A Unifying Theory of Signal Propagation in Deep Transformers

Alessio Giorlandino
International School for Advanced Studies
Trieste, Italy
agiorlan@sissa.it

Sebastian Goldt
International School for Advanced Studies
Trieste, Italy
sgoldt@sissa.it

Abstract

Finding the right initialisation for neural networks is crucial to ensure smooth training and good performance. In transformers, the wrong initialisation can lead to one of two failure modes of self-attention layers: rank collapse, where all tokens collapse into similar representations, and entropy collapse, where highly concentrated attention scores lead to training instability. While previous work has studied different scaling regimes for transformers, an asymptotically exact, down-to-the constant prescription for how to initialise transformers has so far been lacking. Here, we provide an analytical theory of signal propagation through deep transformers with self-attention, layer normalisation, skip connections and MLP. Our theory yields a simple algorithm to compute trainability diagrams that identify the correct choice of initialisation hyper-parameters for a given architecture. We overcome the key challenge, an exact treatment of the self-attention layer, by establishing a formal parallel with the Random Energy Model from statistical physics. We also analyse gradients in the backward path and determine the regime where gradients vanish at initialisation. We demonstrate the versatility of our framework through three case studies. Our theoretical framework gives a unified perspective on the two failure modes of self-attention and gives quantitative predictions on the scale of both weights and residual connections that guarantee smooth training.

1 Introduction

In Transformers [Vaswani et al., 2017], where fully-connected layers alternate with self-attention layers [Bahdanau et al., 2015], the key quantity for measuring information flow through the network is the similarity between tokens in a sequence as it propagates through the network. Signal propagation faces additional challenges due to two key failure modes of self-attention layers. The first is *rank collapse*, where self-attention maps all input tokens to identical representations, producing an output matrix of rank one. This phenomenon manifests as the attention pattern shown in fig. 1(a). Dong et al. [2021] demonstrated that networks composed solely of self-attention layers will inevitably collapse any input sequence into uniform token representations at a rate that is double exponential in the number of layers. This rank collapse fundamentally destroys input sequence information and impedes effective training by inducing vanishing gradients [Noci et al., 2022]. *Entropy collapse* represents the second failure mode, where queries attend to only a small, frozen subset of tokens irrespective of the input, leading to low Shannon entropy of the attention distribution (hence the name) and, more importantly, unstable training [Zhai et al., 2023]. An attention matrix exhibiting this pathology is shown in fig. 1(b).

Previous work has highlighted the importance of skip connections in mitigating rank collapse [Dong et al., 2021, Noci et al., 2022, Wang et al., 2024], and Zhai et al. [2023] suggested a modification of the self-attention layer that helps avoiding entropy collapse. However, a quantitative, unified description of how these two distinct phenomena emerge has been lacking. Our result 1 fills this gap,

EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM).

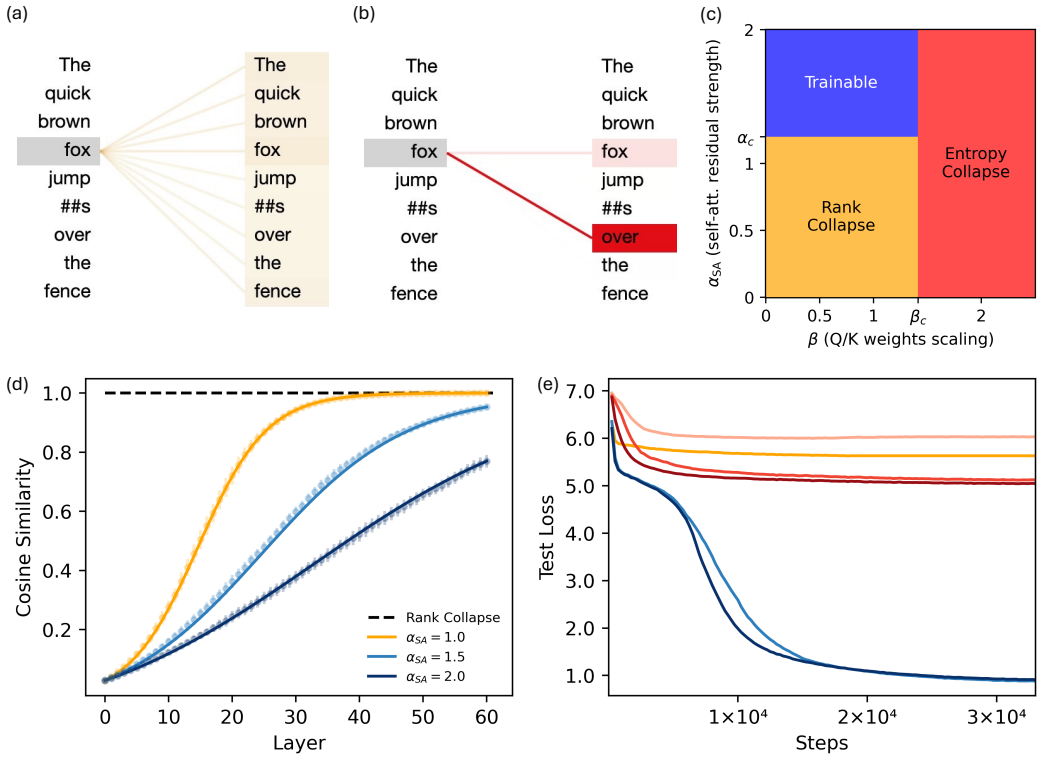

Figure 1: **Two failure modes of Transformers at initialisation, and how to avoid them. (a)** Rank collapse occurs when the self-attention layer attends uniformly to all tokens, mapping all input tokens into the same output token. **(b)** Entropy collapse is a regime of highly saturated attention matrices which attend to random, semantically meaningless patterns, leading to training instability [Zhai et al., 2023]. **(c)** Trainability diagram for a 60-layer BERT transformer, obtained from our analytical theory of signal propagation, see algorithm 1. Depending on the strength of the self-attention residual connections α_{SA} (eq. (8)) and the scale of initial key and query weights β (defined in result 1), we delineate the three regimes of rank collapse, entropy collapse, and the regime where the Transformer is trainable (blue). **(d)** Average cosine similarity between token embeddings of a sequence taken from the TinyStories dataset as it propagates through the layers of a vanilla BERT model for different self-attention residual strengths; empirical measurements (dots) closely follow theoretical predictions (solid lines). Sufficiently large residual connections α_{SA} are key to preventing the similarity between tokens from becoming unity, which would indicate rank collapse. **(e)** Test loss of a 60-layer BERT model on TinyStories for two initialisations from each regime. Models suffering from rank or entropy collapse at initialisation fail to train, as predicted by theory. Full experimental details in section C.1.

providing clear guidance on how to jointly avoid both issues by appropriately choosing the strength of residual connections and the scale of initial weights.

The main challenge in studying information propagation in Transformers arises from the self-attention layers. Previous works either assume uniform attention [Noci et al., 2022] or approximate the average behaviour of a self-attention layer by taking expectations separately over the numerator and denominator of the softmax [Cowsik et al., 2024] – a strong simplification that fails the behaviour of large initial query and key weights, which is responsible for entropy collapse.

In contrast, we analyse the standard softmax transformer in the complementary limit of infinitely long sequences by leveraging tools and concepts from statistical physics. Our theoretical framework provides a unified explanation for – and a practical solution to – the emergence of both failure modes observed in practice: rank collapse and entropy collapse.

