Probing Geometry of Next Token Prediction Using Cumulant Expansion of the Softmax Entropy

Karthik Viswanathan* K.VISWANATHAN@UVA.NL
Institute of Physics, University of Amsterdam, the Netherlands

Sang Eon Park* SANGEON@MIT.EDU
Department of Physics, Massachusetts Institute of Technology, Cambridge, U.S.A.

Abstract

We introduce a cumulant-expansion framework for quantifying how large language models (LLMs) internalize higher-order statistical structure during next-token prediction. By treating the softmax entropy of each layer's logit distribution as a perturbation around its "center" distribution, we derive closed-form cumulant observables that isolate successively higher-order correlations. Empirically, we track these cumulants in GPT-2 and Pythia models on Pile-10K prompts. (i) Structured prompts exhibit a characteristic rise–and–plateau profile across layers, whereas token-shuffled prompts remain flat, revealing the dependence of the cumulant profile on meaningful context. (ii) During training, all cumulants increase monotonically before saturating, directly visualizing the model's progression from capturing variance to learning skew, kurtosis, and higher-order statistical structures. (iii) Mathematical prompts show distinct cumulant signatures compared to general text, quantifying how models employ fundamentally different processing mechanisms for mathematical versus linguistic content. Together, these results establish cumulant analysis as a lightweight, mathematically grounded probe of feature-learning dynamics in high-dimensional neural networks.

1. Introduction

Growing evidence suggests that Large Language Models (LLMs) based on transformer architectures process information in distinct stages [7, 16, 23]. By tracking statistical quantities of latent prediction probabilities [3] such as mean, KL-divergence, and entropy across layers, as well as geometric properties of internal representations like intrinsic dimension [26] and curvature [23], researchers can investigate the emergent representations within these models.

We introduce a novel method for examining information evolution within different LLM layers using the cumulant expansion of the entropy from softmax-generated probability distributions (hereafter referred to as 'softmax entropy'). In both training and inference of LLMs, the softmax function serves as a critical bridge translating between internal model representations and next-token prediction. By analyzing LLM through the lens of information theory, particularly through cumulants, we formalize how information is learned during each training step and how next token predictions are refined across layers. In this paper, we propose that cumulants of softmax entropy effectively capture the emergence of higher-order moments and demonstrate their efficacy in the study of off-the-shelf LLMs through i) experiments with structured and shuffled prompts, and ii) studying the evolution of cumulants during training. We believe that our new observable provides mathematical language for explicitly demonstrating neural network learning higher-order correlations within data.

* Equal contribution

© K. Viswanathan* & S.E. Park*.

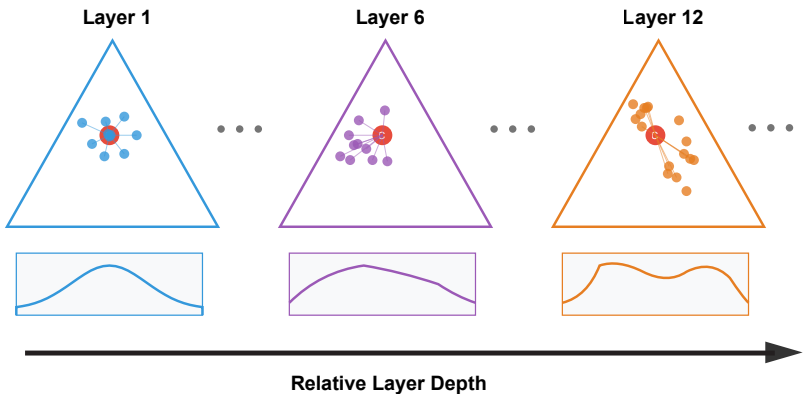

Figure 1: **Schematic of Logit Geometry Across Layers.** Each triangle represents the probability simplex at a given layer, where colored dots correspond to token logits mapped to probabilities. The red circle labeled 'c' indicates the probability of the center of logits. The histograms illustrate the distribution of token-wise deviations from the center.

2. Related Work

Several recent studies have investigated the evolution of LLM representations through their layers, examining information refinement and emergent behaviors within these architectures [7, 16, 23]. Mean field theoretical approaches offer complementary insight into the dynamics in neural network [6, 10, 12, 14, 17, 22, 24]. Analyzing internal representations points to distributional simplicity bias, wherein neural networks learn lower-order statistics of input data before progressing to higher-order statistics [4, 20, 21].

Information theory has emerged as a valuable tool for understanding different phenomena in large language models, as demonstrated in previous works [1, 8, 9, 25]. The cumulant expansion, a well-established method in statistical physics, has found applications across diverse fields, including astrophysics [5] and chemical physics [15]. In particular, [2] expresses KL divergence as an expansion in cumulants for intractable distributions. Complementary to our approach, [27] establishes a relationship between intrinsic dimension and softmax entropy in internal representations, while [13] examines skew and kurtosis of token activations to understand universal neuron properties.

