5073b9c31e3e49a807789c-Paper-Datasets_and_Benchmarks.pdf.

Petrak, D., Moosavi, N. S., and Gurevych, I. Arithmetic-based pretraining improving numeracy of pretrained language models. In Palmer, A. and Camacho-collados, J. (eds.), *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pp. 477–493, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.42. URL https://aclanthology.org/2023.starsem-1.42/.

Razeghi, Y., Logan IV, R. L., Gardner, M., and Singh, S. Impact of pretraining term frequencies on few-shot numerical reasoning. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.59. URL https://aclanthology.org/2022.findings-emnlp.59/.

Shah, R., Marupudi, V., Koenen, R., Bhardwaj, K., and Varma, S. Numeric magnitude comparison effects in large language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6147–6161, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.383. URL https://aclanthology.org/2023.findings-acl.383/.

Siegler, R. S. and Opfer, J. E. The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, 14(3):237–243,

2003. doi: 10.1111/1467-9280.02438. URL https://doi.org/10.1111/1467-9280.02438.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Ravichander, A., Richardson, K., Shen, Z., Strubell, E., Subramani, N., Tafjord, O., Walsh, E., Zettlemoyer, L., Smith, N., Hajishirzi, H., Beltagy, I., Groeneveld, D., Dodge, J., and Lo, K. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Spithourakis, G. and Riedel, S. Numeracy for language models: Evaluating and improving their ability to predict numbers. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2104–2115, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1196. URL https://aclanthology.org/P18-1196/.

Tiblias, F., Bigoulaeva, I., Niu, J., Balloccu, S., and Gurevych, I. Hypothesis-driven feature manifold analysis in LLMs via supervised multi-dimensional scaling. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=vCKZ40YYPr.

Wallace, E., Wang, Y., Li, S., Singh, S., and Gardner, M. Do NLP models know numbers? probing numeracy in embeddings. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5307–5315, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1534. URL https://aclanthology.org/D19-1534/.

Wang, Z., Jiang, Y., Zhou, R., Zhang, B., Zhang, F., Xu, Z., Zhang, Y., and Wang, J. Drivecode: Domain specific numerical encoding for llm-based autonomous driving, 2026. URL https://arxiv.org/abs/2603.00919.

Weber, M., Fu, D. Y., Anthony, Q., Oren, Y., Adams, S., Alexandrov, A., Lyu, X., Nguyen, H., Yao, X., Adams, V., Athiwaratkun, B., Chalamala, R., Chen, K., Ryabinin, M., Dao, T., Liang, P., Ré, C., Rish, I., and Zhang, C. Redpajama: an open dataset for training large language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 116462–116492. Curran Associates, Inc., 2024. doi: 10.52202/079017-3697. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d34497330b1fd6530f7afd86d0df9f76-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Zhang, X., Ramachandran, D., Tenney, I., Elazar, Y., and Roth, D. Do language embeddings capture scales? In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4889–4896, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.439. URL https://aclanthology.org/2020.findings-emnlp.439/.

Zhou, Z., Wang, J., Lin, D., and Chen, K. Scaling behavior for large language models regarding numeral systems: An example using pythia. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3806–3820, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.218. URL https://aclanthology.org/2024.findings-emnlp.218/.

Zhu, F., Dai, D., and Sui, Z. Language models encode the value of numbers linearly. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 693–709, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.47/.

Zipf, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA, 1949. URL https://archive.org/details/in.ernet.dli.2015.90211.

A. Overview of the Proposed Pipeline

Figure 3 summarizes the complete pipeline used in this work. First, we estimate the magnitude-frequency exponent α by extracting integers from pretraining corpora, normalizing them, and fitting a magnitude-based power law to $\mathrm{count}(N)$. Second, we estimate the compression factor β by extracting hidden representations from transformer layers, projecting them with PCA, and analyzing spacing across numerical magnitudes. Finally, we compare corpus-level statistics α with representation geometry β across the matched model families.