Our main contribution is an analytical theory of signal propagation in deep transformers with self-attention layers, skip connections, layer normalisation, and MLPs. Our theoretical framework

yields a simple algorithm to compute the evolution of the typical overlap between token embeddings as sequences propagate through deep, off-the-shelf transformers at initialisation. This enables us to construct trainability diagrams such as the one shown in fig. 1(c) for a 60-layer BERT-style transformer. By varying the strength the scale of initial key and query weights β (defined in result 1) and the strength of the skip connections of the self-attention layer α_{SA}, eq. (8), we identify three regimes for signal propagation: entropy collapse dominance (red), rank collapse dominance (yellow), and a trainable regime characterised by small initial weights and strong skip connections (blue). The critical threshold for query/key variance, β_c, emerges as a global property, while the critical threshold α_c for residual strength depends on network depth: for any given model depth, our theory predicts the minimum residual strength required to guarantee signal propagation and ensure trainability.

We validate our theory in two ways. In fig. 1(d), we show that our theory (solid lines) accurately predicts the average cosine similarity between tokens (averaged over all token pairs in a sequence) when propagating sequences from the TinyStories dataset [Eldan and Li, 2023] through a vanilla BERT model [Devlin et al., 2019] at initialisation (dots). In fig. 1(e), we show the test loss of a 60-layer BERT model on TinyStories for two initialisations from each regime; models suffering rank or entropy collapse at initialisation indeed fail to train, as the trainability diagram predicts. We provide further applicaitons of our theory in section 3.

2 A theory for signal propagation in Transformers

The main challenge in analysing information flow through a transformer is handling the self-attention mechanism with its strong non-linearity and its normalisation step. In this section, we derive a theory for signal propagation by leveraging a formal similarity between self-attention and the *Random energy model* of Derrida [1981]. Using tools and concepts from statistical physics, we give an asymptotically precise characterisation of signal propagation in self-attention layers in the forward pass in section 2.1. We analyse the backward pass in section B.4. Full details of the setup are in section A.

2.1 Forward Signal Propagation through a self-attention layer

Assuming i.i.d. token embeddings, standard concentration results in high dimensions imply that their norms concentrate, while pairwise overlaps scale as $O(d^{-1/2})$. Under self-attention, these tokens mix and their overlaps evolve as the sequence propagates in depth. Two key quantities to describe the evolution of **token similarity** are the overlap and cosine similarity matrices

$$q_{ts} = \frac{1}{d} X_t \cdot X_s \qquad \rho_{ts} = \frac{q_{ts}}{\sqrt{q_{tt} q_{ss}}}, \tag{1}$$

which measure the degree of alignment between two tokens in a sequence. The third important quantity is the inverse participation ratio (IPR) of an attention row, which is defined for all $t \in T$ and $r \in \mathbb{N}$ as

$$Y_t^{(r)} = \sum_{s=1}^{T} A_{ts}^r. \tag{2}$$

For $r = 2$, the IPR estimates the effective number of tokens receiving attention. A small IPR, scaling as $o_T(1)$, indicates attention is evenly spread out over keys, while an IPR that is non-vanishing indicates that only $O_T(1)$ keys are relevant, resulting in sparse self-attention. The connection with Shannon entropy is direct: the IPR captures the collapse of entropy to zero. Higher values of r make the distinction between localised and delocalised attention vectors even sharper.

By tracking the average norm $\mathbb{E}\langle q_{tt} \rangle := q$ and the average overlap $\mathbb{E}\langle q_{ts} \rangle := p$, we describe the evolution of the average cosine similarity ρ and characterize the average IPR of an attention row, which serves as an indicator of entropy collapse. We state here the result and provide a derivation in section B.2.

Result 1 (Average Cosine Similarity Update under Self-Attention) *Let W_Q and W_K be initialised with i.i.d. entries with variance $\sigma_Q^2 = \sigma_K^2 = \beta\sqrt{\log T}/d$, and W_V with variance $\sigma_V^2 = \sigma_v^2/d$.*

For a sequence with average token norm q and average pairwise overlap p, define the critical initialisation scale $\beta_c(q,p) \equiv \sqrt{\frac{2}{q(q-p)}}$. In the limit of infinite sequence length $T \to \infty$, we then have that:

1. *The evolution of the average cosine similarity ρ takes the form:*

$$\Phi_{\mathcal{S}}(\rho) = \frac{\rho}{(1-\rho)Y^{(2)}(\beta) + \rho} = \begin{cases} 1, & \beta < \beta_c(q,p), \\ \frac{\rho}{1-\beta^{-1}\sqrt{2(1-\rho)}}, & \beta > \beta_c(q,p). \end{cases} \tag{3}$$

2. *The average inverse participation ratio (IPR) $Y_t^{(2)}$ satisfies $\forall t \in [T]$:*

$$\lim_{T\to\infty} \mathbb{E}Y_t^{(2)} = Y^{(2)}(\beta) = \begin{cases} 0, & \beta < \beta_c(q,p) \\ 1 - \frac{\beta_c(q,p)}{\beta}, & \beta > \beta_c(q,p) \end{cases} \tag{4}$$

2.1.1 Discussion

For small initialisation $\beta < \beta_c$, self-attention operates in a *'spread attention' phase* where the attention layer effectively outputs the average of all tokens, making the pair-wise cosine similarity between tokens saturate at one, resulting in a rank-one representation matrix. In the absence of skip connections, this behaviour leads to rank collapse, which makes the transformer untrainable. This behaviour is also reminiscent of the "clustering" property analysed by Geshkovski et al. [2023], Bruno et al. [2025], Chen et al. [2025].

For $\beta > \beta_c$, the self-attention layer preserves diversity, and hence information, among input tokens, suggesting this regime as a viable initialisation. However, the non-vanishing value of the IPR for $\beta > \beta_c$, eq. (4), reveals that self-attention is in a *localised phase* in this regime: it only attends to a few tokens which are determined by initialisation, rather than learnt, leading to the training instabilities observed by Zhai et al. [2023]. In other words, for $\beta > \beta_c$ self-attention suffers from entropy collapse, which cannot be avoided using skip connections. This behaviour is a central novelty of our analysis: Noci et al. [2022] assumes the model operates entirely in the rank-collapse phase, thereby overlooking this distinct failure mode, while the "annealed" approximation of Cowsik et al. [2024] does not capture the large-deviation behaviour underlying entropy collapse.

In a nutshell, result 1 implies that the only viable initialisation for self-attention is in the small-variance regime $\beta < \beta_c$ with skip connections to maintain information flow.

Overall, this result unifies two previously observed phenomena, namely *rank* and *entropy collapse*, within a single theoretical framework: the onset of either phenomenon is separated by a sharp phase transition, governed by the variance of the query and key weight initialisation, as parametrised by β.

This framework can also be extended to describe backward signal propagation (see section B.4). Furthermore, integrating the remaining components of a standard transformer block is straightforward (see section B.6). This provides a convenient way to test signal propagation under simple architectural modifications. For instance, we examine a straightforward variant—the gain-controlled transformer—which mitigates both forms of collapse (see section 3).

3 Applications

To showcase the versatility of our approach, we consider three case studies: signal propagation in a standard BERT architecture; comparing different placements of LayerNorm; and comparing variations of the self-attention mechanism itself.