3. Cumulants as Observables

We propose cumulants from the expansion of the softmax entropy as variables to probe the internal representation of the tokens. Let $\mathbf{X} = \{X_1, X_2, ...X_N\}$ be the collection of logits for N tokens, and let $\boldsymbol{\mu}$ be the "center" of the logits defined as the point minimizing the sum of *distances* to the sample points $\mathbf{X}$, where the *distance* is the KL-divergence from the minimizing point to the sample points,

$$\boldsymbol{\mu} = \arg\min_{Y} \sum_{i} D_{KL}\left(p(X_i)||p(Y)\right) \quad \Rightarrow \quad p(\boldsymbol{\mu}) = \frac{1}{N}\sum_{i=1}^{N} p(x_i) \tag{1}$$

Eq. 1 shows that the next token prediction corresponding to the center is the average of the next token prediction of the collection of logits.

Let $S(X_i)$ be the softmax entropy of the token X_i, defined as the entropy of the discrete distribution over the vocabulary after applying the softmax function. Let $\langle S(\mathbf{X})\rangle$ be the arithmetic mean of the softmax entropy of all the tokens in a prompt. Then we can perform the perturbative expansion of $S(\mathbf{X})$ around the softmax entropy of the center $S(\boldsymbol{\mu})$.

$$\langle S(\mathbf{X})\rangle = S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{i=1}^{N} D_{KL}(p_i||p_{\boldsymbol{\mu}}) \tag{2}$$

where $p_i = \mathrm{softmax}(X_i)$ denotes the softmax probability distribution over the vocabulary for token i, and $p_{\boldsymbol{\mu}}$ the softmax probability distribution for the center of logits, as defined in Eq. 1. Using Eq. 2, we can motivate $\frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}(p_i \,\|\, p_{\boldsymbol{\mu}})$ as a measure of interaction between logits and the center. The underlying intuition is that the entropy of the prompt is obtained by subtracting this interaction term from the entropy of the center. This view is supported by experimental results in Section 4. By expanding the KL-divergence, the second term in the sum of the right-hand side of Eq. 2 can be expressed as cumulants of the softmax probability distribution of $\mathbf{X}$.

$$D_{\mathrm{KL}}(p_\beta(X_i)\|p_\beta(\boldsymbol{\mu})) = \sum_{n=2}^{\infty} \frac{\beta^n}{n!}\kappa_n^{p_\beta(X_i)}(-\delta X_i) \tag{3}$$

Eq. 3 shows concretely how the KL divergence can be expressed in terms of cumulants using cumulant generating functions, where δX_i is a random variable associated with the ith token in a prompt, constructed by sampling a vocabulary token j from the softmax probability distribution $p_\beta(X_i)$ and assigning it the value $\delta X_{ij} = X_{ij} - \mu_j$. $\beta = \frac{1}{T}$, T is the temperature. The superscript $p_\beta(X_i)$ denotes the underlying distribution of the random variables X_i used in δX_i.

Plugging this back into Eq. 2, we finally obtain the relation in Eq. 4.

$$\langle S(\mathbf{X})\rangle = S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{n=2}^{\infty} \frac{\beta^n}{n!}\kappa_n^{p_\beta(\mathbf{X})}\left(-\sum_{i=1}^{N}\delta X_i\right) \tag{4}$$

Eq. 4 demonstrates that by observing cumulants, we effectively measure the distribution of logits—specifically, how token logits are distributed relative to the center logit. Furthermore, this equation reveals that token-wise average cumulants correspond to the cumulants of the aggregated random variable $\delta X = \sum_{i=1}^{n}\delta X_i$. We verify this with Monte-Carlo simulation in Appendix B.2.

This implies that prompts with different δX distributions produce distinct cumulants; conversely, identical δX distributions yield the same cumulants. In theory, the average cumulants contain the same information as the distribution $p_\beta(\delta X)$, and this equivalence offers insight into what our observables capture. However, accurately estimating $p_\beta(\delta X)$ directly from δX_i requires many simulations. Hence, computing the average of token-wise cumulants provides a more computationally efficient and reliable way of probing $p_\beta(\delta X)$ in practice.

We could also make an analogy to the perturbative Taylor series expansion, where we are performing the perturbative expansion of the entropy, and by calculating higher-order cumulants, we consider successively higher-order correlations within token logits. The cumulant summary statistics we introduce probe the geometry of the latent probabilities that reside in the probability simplex. In particular, different sets of logits that yield the same set of probabilities post-softmax will produce identical cumulants. We show the full derivation of each equation, and how we calculate each cumulant in practice in Appendix A and C.