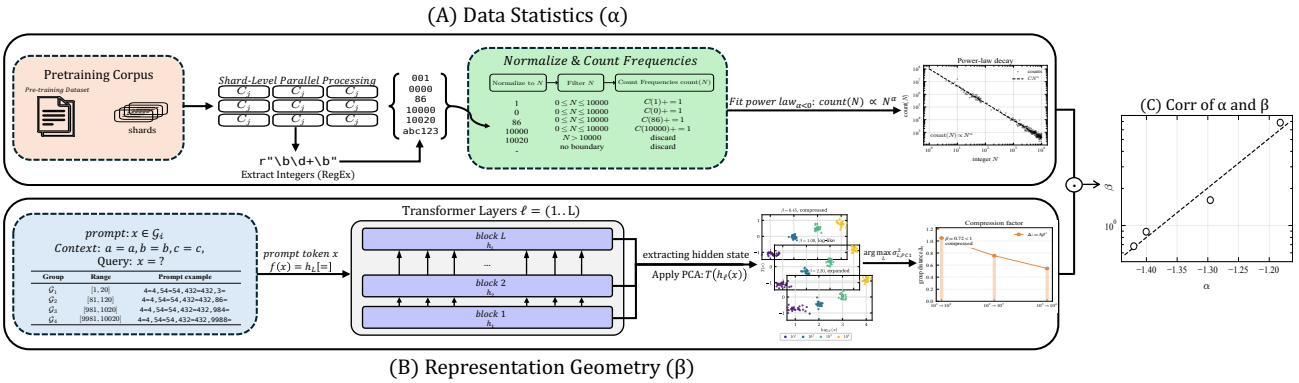

Figure 3. Overview of the proposed pipeline. (A) We estimate the magnitude-frequency exponent α by extracting integers from pretraining corpora, normalizing them, and fitting a power law to $\mathrm{count}(N)$. (B) We obtain the compression factor β by extracting hidden representations from transformer layers, projecting them via PCA, and analyzing spacing across numerical magnitudes. (C) We then analyze the relationship between corpus-level statistics α and representation geometry β across models.

B. Logarithmic Mental Line Hypothesis

The logarithmic mental line hypothesis proposes that numerical magnitudes are not always represented with uniform spacing. Instead, smaller numbers are represented with relatively larger separations, while larger numbers become increasingly compressed. This produces an approximately logarithmic order of magnitude, where equal ratios are represented more similarly than equal absolute differences. Such compression has been observed in human numerical cognition, especially in approximate estimation and early numerical development, and provides a useful reference point for studying whether LLMs organize numbers in a smaller geometric structure.

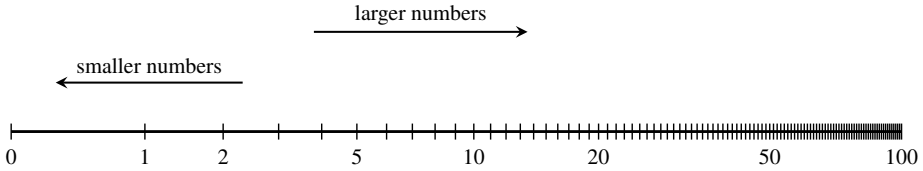

Figure 4. Logarithmically compressed mental number line. Image source (Fritz et al., 2013).

C. Additional Corpus-Frequency Diagnostics

In addition to estimating α, we compare the empirical distribution to a Zipf's law baseline:

$$q(N) = \frac{(N+1)^{-1}}{\sum_{k=0}^{10000}(k+1)^{-1}}.$$

We use $N+1$ to ensure the distribution is defined at $N=0$. Let

$$p(N) = \frac{\mathrm{count}(N)}{\sum_{k=0}^{10000}\mathrm{count}(k)}$$

be the normalized empirical distribution. To measure how different the empirical distribution is from this baseline, we compute the KL divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{N=0}^{10000} p(N) \log \frac{p(N)}{q(N)}. \tag{6}$$

This value serves as a secondary measure, indicating how closely the observed data follows a Zipf-like distribution.