Signal propagation in a vanilla transformer We first used algorithm 1 to analyse signal propagation in a BERT-style transformer [Devlin et al., 2019]. Since BERT uses the post-norm convention for LayerNorm, we state the algorithm 1 for the post-norm architecture; see algorithm 2 for the pre-norm version. The algorithm states the update for the average norm q and average overlap p; the average cosine similarity can be read off after each block simply as $\rho = p/q$. Iterating the algorithm for different values of β and α_{SA} finally yields the trainability diagram shown in fig. 1(d).

Placement of LayerNorm Xiong et al. [2020] showed that placing LayerNorm before the self-attention layer and before the MLP greatly stabilises the training of deep transformers. Comparing signal propagation using our algorithm, we show in fig. 2 that rank collapse, corresponding to an average token similarity of $\langle \rho \rangle = 1$, does indeed occur much later for pre-LN than for post-LN, confirming pre-LN as the more stable choice.

Algorithm 1 Post-norm Block Update

Inputs: $\beta, q, p, \alpha_{\text{SA}}, \alpha_{\text{MLP}}, \sigma_w^2, \sigma_b^2, \sigma_v^2$
▷ Attention layer + residual
$\beta_c \leftarrow \sqrt{\frac{2}{q(q-p)}}$
$Y^{(2)}(\beta) \leftarrow \max(0, 1 - \beta_c/\beta)$
$q \leftarrow \sigma_v^2 \left(p + (q-p) \cdot Y^{(2)}(\beta, q, p) \right) + q \cdot \alpha_{\text{SA}^2}$
$p \leftarrow p \cdot (\sigma_v^2 + \alpha_{\text{SA}}^2)$
▷ Post-norm LN
$p \leftarrow p/q; \quad q \leftarrow 1$
▷ MLP + residual
$q_1 \leftarrow \sigma_w^2 q + \sigma_b^2; \quad p_1 \leftarrow \sigma_w^2 p + \sigma_b^2$
$q_2 \leftarrow \frac{\sigma_w^2}{2} q_1 + \sigma_b^2; \quad p_2 \leftarrow \frac{\sigma_w^2}{2} f(p_1/q_1) q_1 + \sigma_b^2$
$q \leftarrow q_2 + \alpha_{\text{MLP}}^2 q; \quad p \leftarrow p_2 + \alpha_{\text{MLP}}^2 p$
▷ Post-norm LN
$p \leftarrow p/q; \quad q \leftarrow 1$
return (q, p)

Avoiding All Collapses: gain-controlled attention A recent line of work has sought to alleviate rank collapse by directly modifying the self-attention layer itself. Noci et al. [2023], Naderi et al. [2024] proposed to enforce the attention layer to be a perturbation of the identity. Noci et al. [2023] derive an SDE description of the limiting distribution of the overlap (or neural covariance) matrix in the proportional limit where both width and depth go to infinity. To obtain this limit, they introduce a width-dependent temperature parameter and remove layer normalisation. However, layer normalisation is crucial to avoid entropy collapse, see result 1.

Using our framework, we can show that simply removing the mean of the values along the sequence from the output of the standard self-attention layer, also explored in [Naderi et al., 2024] and reminiscent of gain control in neuroscience, can be combined to great effect with either post-LN or pre-LN.

In fig. 2, we show the theoretical prediction for the evolution of the average cosine similarity, illustrating how this modification alleviates rank collapse. Our preliminary experiments with a twenty-layer BERT-style transformer trained on TinyStories show that gain-controlled transformers succeed in regimes where vanilla attention fails, see fig. 5, encouraging further experiments at scale which are however out of the scope for the present paper. As a final note, in order to propagate the signal infinitely deep, one should simply use gain-controlled attention and initialise the MLPs at the edge of chaos.

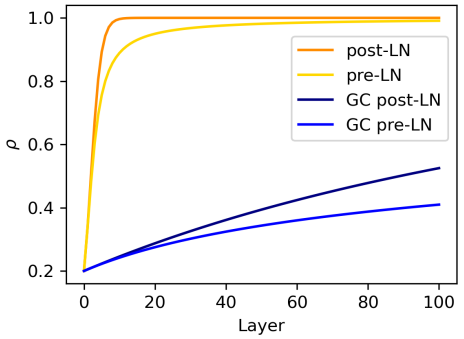

Figure 2: Theoretical prediction of the evolution with depth of the average cosine similarity for the standard transformer and the Gain-controlled Transformer under both LN strategies. Rank collapse is avoided simply by removing the mean value in the self-attention layer. Here, we set $\alpha_{\text{SA}} = \alpha_{\text{MLP}} = 1$.

4 Conclusions

We developed a theory for signal propagation in transformers that unifies the understanding of the two main failure modes in transformer training: rank collapse and entropy collapse. Our framework not only predicts when these phenomena occur for a given set of hyperparameters, but it also provides simple, flexible algorithms for deriving the trainability diagrams of a given architecture. Building on these insights, we find new evidence for the viability of a simple architectural modification of self-attention, the gain-controlled self-attention, that avoids both failure modes, which would be interesting to explore at scale in future work.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference of Learning Representations*, 2015.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse, 2022. URL `https://arxiv.org/abs/2206.03126`.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse, 2023. URL `https://arxiv.org/abs/2303.06296`.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Aditya Cowsik, Tamra Nebabu, Xiao-Liang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *arXiv preprint arXiv:2403.02579*, 2024.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24:2613–2626, 1981. URL `https://api.semanticscholar.org/CorpusID:122288449`.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models, 2025. URL `https://arxiv.org/abs/2410.23228`.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pages 10524–10533. PMLR, 2020.

Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36:54250–54281, 2023.

Alireza Naderi, Thiziri Nait Saada, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in transformers. *arXiv preprint arXiv:2410.07799*, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford Graduate Texts. OUP Oxford, 2009. ISBN 9780198570837. URL `https://books.google.it/books?id=jhCM7i0a6UUC`.

Gérard Ben Arous, Leonid V Bogachev, and Stanislav A Molchanov. Limit theorems for sums of random exponentials. *Probability theory and related fields*, 132(4):579–612, 2005.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL https://www.aclweb.org/anthology/P19-3007.

A Setup and notation

Goal We consider a vanilla Transformer encoder [Vaswani et al., 2017] that processes sequences $\{X_t\}_{t=1,\dots,T}$ of T tokens embedded in a d-dimensional space. We analyse signal propagation through a complete Transformer block comprising self-attention, residual connections, layer normalisation, and a feed-forward MLP.

Layer Norm. We consider layer normalisation [Ba et al., 2016] which centers each token embedding and rescales it by its standard deviation; for simplicity, we omit the affine transformation:

$$\mathrm{LN}(X_t) = \frac{X_t}{\sqrt{\mathrm{Var}[X_t]}} = \frac{X_t}{d^{-1/2}\|X_t\|}, \tag{5}$$

where $\|X_t\|$ is the Euclidean norm. We consider both pre-norm and post-norm variants.

Self-Attention. Given the query, key and value weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$, the attention score between tokens X_t and $X_{t'}$ is

$$a_{tt'} = \frac{1}{\sqrt{d}}(W_Q X_t)^\top (W_K X_{t'}). \tag{6}$$

These scores are normalised via a softmax operation to obtain the attention weights, which are then used to compute the weighted sum of the value projections:

$$A_{tt'} = \frac{e^{a_{tt'}}}{\sum_{\tau=1}^{T} e^{a_{t\tau}}}, \qquad \mathcal{S}(X)_t = \sum_{t'=1}^{T} A_{tt'} W_V X_{t'}. \tag{7}$$

Residual Connections. We write the residual connections for the self-attention and MLP blocks as

$$\mathrm{RES}_{\mathrm{SA}}(X) = \mathcal{S}(X) + \alpha_{\mathrm{SA}} X, \qquad \mathrm{RES}_{\mathrm{MLP}}(X) = \mathrm{MLP}(X) + \alpha_{\mathrm{MLP}} X, \tag{8}$$

where $\alpha > 0$ are the strenghts of the skip connections.