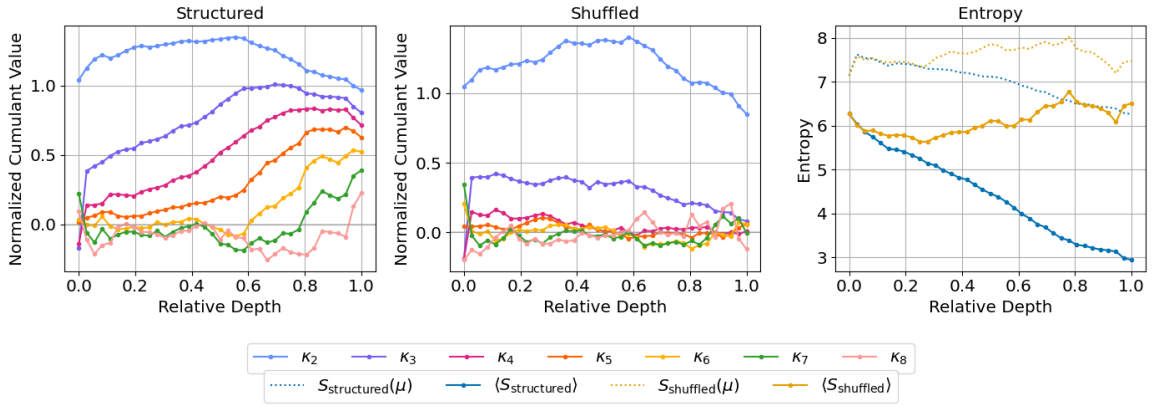

Figure 2: **Cumulants in Structured and Shuffled Prompts.** Left: Cumulants across layers for a single structured prompt (3218 from Pile 10K) in GPT-2 Large. Middle: Cumulants for the shuffled version of the same prompt. Right: Comparing mean softmax entropy (solid lines) and the entropy of the center (dotted lines) for both structured and shuffled prompts.

4. Results

We use the Pile-10K dataset [19], a subset of the Pile dataset [11], as a representative sample of diverse textual data. We first show the result with the 3218th prompt from the Pile-10K dataset, with the Pile set name: *ArXiv*. We then compare the results for two different types of prompts: general text and mathematics prompts. We also show that the results generalize to other prompts in the dataset, as well as different models in Appendix B.

4.1. Cumulants in Structured and Shuffled Prompts

We refer to the original prompts as "structured" prompts, while prompts with randomly permuted token orders are called "shuffled" prompts. To obtain latent predictions from the intermediate layers, we employ TunedLens [3]. For all results, We normalize the nth cumulant by dividing it by $n!$ to reflect its contribution to the softmax entropy as seen in Eq. 4.

We compare the cumulant profiles of a structured prompt (left) and its shuffled counterpart (middle), noting that the second cumulant remains consistently near 1.0 in both cases. The structured prompt exhibits a clear depth-dependent trend for the higher-order cumulants: they increase through intermediate layers, plateau, and then slightly decline near the final layers. In contrast, the cumulants for the shuffled prompt (middle plot) remain relatively constant around zero for higher cumulants, with significantly less variation than those of the structured prompt. We conclude that when the model captures patterns within data, token logit deviations from the center develop a characteristic pattern across layers.

A more striking result appears in the right panel, which compares the mean softmax entropy with the softmax entropy of the center. For structured prompts (blue lines), the gap between the center's softmax entropy (dotted line) and the mean softmax entropy of all tokens (solid line) increases with the layers. The result indicates stronger higher-order relations in deeper layers for

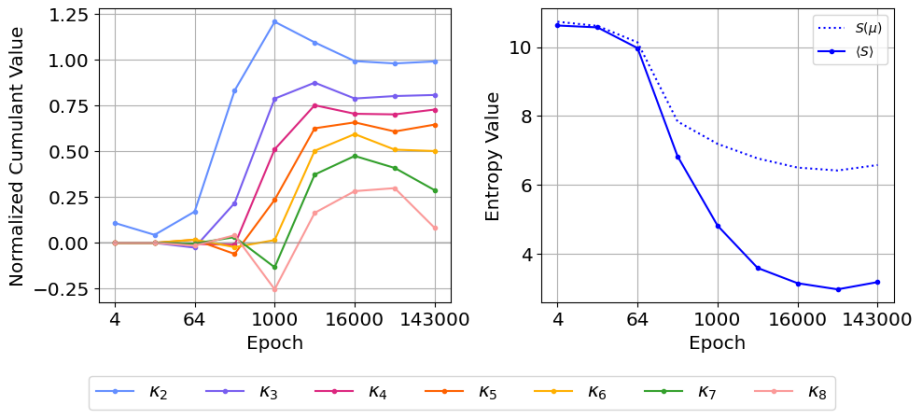

Figure 3: **Evolution of Cumulants During Training.** Left: Cumulants across layers of the Pythia-160M model tracked over training epochs. Right: mean softmax entropy (solid line) and softmax entropy of the center (dotted line) as a function of training.

structured prompts, with stronger interaction between the center and the logits. In contrast, for shuffled prompts (yellow lines), the higher-order cumulants remain mostly constant.

To test for the dependency on prompts and models, we rerun the experiment on different prompts from the Pile dataset and different size models from the GPT-2 family. The pattern we observe is almost identical across all prompts and models, and can be seen in Appendix B.1.

