Pretraining Numerical Frequency and Number-Line in Language Models

Mohammed Ibrahim Awad* Ahmed Elshehaby* Velibor Bojkovic† Hilal AlQuabeh†

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

ICML International Conference On Machine Learning

Overview and Architecture Diagram

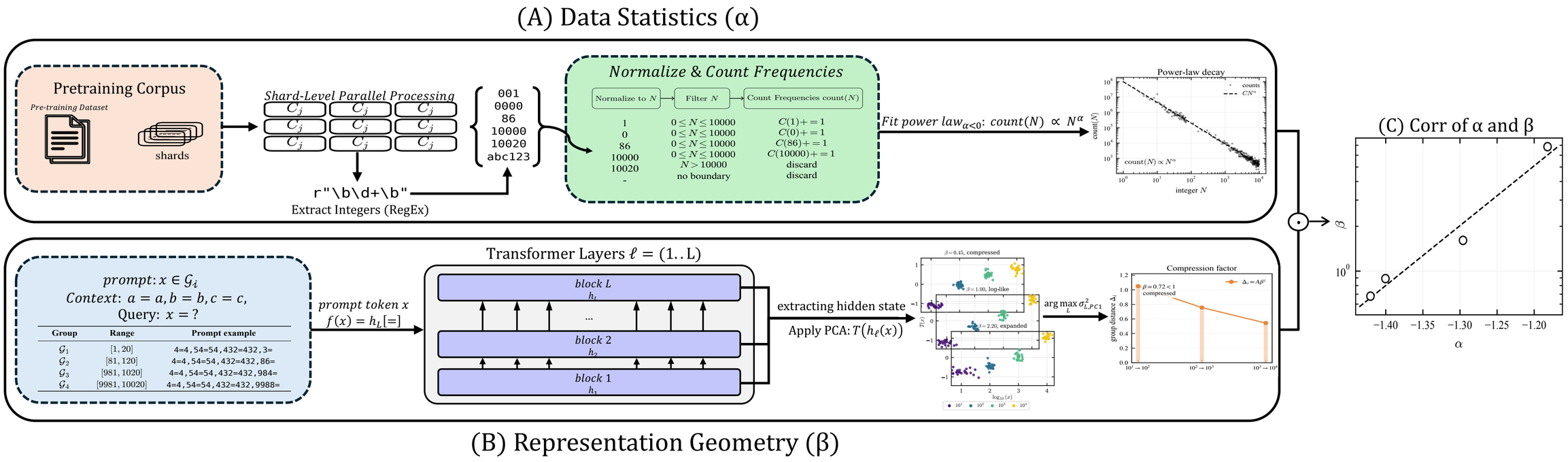

Figure 1. Overview of the full pipeline. (A) Extract digit strings from pretraining corpora, normalize them to integer values, count $N \in [0 : 10{,}000]$, and fit $\mathrm{count}(N) \propto N^{\alpha}$. (B) Run number prompts through each model, project hidden states with PCA, and estimate number-line scaling β. (C) Compare corpus slopes, representation geometry, and number-comparison behavior.

Problem and Hypothesis

Question: do pretraining integer statistics help explain why language models learn different internal number-line geometries?

- LLM hidden states often encode numerical magnitude along a low-dimensional direction.
- The resulting number line is not uniformly spaced: larger magnitudes may be compressed.
- We test whether corpus-level exposure to large integers is associated with the compression factor β.

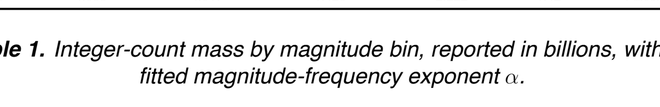

Figure 2. Conceptual logarithmic number line: spacing shrinks as magnitude increases.

Measurement Definitions

Corpus frequency decay

For each corpus, count every integer $N \in [0 : 10{,}000]$ and fit

$$\log \mathrm{count}(N) = c + \alpha \log N + \epsilon_N.$$

More negative α means faster decay in exposure to large numbers.

Representation scaling

For each model, project hidden states to one dimension and fit

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

Smaller β means stronger compression; larger β means more expanded spacing at large magnitudes.