Notation We denote the average over tokens or token pairs with brackets $\langle\cdot\rangle$, whilst $\mathbb{E}[\cdot]$ denotes the average with respect to the initialisation of the parameters W_Q, W_K, W_V.

B Theory Appendix

B.1 Attention Scores Are Correlated Gaussian Variables.

Consider the attention scores defined in eq. (6). By the Central Limit Theorem, they are converge in distribution to Gaussian random variables with zero mean and variance σ_a^2, where d is the embedding dimension (or head dimension in the multi-head case), and σ_a^2 is determined by the initialisation of W_Q and W_K as described in section A. Although the scores $a_{tt'}$ are individually Gaussian, they are not independent; in fact, they are correlated. To quantify these correlations, we compute:

$$\mathrm{Cov}(a_{ts}, a_{\tau\sigma}) = \frac{1}{d} \sum_{i,j,k,l,m,n=1}^{d} X_{ti} X_{sk} X_{\tau l} X_{\sigma n}\, \mathbb{E}[(W_Q)_{ji}(W_Q)_{ml}]\, \mathbb{E}[(W_K)_{jk}(W_K)_{mn}].$$

Since the query and key weights are independently initialised with variances $\sigma_Q^2 = \sigma_K^2 = \sigma_a^2/d$, this simplifies to:

$$\mathbb{E}\left[a_{ts} a_{\tau\sigma}\right] = \sigma_Q^2 \sigma_K^2 \,(X_t \cdot X_\tau)(X_s \cdot X_\sigma) = \sigma_a^2 q_{t\tau} q_{s\sigma}. \tag{9}$$

B.2 Derivation of result 1

B.2.1 Computation of $Y^{(2)}(\beta)$

Consider the partecipation ratio of the t-th row of a self-attention matrix average over the initialisation $W_Q,\ W_K$.

$$\mathbb{E}Y_t^{(2)} = \mathbb{E}\frac{\sum_s e^{2a_{ts}}}{\left(\sum_s e^{a_{ts}}\right)^2}$$

We also define

$$\Phi_t(\beta, h) = \mathbb{E} \log Z_t(\beta, h) \tag{10}$$

where

$$Z_t(\beta, h) = \sum_{s=1}^{T} e^{h a_{ts}}, \quad h \in \mathbb{R}$$

and we recall that β enters in the definition of the covariances between the a's, as they are all proportional to $\sigma_a^2 = \beta^2 \log T$.
We want to exploit Stein's lemma, which yields:

$$\begin{aligned} \partial_h \Phi_t(\beta, h) &= \mathbb{E}\left[\frac{\sum_s a_{ts} e^{h a_{ts}}}{\sum_u e^{h a_{tu}}}\right] \\ &= \sum_{s,s'=1}^{T} \mathbb{E}[a_{ts} a_{ts'}]\, \mathbb{E}\left[\partial_{a_{ts'}}\left(\frac{e^{h a_{ts}}}{\sum_u e^{h a_{tu}}}\right)\right] \end{aligned} \tag{11}$$

with the correlation between attention scores given by eq. (9). We assume $q_{tt} \simeq q$ for all t, due to concentration of measure. For the pairwise overlaps, we proceed as follows: at the first layer, all tokens are approximately orthogonal, making it safe to assume $q_{ts} \simeq 0 \; \forall t \neq s \in [T]$. One then derives the update for this layer and observes that, in the limit of infinite sequence length (as we will briefly show), the update becomes independent of the indices to leading order. By repeating this argument across layers, it is therefore justified to treat the q_{ts} as having a common mean p together with sub-leading Gaussian fluctuations, which we neglect since our analysis focuses on the average overlap. Applying this argument, we get:

$$\begin{aligned} \partial_h \Phi_t(\beta, h) &\simeq h\sigma_a^2 q \sum_s \mathbb{E}\left(q\frac{e^{h a_{ts}}}{\sum_u e^{h a_{tu}}} - q\frac{e^{2h a_{ts}}}{\left(\sum_u e^{h a_{tu}}\right)^2} - p\sum_{s'\neq s}\frac{e^{h a_{ts}} e^{h a_{ts'}}}{\left(\sum_u e^{h a_{tu}}\right)^2}\right) \\ &= h\sigma_a^2 q\,(q-p)\left(1 - \mathbb{E}Y_t^{(2)}\right) \end{aligned}$$

Finally, this leads to:

$$\lim_{T\to\infty} \mathbb{E}Y_t^{(2)} = 1 - \frac{1}{\sigma_a^2 q\,(q-p)} \lim_{h\to 1}\lim_{T\to\infty} \partial_h \Phi_t(\beta, h) \tag{12}$$

To proceed, we are left with computing the expectation $\Phi_t(\beta, h) = \mathbb{E}\left[\log \sum_s e^{h a_{ts}}\right]$.

The remaining computation amounts to evaluating a variant of the Random Energy Model (REM), but with correlated energy levels due to the structure of the a_{ts} variables. This problem can be tackled using the Replica method or, alternatively, via a micro-canonical argument, as discussed in Mézard and Montanari [2009]. Here, we proceed with the Replica method.

We compute the replicated partition function as:

$$\begin{aligned} \mathbb{E}_a Z_t^n(\beta, h) = \mathbb{E}_a\left[\left(\sum_s e^{h a_{ts}}\right)^n\right] &= \sum_{s_1,\dots,s_n=1}^{T} \mathbb{E}_a\left[\exp\left(h\sum_{a=1}^{n} a_{ts_a}\right)\right] \\ &= \sum_{s_1,\dots,s_n=1}^{T} \exp\left(\frac{h^2\sigma_a^2}{2}\sum_{a,b=1}^{n}\mathbb{E}[a_{ts_a} a_{ts_b}]\right) \\ &= \sum_{s_1,\dots,s_n=1}^{T} \exp\left(\frac{h^2\sigma_a^2}{2}\left(q\sum_{a,b}\mathbb{I}(s_a = s_b) + qp\sum_{s\neq s'}\sum_{a,b}\mathbb{I}(s = s_a)\mathbb{I}(s' = s_b)\right)\right) \end{aligned} \tag{13}$$

We now introduce the empirical overlap matrix:

$$Q_{ab} = \mathbb{I}(s_a = s_b)$$

and perform a change of variables from replica indices to overlap structures Q, giving:

$$\mathbb{E}_a Z_t^n(\beta, h) = \sum_Q \sum_{\{s_a\}} \prod_{a,b} \delta(Q_{ab}, \mathbb{I}(s_a = s_b)) \exp\left(\frac{h^2\sigma_a^2}{2} q\,(q-p) \sum_{a,b} Q_{ab} + O(n^2)\right)$$
$$= \sum_Q S(Q) \exp\left(\frac{h^2\sigma_a^2}{2} q\,(q-p) \sum_{a,b} Q_{ab} + O(n^2)\right) \tag{14}$$