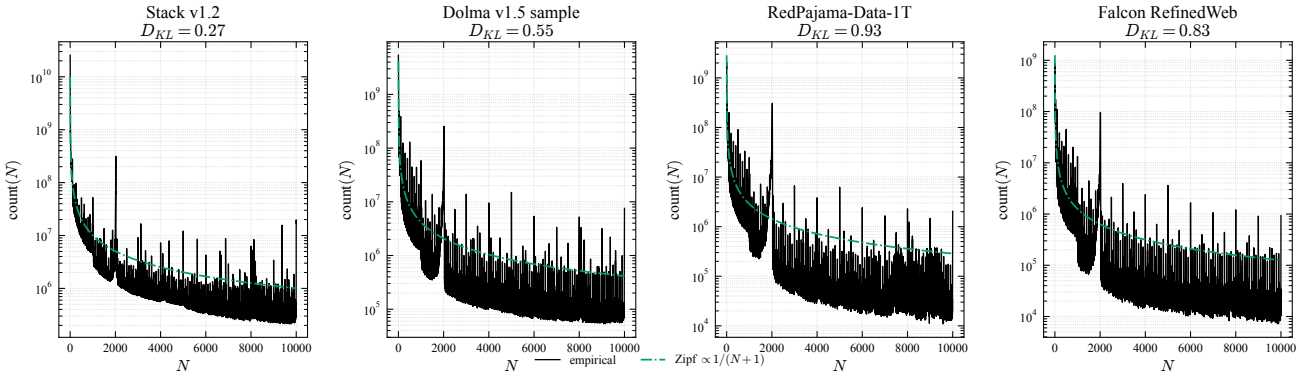

Figure 5. Empirical integer-frequency distributions for the pretraining datasets. The dashed curve shows the Zipf baseline $q(N) \propto (N+1)^{-1}$, and D_{KL} measures how far each empirical distribution deviates from this baseline.

Integer frequencies in natural text are not perfectly smooth and often exhibit systematic spikes due to round numbers, years, dates, and culturally salient values. To characterize these deviations, we analyze round-number categories:

$$N \equiv 0 \pmod{10}, \qquad N \equiv 0 \pmod{100}.$$

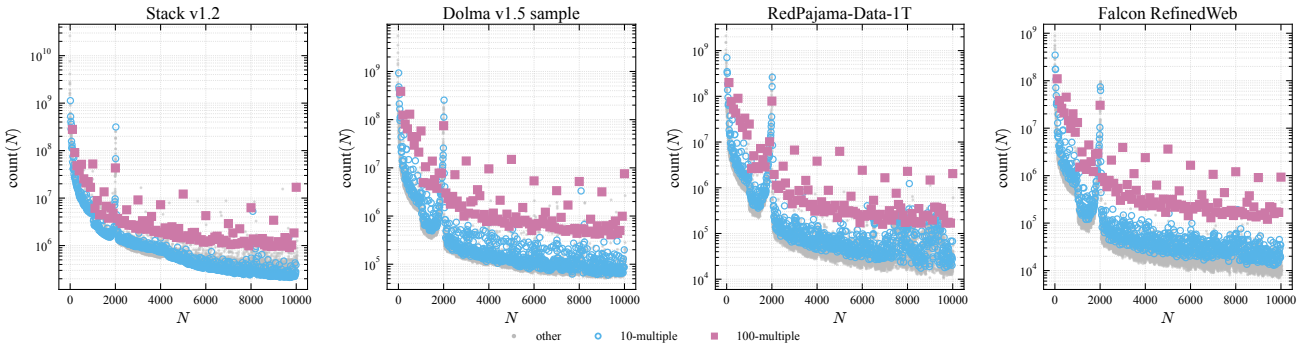

Figure 6. Integer-frequency distributions with round-number and year-like effects. Multiples of 10 and 100 show systematic spikes, with additional peaks around values such as 2000 likely reflecting frequent year mentions.

D. Additional Representation-Geometry Diagnostics

The main paper visualizes the PCA number-line projections. Here, we report the full PCA-based numerical representation analysis across models, including the selected layer, monotonicity score ρ, compression factor β, and explained variance σ^2. The selected layer is the layer with the highest explained variance for the one-dimensional PCA projection.

Aggregating the PCA scaling factors by matched corpus–model family gives the corpus-level relationship shown in Figure 7. Less negative α values correspond to larger β, supporting the same trend discussed in the main paper: flatter integer-frequency decay is associated with less compressed number-line geometry.