Corpora, Counts, and Matched Models

Corpus	$[0, 10)$	$[10, 10^2)$	$[10^2, 10^3)$	$[10^3, 10^4]$	α	Matched models
Stack v1.2	51.530	23.656	13.420	10.050	-1.18	SCB-1B/3B/7B
Dolma v1.5	18.879	11.297	4.701	5.725	-1.30	OLMo-7B/7B-Twin
RedPajama	9.849	8.402	3.059	6.741	-1.40	INCITE-3B/7B
RefinedWeb	4.633	4.017	1.426	2.222	-1.42	Falcon-RW-1B/7B

Table 1. Integer-count mass by magnitude bin, reported in billions, with the fitted magnitude-frequency exponent α.

Empirical Integer Frequencies

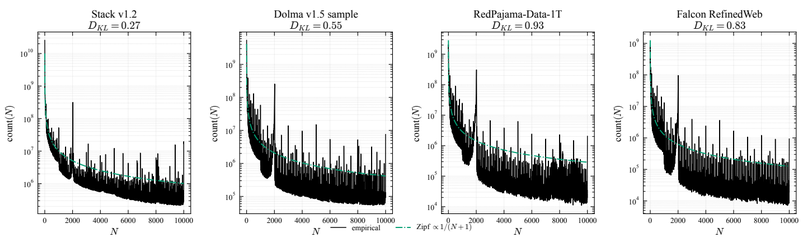

Figure 3. All four corpora are heavy-tailed, but their slopes differ: Stack has the flattest decay and RefinedWeb/RedPajama have steeper decay.

Round-Number and Year Effects

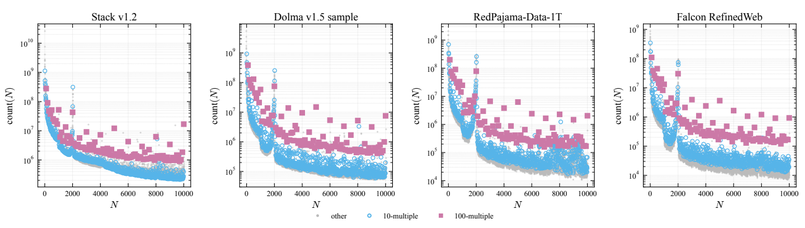

Figure 4. Integer frequency is not a smooth Zipf curve: multiples of 10/100 and year-like values remain visible after log scaling.

Corpus Diagnostics

Corpus	α	R^2	D_{KL}
Stack v1.2	-1.18	0.91	0.27
Dolma v1.5	-1.30	0.81	0.55
RedPajama	-1.40	0.68	0.93
RefinedWeb	-1.42	0.79	0.83

Table 2. α captures signed magnitude decay, R^2 measures log-log fit quality, and D_{KL} captures deviation from the Zipf-like baseline shown in Figure 3.

Hidden-State Number Lines

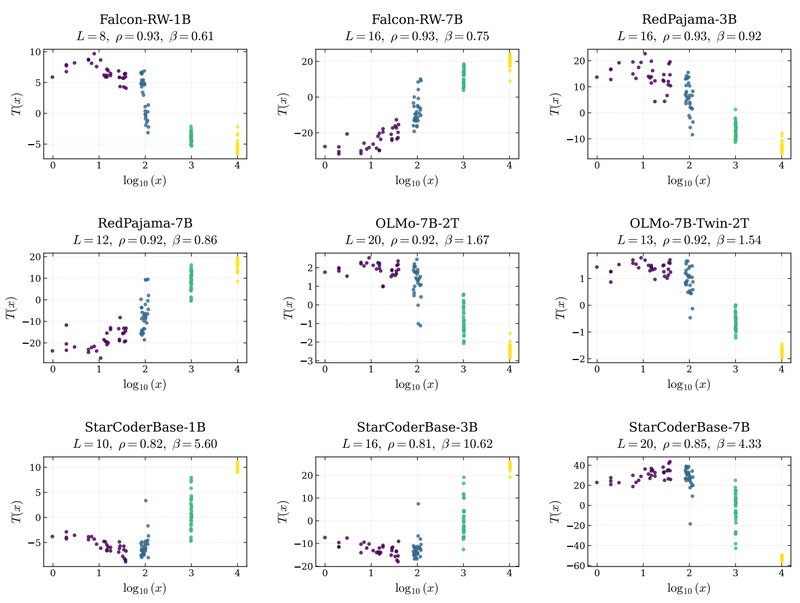

Figure 5. Selected-layer PCA projections form ordered numerical directions for all nine matched models, but spacing changes sharply across model families.