Now we take the 1-RSB ansatz for Q: the n replicas are diveded into $\frac{n}{x}$ groups of x elements which are in the same energy configurations.
Moreover, we need to consider exponentially long sequences, i.e. we take $T = e^N$ and control N. This implies:

$$S(Q) \simeq e^{N\frac{n}{x}} \qquad \sum_{ab}^{n} Q_{ab} = nx$$

So exploiting the replica trick and recalling the definition of σ_a^2, we get:

$$\Phi_t(\beta, h)/N = \max_{x<1} \frac{\beta^2 h^2}{2} q\,(q-p)\,x + \frac{1}{x} \tag{15}$$

which leads to the existence of a critical temperature $\beta_c(h, q, p) = \frac{1}{h}\sqrt{\frac{2}{q(q-p)}}$ where a condensation phase transition takes place. In particular we have:

$$\Phi_t(\beta, h)/N = \begin{cases} 1 + \frac{\beta^2 h^2}{2} q\,(q-p) & \beta < \beta_c(h, q, p) \\ \beta h \sqrt{2q\,(q-p)} & \beta > \beta_c(h, q, p) \end{cases}$$

Due to the fact that we took the expectation over all the a's there is no dependence on v, rather only on the similarity matrix, and so we can drop the subscript from Φ_t. Finally we can put all together, and we get:

$$\lim_{T\to\infty} \mathbb{E} Y_t(\beta) = 1 - \frac{1}{\sigma_a^2 q\,(q-p)} \lim_{h\to 1} \lim_{T\to\infty} \partial_h \Phi(\beta, h) := Y^{(2)}(\beta) = \begin{cases} 0 & \beta < \beta_c(q, p) \\ 1 - \frac{\beta_c(q,p)}{\beta} & \beta > \beta_c(q, p) \end{cases}$$

where $\beta_c(q, p) = \beta_c(h = 1, q, p) = \sqrt{\frac{2}{q(q-p)}}$. We plot in fig. 3 the result of the computation and some numerical simulations.

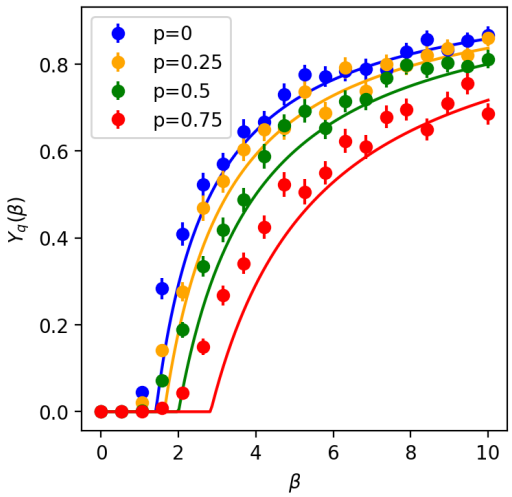

Figure 3: Theory and experiments ($T = 10^5$) comparison of the computation of $Y^{(2)}(\beta)$, finite size effects are visible around the phase transition.

B.2.2 Computation of $Y_p(\beta)$

We need to compute:

$$Y_p(\beta) = \mathbb{E}\frac{\sum_s e^{a_{vs}+a_{us}}}{\sum_s e^{a_{vs}}\sum_{s'} e^{a_{us'}}}$$

Consider the partition function with auxiliary fields $\mathbf{h} = (h_{ss'})_{s,s'=1}^T$:

$$Z_{v,u}(\beta, \mathbf{h}) = \sum_{ss'} e^{h_{ss'}(a_{vs}+a_{us'})}$$

Let's observe that:

$$\begin{aligned}\sum_\sigma \partial_{h_{\sigma\sigma}} \mathbb{E}\log Z_{v,u}(\beta,\mathbf{h}) &= \mathbb{E}\left[\frac{\sum_\sigma (a_{v\sigma}+a_{u\sigma})e^{h_{\sigma\sigma}(a_{v\sigma}+a_{u\sigma})}}{\sum_{s,s'} e^{h_{ss'}(a_{vs}+a_{us'})}}\right] \\ &= \mathbb{E}\left[\frac{\sum_\sigma (a_{v\sigma}+a_{u\sigma})e^{h_{\sigma\sigma}(a_{v\sigma}+a_{u\sigma})}}{\sum_\sigma e^{h_{\sigma\sigma}(a_{v\sigma}+a_{u\sigma})}} \cdot \frac{\sum_\sigma e^{h_{\sigma\sigma}(a_{v\sigma}+a_{u\sigma})}}{\sum_{s,s'} e^{h_{ss'}(a_{vs}+a_{us'})}}\right] \end{aligned} \tag{16}$$

Assuming the first term is self-averaging, we approximate:

$$\sum_\sigma \partial_{h_{\sigma\sigma}} \mathbb{E}\log Z_{v,u}(\beta,\mathbf{h}) \simeq \mathbb{E}\left[\frac{\sum_\sigma (a_{v\sigma}+a_{u\sigma})e^{h_{\sigma\sigma}(a_{v\sigma}+a_{u\sigma})}}{\sum_\sigma e^{h_{\sigma\sigma}(a_{v\sigma}+a_{u\sigma})}}\right] \cdot \mathbb{E}\left[\frac{\sum_\sigma e^{h_{\sigma\sigma}(a_{v\sigma}+a_{u\sigma})}}{\sum_{s,s'} e^{h_{ss'}(a_{vs}+a_{us'})}}\right]$$

Under this approximation:

$$Y_p(\beta) = \frac{\lim_{\mathbf{h}\to\mathbf{1}} \sum_\sigma \partial_{h_{\sigma\sigma}} \mathbb{E}\log \sum_{s,s'} e^{h_{ss'}(a_{vs}+a_{us'})}}{\lim_{h\to 1} \partial_h \mathbb{E}\log \sum_s e^{h(a_{vs}+a_{us})}} \tag{17}$$

The computation of the free entropy in the denominator is straightforward. It closely resembles the derivation for the free entropy appearing in the calculation of $Y_q(\beta)$.

One gets:

$$\frac{1}{N}\mathbb{E}\log\sum_s e^{h(a_{vs}+a_{us})} = \begin{cases} 1 + \left(1-\frac{p^2}{q^2}\right)\beta^2 h^2 & \beta < \frac{1}{h\sqrt{1-\frac{p^2}{q^2}}} \\ 2\beta h\sqrt{1-\frac{p^2}{q^2}} & \beta > \frac{1}{h\sqrt{1-\frac{p^2}{q^2}}} \end{cases} \tag{18}$$

which gives

$$\lim_{h\to 1}\partial_h \mathbb{E}\log\sum_s e^{h(a_{vs}+a_{us})} = \begin{cases} 2\beta^2\left(1-\frac{p^2}{q^2}\right) & \beta < \frac{1}{\sqrt{1-\frac{p^2}{q^2}}} \\ 2\beta\sqrt{1-\frac{p^2}{q^2}} & \beta > \frac{1}{\sqrt{1-\frac{p^2}{q^2}}} \end{cases} \tag{19}$$

The calculation of the free entropy at the numerator is a bit more involved, but we can observe the following:

$$\begin{aligned}\lim_{\mathbf{h}\to\mathbf{1}} \sum_\sigma \partial_{h_{\sigma\sigma}} \mathbb{E}\log \sum_{s,s'} e^{h_{ss'}(a_{vs}+a_{us'})} &= \frac{\sum_s (a_{vs}+a_{us})e^{a_{vs}+a_{us}}}{\sum_{s,s'} e^{a_{vs}+a_{us'}}} \\ &= \frac{\sum_{s,s'}(a_{vs}+a_{us'})e^{a_{vs}+a_{us'}} - \sum_{\substack{s,s'=1 \\ s\neq s'}}(a_{vs}+a_{us'})e^{a_{vs}+a_{us'}}}{\sum_{s,s'} e^{a_{vs}+a_{us'}}}\end{aligned}$$