Model	Dataset	PCA			
		Layer	ρ	β	σ^2
Falcon-RW-1B	RefinedWeb	8	0.93 ± 0.01	0.61 ± 0.01	0.31 ± 0.00
Falcon-RW-7B		16	0.93 ± 0.01	0.75 ± 0.03	0.33 ± 0.01
INCITE-3B	RedPajama	16	0.93 ± 0.01	0.92 ± 0.06	0.39 ± 0.01
INCITE-7B		12	0.92 ± 0.01	0.86 ± 0.07	0.37 ± 0.01
OLMo-7B-2T	Dolma	20	0.92 ± 0.01	1.67 ± 0.23	0.49 ± 0.02
OLMo-7B-Twin-2T		13	0.92 ± 0.00	1.54 ± 0.12	0.50 ± 0.01
StarCoderBase-1B	Stack	10	0.82 ± 0.01	5.60 ± 2.86	0.37 ± 0.01
StarCoderBase-3B		16	0.81 ± 0.01	10.62 ± 6.07	0.37 ± 0.01
StarCoderBase-7B		20	0.85 ± 0.01	4.33 ± 1.10	0.37 ± 0.00

Table 3. PCA-based analysis of numerical representations across models, reporting the selected layer, monotonicity score ρ, compression factor β, and explained variance σ^2.

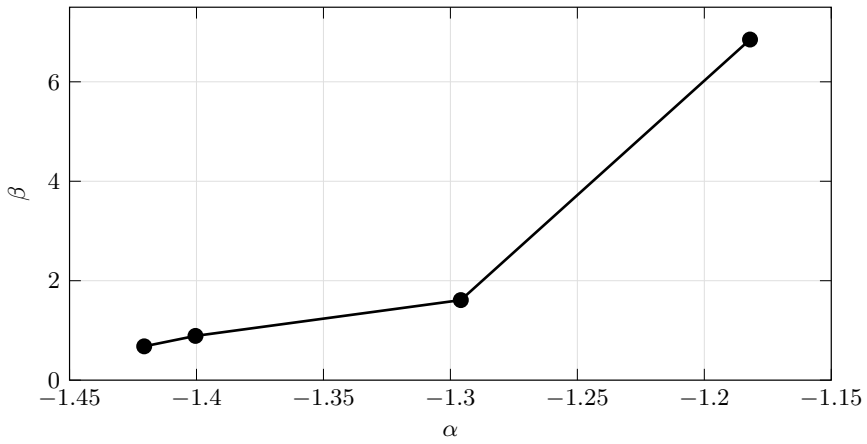

Figure 7. Corpus exponent α vs. PCA scaling factor β. Counts are fit as $C(N) \propto N^{\alpha}$, where more negative α indicates faster integer-frequency decay. Each point aggregates one matched corpus–model family.

E. Additional Details for the Number-Comparison Probe

This appendix provides additional details for the controlled number-comparison probe introduced in Section 4. The probe measures whether a model assigns higher likelihood to the numerically correct candidate in a pairwise comparison, rather than evaluating unrestricted mathematical generation.

E.1. Synthetic Task Construction

Each example contains two integers (a, b) and an instruction asking for either the larger or the smaller value. We generate pairs as

$$(a, b) = (n, n + d), \tag{7}$$

where n is sampled from one of four magnitude groups and d is a fixed absolute gap. The magnitude groups are

$$\mathcal{G}_1 = [10, 99], \quad \mathcal{G}_2 = [100, 999], \quad \mathcal{G}_3 = [1000, 9999], \quad \mathcal{G}_4 = [10000, 99999]. \tag{8}$$

We use two gap regimes:

$$\mathrm{SG} = \{1, 2, 3, 4, 5\}, \qquad \mathrm{MG} = \{6, 7, 8, 9, 10\}. \tag{9}$$

For each magnitude group and exact gap, we sample 32 starting values with a fixed seed. We include both candidate orderings, (a, b) and (b, a), and both query directions, "larger" and "smaller", so the benchmark is balanced against positional and lexical shortcuts.

E.2. Likelihood-Based Scoring

For a prompt q, the valid continuations are the two candidate numbers c_a and c_b. We score each candidate using length-normalized log-likelihood:

$$s(c \mid q) = \frac{1}{|c|} \sum_{t=1}^{|c|} \log p(c_t \mid q, c_{<t}), \tag{10}$$

where $|c|$ is the number of tokens in the candidate continuation. The model prediction is

$$\hat{c} = \arg \max_{c \in \{c_a, c_b\}} s(c \mid q). \tag{11}$$

Accuracy is then

$$\text{Acc} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}[\hat{c}_j = c_j^\star], \tag{12}$$

with $M = 15{,}360$ examples per model-shot condition. This evaluation avoids sampling variance and measures the model's relative preference between the two numerical candidates.