Model-Level Geometry

Model	Corpus	Layer	ρ	β	σ^2
Falcon-RW-1B	RefinedWeb	8	0.93	0.61	0.31
Falcon-RW-7B	RefinedWeb	16	0.93	0.75	0.33
INCITE-3B	RedPajama	16	0.93	0.92	0.39
INCITE-7B	RedPajama	12	0.92	0.86	0.37
OLMo-7B-2T	Dolma	20	0.92	1.67	0.49
OLMo-7B-Twin	Dolma	13	0.92	1.54	0.50
SCB-1B	Stack	10	0.82	5.60	0.37
SCB-3B	Stack	16	0.81	10.62	0.37
SCB-7B	Stack	20	0.85	4.33	0.37

Table 3. Best-layer PCA summary from the paper. ρ measures monotonic order; β measures number-line spacing across scales.

Corpus Frequency Predicts Geometry

Less negative α values correspond to larger β values: flatter integer-frequency decay is associated with less compressed or more expanded internal number-line spacing.

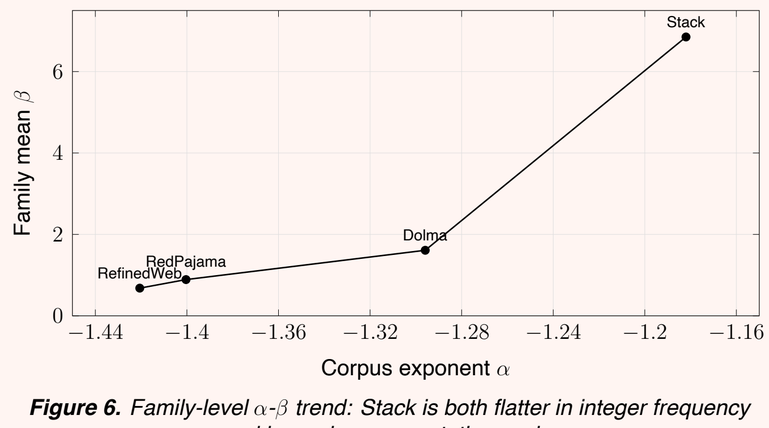

Figure 6. Family-level α-β trend: Stack is both flatter in integer frequency and larger in representation scale.

Layer-Wise Checks

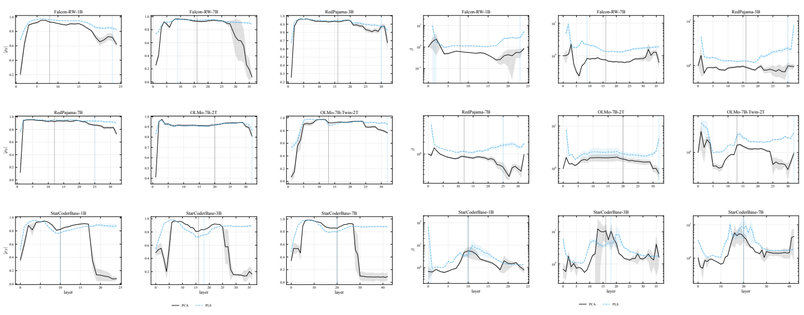

Figure 7. Layer-wise checks: left panel = monotonicity $|\rho_s|$; right panel = scaling β, across model layers.