We define:

$$\langle a_{vs} \rangle_s = \frac{\sum_s a_{vs} e^{a_{vs}}}{\sum_s e^{a_{vs}}}$$

$$\langle (a_{vs} + a_{us'}) \rangle_{s \neq s'} = \frac{\sum_{\substack{s,s'=1 \\ s \neq s'}} (a_{vs} + a_{us'}) e^{a_{vs} + a_{us'}}}{\sum_{\substack{s,s'=1 \\ s \neq s'}} e^{a_{vs} + a_{us'}}}$$

Thus, we obtain the approximation:

$$\lim_{\mathbf{h} \to \mathbf{1}} \sum_{\sigma} \partial_{h_{\sigma\sigma}} \mathbb{E} \log \sum_{s,s'} e^{h_{ss'}(a_{vs} + a_{us'})} \leq \langle a_{vs} \rangle_s + \langle a_{us} \rangle_s - \langle (a_{vs} + a_{us'}) \rangle_{s \neq s'} \simeq 0$$

There is also an intuitive way to see this. Consider the two limiting cases:

- As $\beta \to 0$: The attention weights become uniform, i.e., $A_{vs} \to \frac{1}{T}$. Then,

  $$\frac{1}{d} \mathbb{E}_{QKV} \left[\mathcal{S}(X)_v \cdot \mathcal{S}(X)_w \right] = \mathbb{E}_{KQ} \sum_{s,\sigma} A_{vs} A_{w\sigma} S_{s\sigma} \to \frac{1}{T^2} \sum_{s,\sigma} S_{s\sigma} = \frac{p}{q} + \mathcal{O}(T^{-1}),$$

  meaning both q^{att}, $p^{\text{att}} \to \frac{p}{q}$.
- As $\beta \to \infty$: The attention becomes fully peaked:

  $$A_{vs} \to \delta_{v,s^*(v)}, \quad \text{with} \quad s^*(v) = \arg\max_s a_{vs}.$$

  In this limit, the dot product becomes:

  $$q^{\text{att}} \to S_{s^* s^*} = 1, \quad p^{\text{att}} \to S_{s^*(v), \sigma^*(w)} = \frac{p}{q},$$

  since $s^*(v) \neq \sigma^*(w)$ with probability $\pi = 1 - 1/T$.

Hence, as we vary β from 0 to ∞, the quantity $\text{SA}(q)$ interpolates between $\frac{p}{q}$ and 1, while $\text{SA}(p)$ remains nearly constant.

B.2.3 Update map of the average average cosine similarity.

We begin our analysis by averaging over the value projection matrix W_V. Since W_V is independent from W_Q, Q_K at initialisation, the scalar product between self-attention outputs then becomes:

$$\frac{1}{d} \mathbb{E} \left[\mathcal{S}(X)_t \cdot \mathcal{S}(X)_{t'} \right] = \frac{1}{d} \mathbb{E} \left[\sum_{s,\sigma} A_{ts} A_{t'\sigma} X_s^\top \, \mathbb{E}[W_V^\top W_V] \, X_\sigma \right].$$

Choosing, without loss of generality, the variance of the value projection weights as $\sigma_V^2 = 1/d$, we obtain $\mathbb{E}_V[W_V^\top W_V] = \mathbb{I}_d$, so the expression simplifies to:

$$\mathbb{E} \left[\sum_{s,\sigma} A_{ts} A_{t'\sigma} q_{s\sigma} \right]. \tag{20}$$

With the limit $T \to \infty$ in mind, and since we are concerned only with the average overlap, we neglect sub-leading fluctuations and focus directly on the leading terms:

$$q_{s\sigma} \simeq p \quad \text{for } s \neq \sigma, \qquad q_{ss} \simeq q \,.$$

$$\mathbb{E} \left[q \sum_s A_{ts} A_{t's} + p \sum_{s \neq \sigma} A_{ts} A_{t'\sigma} \right]. \tag{21}$$

Since each row of the attention matrix sums to one, we can simplify this expression further:

$$q \sum_s A_{ts} A_{t's} + p \left(\sum_{s,\sigma} A_{ts} A_{t'\sigma} - \sum_s A_{ts} A_{t's} \right) = (q-p) \sum_s A_{ts} A_{t's} + p.$$

Now, upon averaging and taking the limit $T \to \infty$, two distinct quantities emerge depending on whether $t = t'$ or $t \neq t'$:

$$Y^{(2)}(\beta) = \lim_{T\to\infty} \mathbb{E}\left[\sum_s A_{ts}^2\right], \qquad Y_p(\beta) = \lim_{T\to\infty} \mathbb{E}\left[\sum_s A_{ts} A_{t's}\right] \quad \text{for } t \neq t', \tag{22}$$

which quantify the self- and cross-overlap of attention distributions. The expression for the first one is derived in section B.2.1 and in section B.2.2 we show that the second term is sub-leading.

Substituting the result, we get the update for the average norm:

$$q \xleftarrow{\mathcal{S}} p + (q-p)\, Y_q(\beta) = \begin{cases} p & \beta < \beta_c(q,p) \\ p + (q-p)\left(1 - \frac{\beta_c(q,p)}{\beta}\right) & \beta > \beta_c(q,p) \end{cases} \tag{23}$$

where $\beta_c(q,p) = \sqrt{\frac{2}{q(q-p)}}$.
On the other hand, the scalar product p is not updated, as $Y_p(\beta)$ is sub-leading. Taking the ratio between the updates yields, at leading order in d, the following update for the average cosine similarity:

$$\Phi_{\mathcal{S}}(\rho) = \frac{\rho}{1 + (1-\rho) Y^{(2)}(\beta)}\,.$$

B.3 Finite-Size Effects

Here we give a non-rigorous argument on the finite size effects that afflict our asymptotic theory. In the low-β regime, the attention is spread approximately uniformly over a number $T^* = e^{S(\beta,\rho)}$ of keys, given by an entropic quantity $S(\beta,\rho) = \Phi(\beta,\rho) - \beta\partial_\beta\Phi(\beta,\rho)$ (where the free entropy Φ was defined in eq. (10)). A derivation of the entropy for the REM, very much related to our problem, is explained in depth by Mézard and Montanari [2009]. For $\beta < \beta_c(\rho)$, this turns out to be:

$$S(\beta,\rho) = N\left(1 - \frac{\beta^2}{\beta_c(\rho)}\right) \tag{24}$$

Since the IPR is the inverse number of the expected number of state that matter:

$$Y^{(2)}(\beta,\rho) \simeq e^{-S(\beta,\rho)} \tag{25}$$

In the limit $N = \mathcal{O}(\log T)$, this non-rigorous argument suggests that the corrections to eq. (4) scale are $O\left(T^{-1+\frac{\beta^2}{\beta_c(\rho)^2}}\right)$. As long as $\beta < \tilde{\beta}_c = \beta_c/2$, these fluctuations can be neglected since they remain Gaussian. For $\beta_c/2 < \beta < \beta_c$, however, the fluctuations around the asymptotic solution become non-Gaussian, as the central limit theorem breaks down at $\beta_c/2$ [Ben Arous et al., 2005].