4.2. Evolution of Cumulants during Training

We further investigate the evolution of cumulants during training and what they reveal about neural network learning dynamics. The left panel of Fig 3 shows that all cumulants increase before plateauing as training progresses. This indicates that the logit distribution—specifically, the distribution of deviations from the center—becomes increasingly complex. We observe larger spread (variance), greater asymmetry (skewness, related to the third moment), heavier tails (kurtosis, from the fourth moment), and other higher-order statistical properties.

The right panel of Fig 3 demonstrates that as training progresses, the discrepancy between the center's softmax entropy and the mean softmax entropy of all token logits increases. This growing gap indicates that the center becomes less capable of explaining the total entropy, requiring higher-order information to accurately characterize the logit distribution. This pattern explicitly demonstrates that models progressively learn higher-order correlations within data during training.

4.3. DM Mathematics vs Pile-CC: A Cumulant Perspective

Prior research [18, 28] has established that contexts involving logic and language produce distinct geometric properties in neural network representations. To compare how code/math prompts are processed differently from general text processing in language models, we analyze the cumulant-based statistical properties of next-token prediction as it evolves across layers. We compare prompts from DM Mathematics against Pile-CC, representing general web text from the Pile-10k [19] dataset in Figure 4, where we notice distinct statistical signatures between mathematical and general text

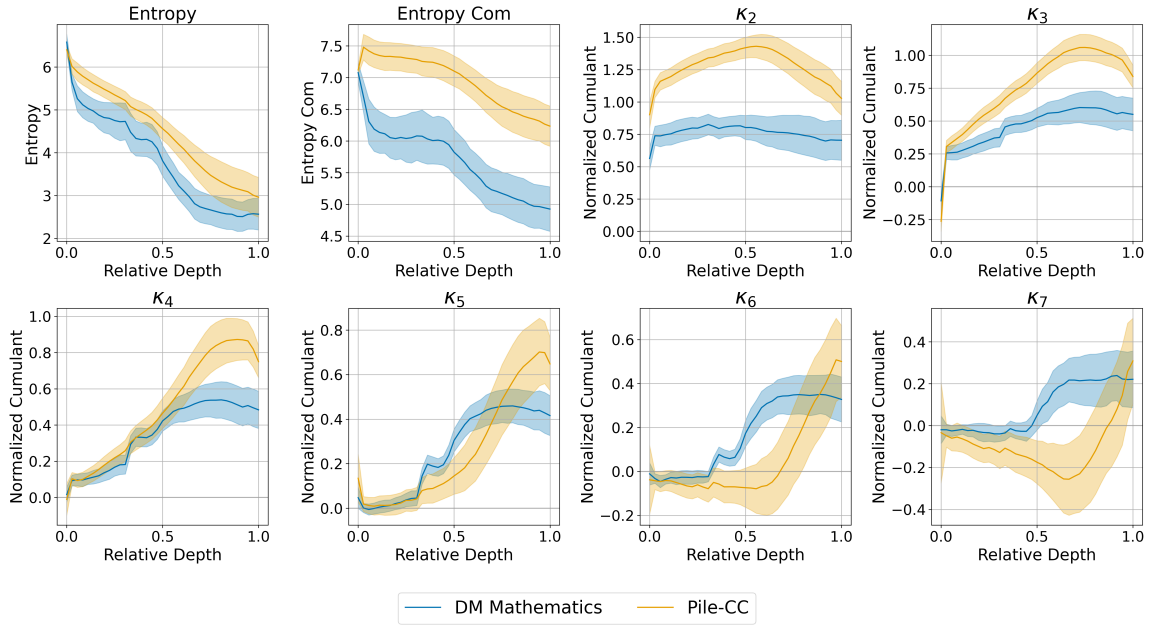

Figure 4: **Cumulants in DM Mathematics and Pile-CC Prompts.** Comparison of normalized cumulants (κ_2 through κ_7) and entropy measures across model layers for mathematical prompts (DM Mathematics topic with 99 prompts) versus general web text (Pile-CC topic with 570 prompts) in GPT-2 Large. Each plot shows the mean (solid line) and standard deviation (shaded region) computed across multiple prompts from each dataset. Mathematical prompts exhibit distinct cumulant profiles compared to general text.

processing through cumulant analysis. Mathematical prompts show lower entropy than Pile-CC prompts across layers, though the separability between distributions remains modest. We observe clearer differences in the entropy of the center $S(\mu)$ and lower-order cumulants κ_2 and κ_3 across layers, with mathematical prompts exhibiting lower values than Pile-CC. The higher-order cumulants κ_6 and κ_7 reveal a qualitative shift in the middle layers: mathematical prompts show higher values than Pile-CC in these layers. Together, these statistical signatures quantify how the model employs fundamentally different mechanisms for mathematical versus general text processing.