E.3. Few-Shot Results

Shots	ρ_s	Pearson r	R^2	Mean Acc.
0	0.84	0.80	0.64	0.64
1	0.83	0.80	0.64	0.63
2	0.74	0.79	0.62	0.62
3	0.80	0.81	0.65	0.62
4	0.82	0.85	0.72	0.62

Table 4. Correlation between PCA scaling factor β and number-comparison accuracy across shot settings. Correlations are computed across the 14 evaluated base models.

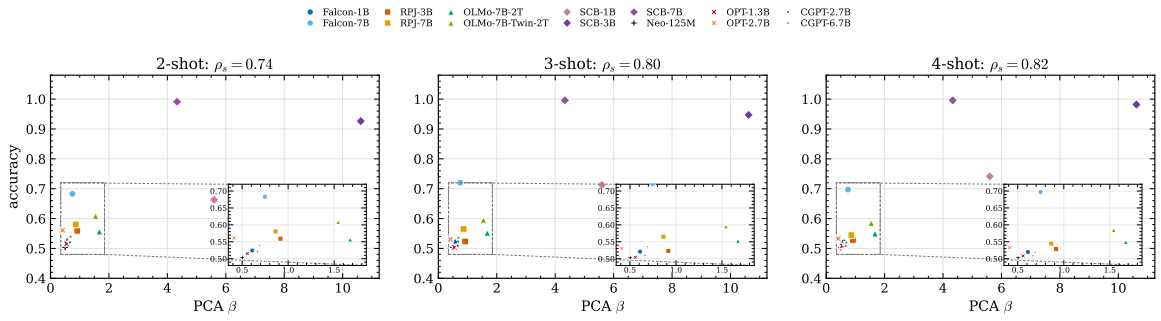

Figure 8. Number-comparison accuracy versus PCA scaling factor β for 2-, 3-, and 4-shot settings. Each point is one model; the inset zooms into the low-β region.

The positive Spearman correlations in Table 4 show that the β–accuracy relationship is not specific to the zero-shot prompt. The relationship remains positive from 1-shot through 4-shot prompting, and the 2–4 shot panels in Figure 8 show the same qualitative pattern as the main 0–1 shot figure.

Model	Family	β	Type	0-shot	1-shot	2-shot	3-shot	4-shot
Falcon-1B	Falcon	0.61	baseline	0.519	0.530	0.524	0.521	0.520
Falcon-7B	Falcon	0.75	baseline	0.633	0.647	0.683	0.720	0.697
RPJ-3B	RedPajama	0.92	baseline	0.556	0.566	0.559	0.523	0.529
RPJ-7B	RedPajama	0.86	baseline	0.528	0.599	0.580	0.565	0.545
OLMo-7B-2T	OLMo	1.67	baseline	0.686	0.582	0.557	0.552	0.550
OLMo-7B-Twin-2T	OLMo	1.54	baseline	0.694	0.614	0.609	0.596	0.585
SCB-1B	StarCoderBase	5.60	baseline	0.737	0.676	0.663	0.713	0.741
SCB-3B	StarCoderBase	10.62	baseline	0.897	0.922	0.926	0.947	0.982
SCB-7B	StarCoderBase	4.33	baseline	0.974	0.991	0.991	0.996	0.996
CGPT-2.7B	Cerebras-GPT	0.67	candidate	0.568	0.537	0.521	0.509	0.508
CGPT-6.7B	Cerebras-GPT	0.69	candidate	0.599	0.542	0.539	0.535	0.519
Neo-125M	GPT-Neo	0.50	candidate	0.500	0.502	0.503	0.503	0.503
OPT-1.3B	OPT	0.56	candidate	0.541	0.554	0.515	0.504	0.508
OPT-2.7B	OPT	0.41	candidate	0.534	0.566	0.561	0.530	0.533

Table 5. Per-model number-comparison accuracy across shot settings. Candidate models are included only in the behavioral comparison because their exact pretraining corpora are not matched in the corpus-frequency analysis.