Full Paper & Code

Behavioral Probe

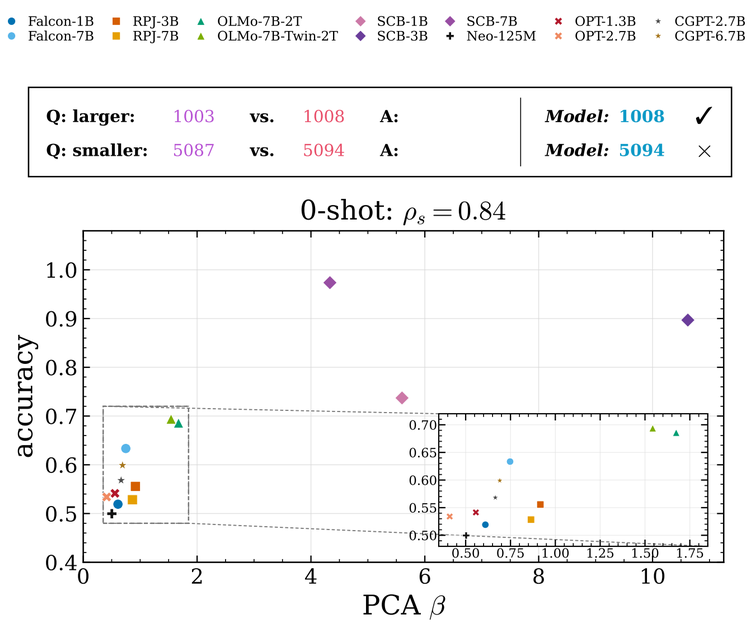

Figure 8. 0-shot comparison accuracy rises with β ($\rho_s = 0.84$); less compressed geometry is associated with better numerical discrimination.

Few-Shot Robustness

The β-accuracy relationship is not specific to the 0-shot prompt: correlations remain positive from 1-shot through 4-shot prompting.

Shots	0	1	2	3	4
Spearman ρ_s	.84	.83	.74	.80	.82
Pearson r	.80	.80	.79	.81	.85
R^2	.64	.64	.62	.65	.72
Mean Acc.	.64	.63	.62	.62	.62

Table 4. Correlation between PCA β and comparison accuracy across shot settings over 14 evaluated base models.

Per-Model Behavioral Scores

Model	β	0s	1s	4s
Falcon-1B	0.61	.519	.530	.520
Falcon-7B	0.75	.633	.647	.697
RPJ-3B	0.92	.556	.566	.529
RPJ-7B	0.86	.528	.599	.545
OLMo-7B-2T	1.67	.686	.582	.550
OLMo-7B-Twin	1.54	.694	.614	.585
SCB-1B	5.60	.737	.676	.741
SCB-3B	10.62	.897	.922	.982
SCB-7B	4.33	.974	.991	.996
CGPT-2.7B	0.67	.568	.537	.508
CGPT-6.7B	0.69	.599	.542	.519
Neo-125M	0.50	.500	.502	.503
OPT-1.3B	0.56	.541	.554	.508
OPT-2.7B	0.41	.534	.566	.533

Table 5. Likelihood-based comparison accuracy for the 0-, 1-, and 4-shot settings. Candidate rows are used only for behavior, not corpus matching.

Claim-Evidence Map

- **Corpus to geometry:** Stack has the flattest α and largest β; steeper corpora have smaller β.
- **Geometry to behavior:** larger β predicts higher controlled comparison accuracy across 0–4 shots.
- **Scope:** matched models test the α-β link; extra candidate models only test behavior.

Limitations

- Correlational evidence: tokenization, architecture, scale, and optimization may also shape number representations.
- Some corpus matches are public proxies for private mixtures.
- Digit-string comparison does not cover dates, units, written numbers, or full arithmetic generation.

References

H. V. AlquBoj, H. AlQuabeh, V. Bojkovic, T. Hiraoka, A. O. El-Shangiti, M. Nwadike, and K. Inui. Number Representations in LLMs: A Computational Parallel to Human Perception. arXiv:2502.16147, 2025.

Y. Razeghi, R. L. Logan IV, M. Gardner, and S. Singh. Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning. Findings of EMNLP 2022, pages 840–854.

J. Merullo, N. A. Smith, S. Wiegreffe, and Y. Elazar. On Linear Representations and Pretraining Data Frequency in Language Models. ICLR 2025.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. SIAM Review, 51(4):661–703, 2009.

D. Kocetkov et al. The Stack: 3 TB of Permissively Licensed Source Code. Transactions on Machine Learning Research, 2023.

*Equal contribution †Equal supervision