5. Conclusion

We propose cumulants as a probe for examining emergent statistical properties within LLMs and demonstrate their effectiveness in revealing representational patterns across model layers. Our new observables provide a mathematical framework for explicitly showing how neural networks learn higher-order correlations within data. Our experiments reveal that in later LLM layers, higher-order cumulants increase and vary significantly for structured prompts but much less for shuffled prompts. We also show during training, the model progressively learn higher-order correlations within data during training. Future research could explore the correlation between loss and cumulants, as well as the causal effects of directly manipulating cumulants.

6. Reproducibility

All the results contained in this work are reproducible by means of a GitHub repository that can be found at https://github.com/karthikviswanathn/cumulant-expansion-llm.

Acknowledgements

We thank Boris Post for helpful discussions on the cumulant expansion of softmax entropy and Matteo Biagetti for the application of this framework to large language models. We acknowledge the Dutch Research Council (NWO) in The Netherlands for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through the 'Computing Time on National Computer Facilities' call. We used LUMI to run the computations required for this project.

References

[1] Riccardo Ali, Francesco Caso, Christopher Irwin, and Pietro Liò. Entropy-lens: The information signature of transformer computations. *arXiv preprint arXiv:2502.16570*, 2025.

[2] David Barber and Piërre van de Laar. Variational cumulant expansions for intractable distributions. *J. Artif. Int. Res.*, 10(1):435–455, June 1999. ISSN 1076-9757.

[3] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[4] Nora Belrose, Quintin Pope, Lucia Quirke, Alex Mallen, and Xiaoli Fern. Neural networks learn statistics of increasing complexity. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

[5] Aoife Boyle, Alexandre Barthelemy, Sandrine Codis, Cora Uhlemann, and Oliver Friedrich. The cumulant generating function as a novel observable to cumulate weak lensing information. *The Open Journal of Astrophysics*, 6, jul 17 2023. doi: 10.21105/astro.2212.10351.

[6] Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[7] Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. *CoRR*, abs/2405.15471, 2024. URL https://doi.org/10.48550/arXiv.2405.15471.

[8] Henry Conklin and Kenny Smith. Representations as language: An information-theoretic framework for interpretability, 2024. URL https://arxiv.org/abs/2406.02449.

[9] Ann-Kathrin Dombrowski and Guillaume Corlouer. An information-theoretic study of lying in LLMs. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/forum?id=9AM5i1wWZZ.

[10] Kirsten Fischer, Alexandre René, Christian Keup, Moritz Layer, David Dahmen, and Moritz Helias. Decomposing neural networks as mappings of correlation functions. *Phys. Rev. Res.*, 4:043143, Nov 2022. doi: 10.1103/PhysRevResearch.4.043143. URL https://link.aps.org/doi/10.1103/PhysRevResearch.4.043143.

[11] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[12] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, 2024. URL https://arxiv.org/abs/2312.10794.

[13] Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in gpt2 language models, 2024. URL https://arxiv.org/abs/2401.12181.

[14] Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*. Springer International Publishing, 2020. ISBN 9783030464448. doi: 10.1007/978-3-030-46444-8. URL http://dx.doi.org/10.1007/978-3-030-46444-8.

[15] Jürgen Köfinger and Gerhard Hummer. Encoding prior knowledge in ensemble refinement. *The Journal of Chemical Physics*, 160(11):114111, 03 2024. ISSN 0021-9606. doi: 10.1063/5.0189901. URL https://doi.org/10.1063/5.0189901.

[16] Vedang Lad, Wes Gurnee, and Max Tegmark. The remarkable robustness of llms: Stages of inference? In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

[17] Moritz Layer, Moritz Helias, and David Dahmen. Effect of synaptic heterogeneity on neuronal coordination. *PRX Life*, 2:013013, Mar 2024. doi: 10.1103/PRXLife.2.013013. URL https://link.aps.org/doi/10.1103/PRXLife.2.013013.

[18] Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, March 2025. ISSN 1099-4300. doi: 10.3390/e27040344. URL http://dx.doi.org/10.3390/e27040344.

[19] Neel Nanda. Pile-10k dataset, 2022. URL https://huggingface.co/datasets/NeelNanda/pile-10k.

[20] Maria Refinetti, Alessandro Ingrosso, and Sebastian Goldt. Neural networks trained with sgd learn distributions of increasing complexity. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[21] Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. A distributional simplicity bias in the learning dynamics of transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=GgV6UczIWM.

[22] Kai Segadlo, Bastian Epping, Alexander van Meegen, David Dahmen, Michael Krämer, and Moritz Helias. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103401, oct 2022. doi: 10.1088/1742-5468/ac8e57. URL https://dx.doi.org/10.1088/1742-5468/ac8e57.

[23] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

[24] Eszter Szekely, Lorenzo Bardone, Federica Gerace, and Sebastian Goldt. Learning from higher-order correlations, efficiently: hypothesis tests, random features, and neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=uHml6eyoVF.

[25] Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory, 2024. URL https://arxiv.org/abs/2411.11984.

[26] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 51234–51252. Curran Associates, Inc., 2023.