E.4. Interpretation of the Behavioral Probe

The number-comparison probe should be interpreted as a controlled preference test. Since both candidate answers are scored directly, the result is not affected by stochastic decoding or by whether the model chooses to emit explanatory text. However, it is still not a complete measure of mathematical reasoning: it tests pairwise numerical discrimination under a fixed prompt format. The consistent positive relationship between β and accuracy in Tables 4 and 5 supports the main claim that less compressed number-line geometry is associated with better numerical comparison behavior.

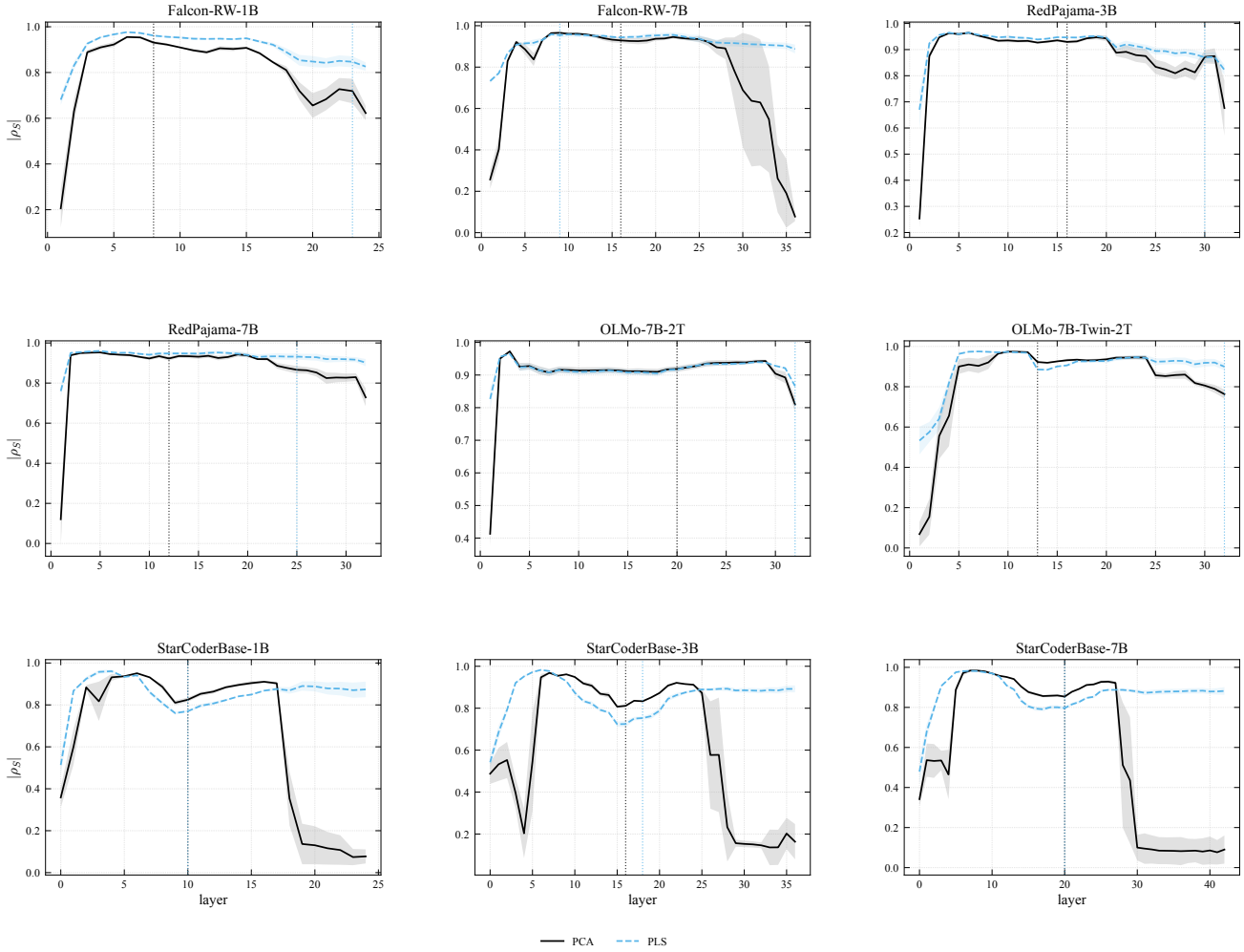

Figure 9. Layer-wise monotonicity of numerical representations using PCA and PLS. The curves show the absolute Spearman correlation $|\rho_s|$ across layers, and the dotted vertical lines indicate the layer with the highest explained variance selected for extracting the final representation.