B.4 The backward pass

To complete our theory of signal propagation, we derive the following result on the norm of the gradients of query and key weights at initialisation (see section B.5 for the derivation):

Result 2 *(Query/Key Gradient Analysis)* In the limit $T \to \infty$, under the same hypothesis of result 1, the expected squared Frobenius norm of the query gradient, and analogously for the key, is given by

$$\frac{T}{d^2\sqrt{\log T}} \mathbb{E}\left\| \frac{\partial \mathcal{L}}{\partial W_Q} \right\|_F^2 = C\beta\sigma_v^2\, q(q-p)\Big[(q-p)\big(Y^{(2)} - 2Y^{(3)}\big) + p\big(Y^{(2)}\big)^2\Big], \tag{26}$$

where C is a constant independent of T and d. The same result holds for W_K.

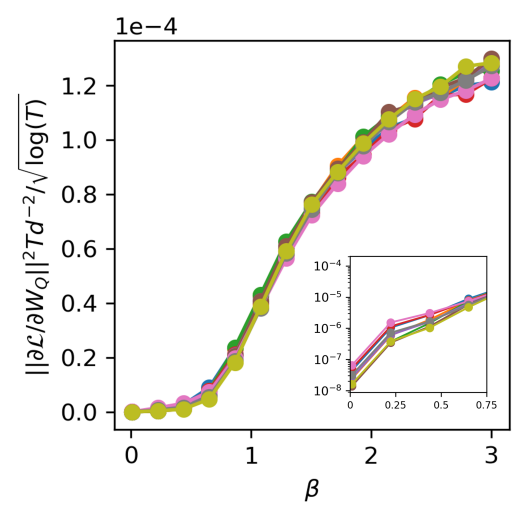

Figure 4: **Norm of the query gradient.** Frobenius norm of the gradient of the loss with respect to query weights for various combinations of sequence length $T = 2048, 4096, 8192$ and embedding dimension $d = 256, 512, 1024$. As predicted by result 2, gradients collapse for different T and d, and vanishing gradients afflict the low-β regime.

Result 2 shows that gradients can vanish under two conditions. First, we see that if $q = p$, i.e. if attention is uniform and maps all input tokens into the same output token, gradients vanish. In this case, we recover the well-known result of Noci et al. [2022] that showed that if the inputs to the self-attention layer are already collapsed, and hence all the tokens are the same, gradients vanish. Result 2 goes beyond their result to show that even if input tokens are diverse, i.e. $p \neq q$, gradients vanish if $\beta < \beta_c$ because in that regime, the inverse participation ratios $Y^{(2)}$ and $Y^{(3)}$ tend to zero as the sequence length $T \to \infty$. For long but finite sequences, finite-size effects reduce the threshold down to $\beta < \tilde{\beta}_c = \beta_c/2$ (see section B.3). We numerically verify result 2 in fig. 4, where we show that (1) the curves of gradient norm versus initialisation strength collapse with the scaling suggested by result 2, and that (2) gradients do indeed tend to zero for $\beta < \tilde{\beta}_c$.

Our analysis of the backward pass raises a paradox: initialise with $\beta < \beta_c$, and gradients vanish; initialise with $\beta > \beta_c$, and self-attention is stuck with entropy collapse. We discussed in the previous section that entropy collapse can only be avoided by initialising with small weights $\beta < \beta_c$ and adding skip connections. So where do the gradients come from? Result 2 depends on the assumption that we are considering i.i.d. embeddings. While this assumption is true at initialisation, the embeddings do receive a non-zero gradient through the skip connections already at initialisation, independently of the value of β. The changes in the embeddings will break the i.i.d. assumption on the embeddings, and hence enable gradients to flow even in the trainable regime of $\beta < \tilde{\beta}_c$.

B.5 Derivation of result 2

Consider square matrices $A, a, Q \in \mathbb{R}^{T\times T}$, where $A = \mathrm{softmax}(a)$ is the standard attention matrix computed from logits a and $Q = (q_{ts})_{(ts)}$ is the overlap matrix. Let I_T denote the $T \times T$ identity matrix. Also, consider matrices $X \in \mathbb{R}^{T\times d}$, $W_V, W_Q, W_K \in \mathbb{R}^{d\times d}$.

Define the attention operation

$$\mathcal{S}(X) = AXW_V. \tag{27}$$

We are interested in computing the squared Frobenius norm of the gradient of $\mathcal{S}(X)$ with respect to the query matrix W_Q:

$$\left\| \frac{\partial \mathcal{S}(X)}{\partial W_Q} \right\|_F^2 = \mathrm{tr}\left(\frac{\partial \mathcal{S}(X)}{\partial W_Q} \left(\frac{\partial \mathcal{S}(X)}{\partial W_Q} \right)^\top \right). \tag{28}$$

The first part of the derivation parallels the proof of a related result in Noci et al. [2022]. However, unlike their approach, we do not assume uniform attention. Instead, we retain the attention explicitly, which allows the previously derived participation ratio to naturally emerge.

Chain rule decomposition:

$$\frac{\partial \mathcal{S}(X)}{\partial W_Q} = \left(I_T \otimes W_V^\top X^\top\right) \frac{\partial A}{\partial a} \left(\frac{1}{\sqrt{d}} X \otimes X W_K\right).$$

Consequently, the squared Frobenius norm equals the trace of

$$\begin{aligned}&\left(I_T \otimes W_V^\top X^\top\right) \frac{\partial A}{\partial a} \left(\frac{1}{\sqrt{d}} X \otimes X W_K\right) \\ &\quad \times \left(\frac{1}{\sqrt{d}} X^\top \otimes W_K^\top X^\top\right) \left(\frac{\partial A}{\partial a}\right)^\top \left(I_T \otimes X W_V\right).\end{aligned}$$

Simplification of the middle terms:

$$\begin{aligned}&\left(\frac{1}{\sqrt{d}} I_T \otimes X W_K\right) \left(\frac{1}{\sqrt{d}} I_T \otimes W_K^\top X^\top\right) \\ &= \frac{1}{d} I_T \otimes \left(X W_K W_K^\top X^\top\right).\end{aligned}$$

Assuming $W_K W_K^\top$ concentrates as

$$W_K W_K^\top \approx d\sigma_K^2 I_d,$$

we obtain

$$\frac{1}{d} X X^\top \otimes \left(X (d\sigma_K^2 I_d) X^\top\right) = \sigma_K^2 \left(X X^\top \otimes X X^\top\right).$$

Taking the trace and using its cyclic property, we define

$$G = \left(\frac{\partial A}{\partial a}\right) \left(X X^\top \otimes X X^\top\right) \left(\frac{\partial A}{\partial a}\right)^\top,$$

and write

$$\left\| \frac{\partial \mathcal{S}(X)}{\partial W_Q} \right\|_F^2 = \sigma_K^2 \operatorname{tr}\left(G \left(I_T \otimes X W_V\right) \left(I_T \otimes W_V^\top X^\top\right)\right).$$

Assuming $W_V W_V^\top$ concentrates as

$$W_V W_V^\top \approx d\sigma_V^2 I_d,$$

we simplify further as

$$\left\| \frac{\partial \mathcal{S}(X)}{\partial W_Q} \right\|_F^2 = d\sigma_K^2 \sigma_V^2 \operatorname{tr}\left(G \left(I_T \otimes X X^\top\right)\right).$$

Recall the definition of the overlap matrix

$$Q = \frac{1}{d} X X^\top \in \mathbb{R}^{T \times T},$$

we get the compact expression

$$\left\| \frac{\partial \mathcal{S}(X)}{\partial W_Q} \right\|_F^2 = d^4 \sigma_K^2 \sigma_V^2 \operatorname{tr}\left(G (I_T \otimes Q)\right),$$

where

$$G = \left(\frac{\partial A}{\partial a}\right)(Q \otimes Q)\left(\frac{\partial A}{\partial a}\right)^{\top}.$$

This expression reveals how the gradient norm depends on the structure of Q, the Jacobian of the attention, and the variance parameters associated with the key and value projections.