[27] Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. The geometry of tokens in internal representations of large language models, 2025. URL https://arxiv.org/abs/2501.10573.

[28] George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SUc1UOWndp.

Appendix A. Derivations

We use the definition of the mean, the softmax function, and the KL-divergence.

$$\begin{aligned}
\langle S(\mathbf{X})\rangle &= \frac{1}{N}\sum_{i=1}^{N} S(X_i) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\sum_{\alpha=1}^{\mathcal{V}} p_{i\alpha}\log p_{i\alpha} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\sum_{\alpha=1}^{\mathcal{V}} p_{i\alpha}\log p_{i\alpha} - p_{i\alpha}\log p_{\mu\alpha} - \frac{1}{N}\sum_{i=1}^{N}\sum_{\alpha=1}^{\mathcal{V}} p_{i\alpha}\log p_{\mu\alpha} \\
&= -\sum_{\alpha=1}^{\mathcal{V}}\log p_{\mu\alpha}\left(\frac{1}{N}\sum_{i=1}^{N} p_{i\alpha}\right) - \frac{1}{N}\sum_{i=1}^{N} D(p_i||p_\mu) \\
&= -\sum_{\alpha=1}^{\mathcal{V}} p_{\mu\alpha}\log p_{\mu\alpha} - \frac{1}{N}\sum_{i=1}^{N} D(p_i||p_\mu) \\
&= S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{i=1}^{N} D(p_i||p_\mu)
\end{aligned}$$

Therefore we can obtain Eq. 2

$$\boxed{\langle S(\mathbf{X})\rangle = S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{i=1}^{N} D(p_i||p_\mu)} \tag{2}$$

Now, using the definition of the KL-divergence, and plugging in the distribution of the distribution of tokens from the softmax function, where the "partition function" $Z_\beta(\mathbf{X}) = \sum_{\alpha=1}^{\mathcal{V}} e^{\beta x_\alpha}$, and the definition of the cumulant generating function,

$$\begin{aligned}
D_{\mathrm{KL}}(p_\beta(X)\|p_\beta(\boldsymbol{\mu})) &= \sum_{\alpha=1}^{\mathcal{V}} \frac{e^{\beta x_\alpha}}{Z_\beta(\mathbf{X})}\log\frac{\frac{e^{\beta x_\alpha}}{Z_\beta(\mathbf{X})}}{\frac{e^{\beta \mu_\alpha}}{Z_\beta(\boldsymbol{\mu})}} \\
&= \sum_{\alpha=1}^{\mathcal{V}} \frac{e^{\beta x_\alpha}}{Z_\beta(\mathbf{X})}\left(\beta(x_\alpha - \mu_\alpha) + \log\frac{Z_\beta(\boldsymbol{\mu})}{Z_\beta(\mathbf{X})}\right) \\
&= \sum_{\alpha=1}^{\mathcal{V}} \frac{e^{\beta x_\alpha}}{Z_\beta(\mathbf{X})}\beta\delta x_\alpha + \log\sum_{\alpha=1}^{\mathcal{V}} \frac{e^{\beta x_\alpha}}{Z_\beta(\mathbf{X})} e^{-\beta\delta x_\alpha} \\
&= \mathbb{E}_{p_\beta(X)}\left[\beta\delta x\right] + \log\mathbb{E}_{p_\beta(X)}\left[e^{-\beta\delta x}\right] \\
&= \sum_{n=2}^{\infty}\frac{\beta^n}{n!}\kappa_n^{p_\beta(\mathbf{X})}(-\delta x) \quad (\because \text{ definition of cumulant generating function})
\end{aligned}$$

Therefore we can obtain Eq. 3

$$\boxed{D_{\text{KL}}(p_\beta(X)\|p_\beta(\boldsymbol{\mu})) = \sum_{n=2}^{\infty} \frac{\beta^n}{n!} \kappa_n^{p_\beta(\mathbf{X})}(-\delta X)} \tag{3}$$

Now, we use several properties of cumulants to express the deviation of the mean softmax entropy from entropy of the center as a function of cumulants of the sum of the logits' deviation from the center.

$$\begin{aligned}
\langle S(\mathbf{X})\rangle &= S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{i=1}^{N} D(p_i||p_\mu) \\
&= S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{i=1}^{N}\sum_{n=2}^{\infty} \frac{\beta^n}{n!} \kappa_n^{p_\beta(X_i)}(-\delta X_i) \\
&= S(\boldsymbol{\mu}) - \sum_{n=2}^{\infty} \frac{\beta^n}{n!} \langle \kappa_n^{p_\beta(X_i)}(-\delta X_i)\rangle \\
&= S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{n=2}^{\infty}\sum_{i=1}^{N} \frac{\beta^n}{n!} \kappa_n^{p_\beta(X_i)}(-\delta X_i) \\
&= S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{n=2}^{\infty} \frac{\beta^n}{n!} \kappa_n^{p_\beta(\mathbf{X})}\left(-\sum_{i=1}^{N}\delta X_i\right) \quad (\because \text{ additive property of cumulants})
\end{aligned}$$