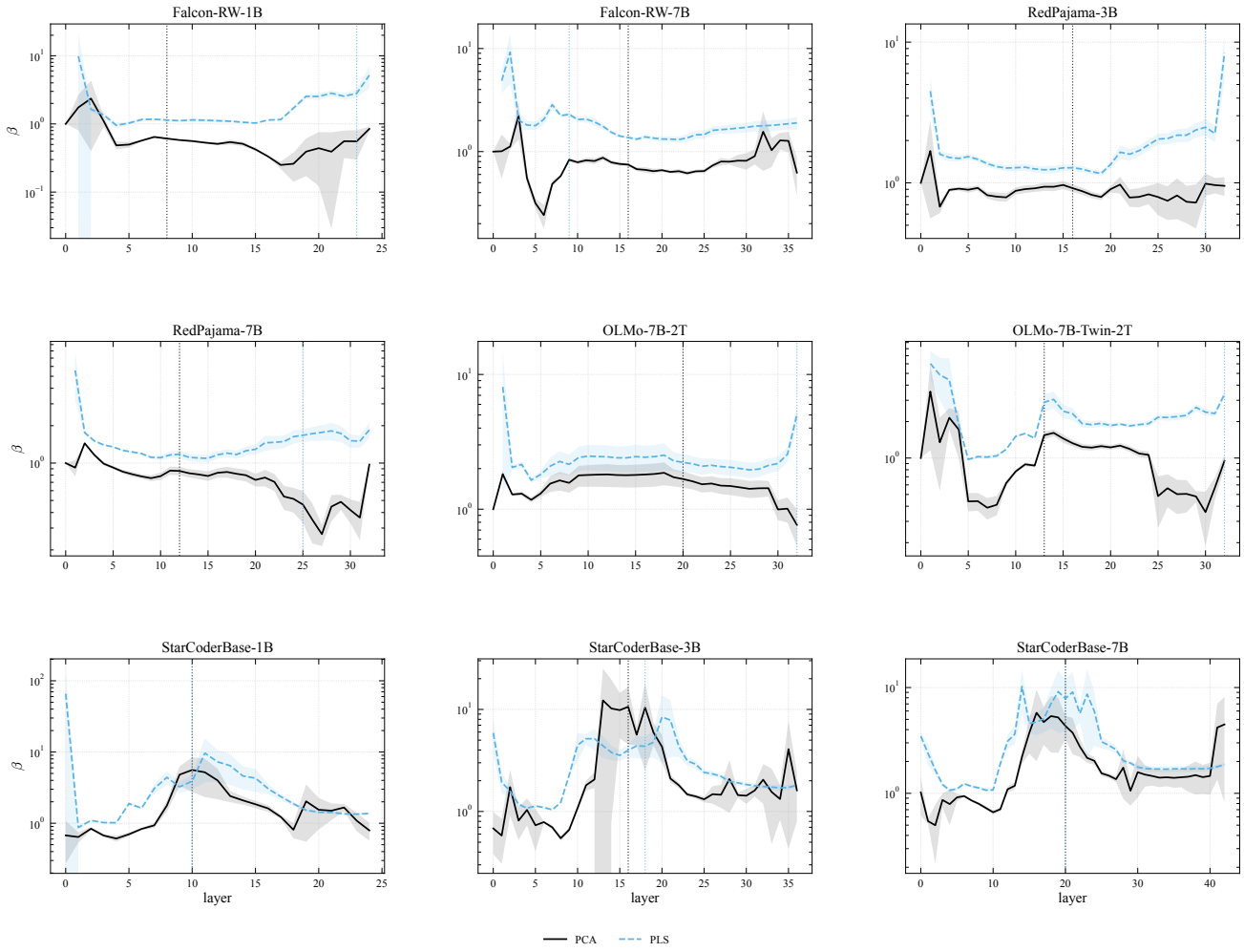

Figure 10. Layer-wise scaling factor β for PCA and PLS numerical projections. The curves show how the estimated number-line compression factor varies across layers, and the dotted vertical lines mark the highest-explained-variance layer used to extract the final representation.