Therefore we obtain Eq. 4

$$\boxed{\langle S(\mathbf{X})\rangle = S(\boldsymbol{\mu}) - \frac{1}{N}\sum_{n=2}^{\infty} \frac{\beta^n}{n!} \kappa_n^{p_\beta(\mathbf{X})}\left(-\sum_{i=1}^{N}\delta X_i\right)} \tag{4}$$

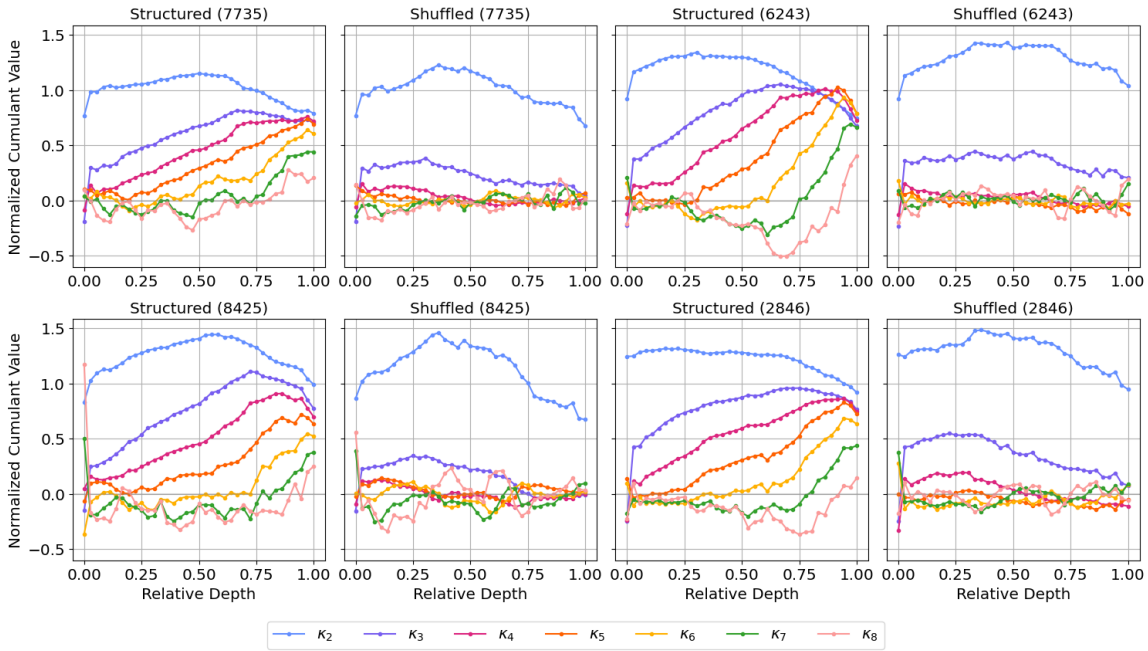

Figure 5: **Cumulants Across Prompts.** Cumulants of structured and shuffled versions of four randomly selected prompts from the Pile-10k dataset [19]. Each panel corresponds to a different prompt, and the numbers on the axis title represent the prompt number in the Pile-10K dataset. Structured prompts consistently show richer and more structured cumulant profiles compared to their shuffled counterparts.

Appendix B. Supplementary Figures

B.1. Cumulants for the prompts in the Pile dataset, on GPT2-Large

Fig.4 show the cumulants of structured and shuffled prompts for four different randomly selected prompts within the Pile-10k dataset [19], and Fig.5 show the result of a structured and shuffled prompt for different size models from the GPT2 family.

B.2. Verification of Cumulants

In this section we verify that the cumulants of $\delta X = \sum_{i=1}^{n} \delta X_i$ are indeed equal to the sum of the per-token cumulants δX_i following the discussion in Section 3. We fix a prompt and focus on a single layer L. For each of N Monte Carlo trials, we:

1. For each token i, sample a vocabulary index t_i according to the $p_\beta(X_i)$.
2. Calculate how "far" X_i is from the center along the direction t_i, by measuring $\delta X_{i,t_i} = X_{i,t_i} - \mu_{t_i}$.
3. Sum over tokens to obtain a single realization of the aggregate deviation: $\delta X = \sum_i \delta X_{i,t_i}$

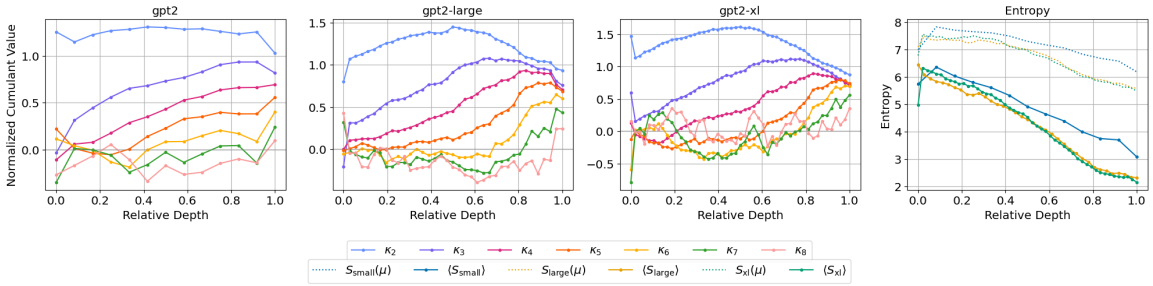

Figure 6: **Cumulants and Entropy Across Model Sizes.** Left three panels: cumulants of structured prompts across layers for three models from the GPT-2 family of increasing size (small, large, xl). Right panel: comparing the mean softmax entropy and the softmax entropy of the center across the same models. The structured prompts exhibit more pronounced cumulant variation and a larger entropy gap, consistent across model size.

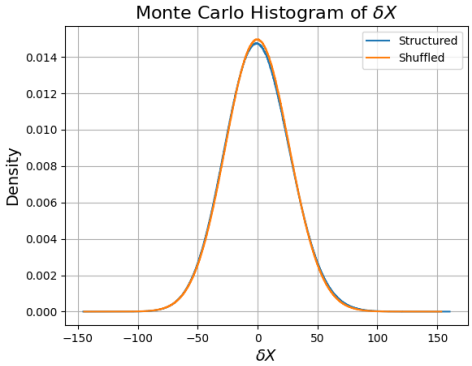

Figure 7: **Histogram of δX.** We use $N = 2 \times 10^8$ Monte Carlo simulations for prompt 3218 at layer 20 of GPT2-Large for the structured (blue) and the shuffled (orange) prompts.

We plot the histogram of δX obtained from $N = 2 \times 10^8$ Monte Carlo simulations for both the structured and shuffled versions of prompt 3218 from the Pile-10K dataset, using the GPT2-Large model at layer 20 (Figure 7). We compare the cumulant values from the two methods in Table 1.

Type	**2nd Cumulant**		**3rd Cumulant**		**4th Cumulant**	
	Monte Carlo	Average	Monte Carlo	Average	Monte Carlo	Average
Structured	1.432	1.432	0.986	0.981	0.623	0.616
Shuffled	1.351	1.430	0.238	0.294	-0.015	-0.128

Table 1: **Monte Carlo vs. Token-Wise Averaged Cumulants.** Comparison of cumulant estimates obtained via Monte Carlo sampling and averaging cumulants over tokens for structured and shuffled prompts. The columns correspond to the (2nd, 3rd, 4th) cumulants.

Appendix C. Cumulant calculation

Algorithm 1: Layer-wise Cumulant Computation for a Text Prompt

Input: Language Model f with L layers, prompt P with T tokens, max cumulant order K.
Output: Cumulants $\kappa_n^{(\ell)}$ for $n = 1, \ldots, K$ and $\ell = 1, \ldots, L$.
Tokenize P into $w_{1:T}$ and run forward pass of f on $w_{1:T}$;
Extract for each layer $\ell = 1, \ldots, L$ the logits from TunedLens $X_t^{(\ell)} \in \mathbb{R}^V$ at each position t;
for $\ell \leftarrow 1$ **to** L **do**
  for $t \leftarrow 1$ **to** T **do**
    $p_t^{(\ell)} \leftarrow \mathrm{softmax}\big(X_t^{(\ell)}\big)$;
  end
  $p_\mu^{(\ell)} \leftarrow \frac{1}{T} \sum_{t=1}^{T} p_t^{(\ell)}$; `// Compute center distribution` p_μ
  $\mu^{(\ell)} \leftarrow \log p_\mu^{(\ell)}$;
  for $t \leftarrow 1$ **to** T **do**
    $\delta X_t^{(\ell)} \leftarrow X_t^{(\ell)} - \mu^{(\ell)}$;
  end
  `// Compute token-wise moments and cumulants`
  for $t \leftarrow 1$ **to** T **do**
    for $n \leftarrow 1$ **to** K **do**
      $m_n^{(\ell)}(t) \leftarrow \sum_{v=1}^{V} p_t^{(\ell)}(v) \left[\delta X_t^{(\ell)}(v)\right]^n$;
      `// Obtaining cumulants from moments where` $B_{n,k}$ `are incomplete Bell polynomials.`
      $\kappa_n^{(\ell)}(t) = \sum_{k=1}^{n} (-1)^{k-1}(k-1)! B_{n,k}\left(m_1^{(\ell)}(t), \ldots, m_n^{(\ell)}(t)\right)$;
    end
  end
  `// Average cumulants across tokens`
  for $n \leftarrow 1$ **to** K **do**
    $\kappa_n^{(\ell)} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \kappa_n^{(\ell)}(t)$;
  end
end
return $\{\kappa_n^{(\ell)}\}_{\substack{\ell=1..L \\ n=1..K}}$;