Let's compute the trace term:

$$\mathrm{tr}\left(\frac{\partial A}{\partial a}(Q \otimes Q)\left(\frac{\partial A}{\partial a}\right)^{\top}(I_T \otimes Q)\right)$$

where $\frac{\partial A}{\partial a}$ is the Jacobian matrix of size $T^2 \times T^2$. Let's write the Jacobian in components. Using the fact that $A = \text{softmax(a)}$, so the Jacobiam components are:

$$D_{(ij),(kl)} := \frac{\partial A_{ij}}{\partial a_{kl}} = \delta_{ik}\delta_{jl}A_{ij} - \delta_{ik}A_{ij}A_{il}. \tag{29}$$

Now, the trace can be written as

$$\mathrm{tr} = \sum_{i,j,r,s=1}^{T}\Big[\frac{\partial A}{\partial a}(Q \otimes Q)\left(\frac{\partial A}{\partial a}\right)^{\top}(I_T \otimes Q)\Big]_{(ij),(tu)}\delta_{it}\delta_{ju}.$$

Expanding indices leads to

$$\mathrm{tr} = \sum_{i,j,k,l,m,n,r,s,t,u=1}^{T} D_{(ij),(kl)}q_{km}q_{ln}D_{(rs),(mn)}\delta_{rt}q_{su}\delta_{it}\delta_{ju}$$

Simplifying the deltas:

$$\mathrm{tr} = \sum_{i,j,k,l,m,n,s=1}^{T} D_{(ij),(kl)}q_{km}q_{ln}D_{(is),(mn)}q_{sj}$$

Now we can substitute the Jacobian components given by eq. (29).

Assume Einstein's notation.

$$q_{i,i}\Big[A_{ij}A_{in}q_{jn}^2 - A_{ij}A_{is}A_{in}q_{jn}q_{sj} - A_{ij}A_{il}A_{in}q_{ln}q_{nj} + A_{ij}A_{il}A_{is}A_{in}q_{ln}q_{sj}\Big]$$

To leading order in d we can substitute Q with its expectation value,

$$q_{ts} \simeq p + (q-p)\delta_{ts}.$$

Expanding each contribution and taking the limit:

$$\begin{aligned}
&(1)\ A_{ij}A_{in}q_{jn}^2 \ \longrightarrow\ p^2 + \big(2p(q-p) + (q-p)^2\big)Y^{(2)},\\
&(2)\ A_{ij}A_{is}A_{in}q_{jn}q_{sj} \ \longrightarrow\ p^2 + 2p(q-p)Y^{(2)} + (q-p)^2Y^{(3)},\\
&(3)\ A_{ij}A_{il}A_{in}q_{ln}q_{nj} \ \longrightarrow\ p^2 + 2p(q-p)Y^{(2)} + (q-p)^2Y^{(3)},\\
&(4)\ A_{ij}A_{il}A_{is}A_{in}q_{ln}q_{sj} \ \longrightarrow\ p^2 + 2p(q-p)Y^{(2)} + p(q-p)\,(Y^{(2)})^2.
\end{aligned}$$

putting all together:

$$tr = q(q-p)\Big[(q-p)\big(Y^{(2)} - 2Y^{(3)}\big) + p\big(Y^{(2)}\big)^2\Big].$$

Finally:

$$\mathbb{E}\left\|\frac{\partial \mathcal{S}(X)}{\partial W_Q}\right\|_F^2 = d^4\sigma_K^2\sigma_V^2\, q\sum_{ij} A_{ij}^2(q-p)^2 + q\sum_{ij} A_{ij}^3(q-p)^2 + q(q-p)^2\sum_{ijn} A_{ij}^2 A_{in}^2 \tag{30}$$

Recall that we took $\sigma_K^2 = \beta\sqrt{\log(T)}/d$ and $\sigma_V^2 = \sigma_v^2/d$, so:

$$\frac{1}{d^2}\mathbb{E}\left\|\frac{\partial \mathcal{S}(X)}{\partial W_Q}\right\|_F^2 = \beta\sqrt{\log(T)}\sigma_v^2\, q(q-p)\Big[(q-p)\big(Y^{(2)} - 2Y^{(3)}\big) + p\big(Y^{(2)}\big)^2\Big]. \tag{31}$$

Now if consider instead the gradient of the loss:

$$\frac{1}{d^2}\mathbb{E}\left\|\frac{\partial \mathcal{L}}{\partial W_Q}\right\|_F^2 \leq \mathcal{B}(X)\beta\sqrt{\log(T)}\sigma_v^2\, q(q-p)\Big[(q-p)\big(Y^{(2)} - 2Y^{(3)}\big) + p\big(Y^{(2)}\big)^2\Big]. \tag{32}$$

where $\mathcal{B}(X)$ is a bounded a quantity of X (see the proof of theorem 3.2 Noci et al. [2022]).

We assume that this quantity scales like $O_T(T^{-1})$ and check it numerically in fig. 4. So putting all together:

$$\frac{T}{d^2\sqrt{\log T}}\mathbb{E}\left\|\frac{\partial \mathcal{L}}{\partial W_Q}\right\|_F^2 \propto \beta\sigma_v^2\, q(q-p)\Big[(q-p)\big(Y^{(2)} - 2Y^{(3)}\big) + p\big(Y^{(2)}\big)^2\Big]. \tag{33}$$

Since the $Y^{(r)} \to 0$ in the small β regime [Mézard and Montanari, 2009], let's check our predictions for the case where $q = 1$, $p \simeq 0$. In this case $\beta_c = \sqrt{2}$, but in light of the discussion on finite size effects in section B.3, we actually expect our prediction to be sharp up to $\beta_c/2 \approx 0.7$ and then a crossover between $\beta_c/2$ and β_c to the other solution.

B.6 Full Transformer Block Analysis

We now briefly describe how to treat signal propagation through skip connections, Layer Norm, and the MLP, and then integrate our results into the simple iterative algorithm 1 that practitioners can use to predict the evolution of the average cosine similarity across Transformer layers and obtain trainability diagrams that identify viable hyper-parameter choices for a given Transformer variant. We then give three applications of the algorithm in section 3.

Putting it all together We derive the change to the average cosine similarity of tokens due to skip connections in section B.6.1, where we find an update equation that reads

$$\Phi_{\text{RES}_{\text{SA}}}(\rho) = \frac{p(\sigma_v^2 + \alpha_{\text{SA}}^2)}{\sigma_v^2\big(p + (q-p)Y^{(2)}(\beta)\big) + \alpha_{\text{SA}}^2 q} \tag{34}$$

where we recall that $\mathbb{E}[W_V^\top W_V] = \sigma_v^2 I_d$. The update equations to describe propagation through MLP layers follows Poole et al. [2016], Schoenholz et al. [2016]; for completeness, we restate these recursions in section B.6.2. The activation function only enters through a kernel $f(\rho)$ that updates the overlap; for ReLU for example, we have that $f(\rho) = \pi^{-1}\left[\sqrt{1-\rho^2} + \rho\,(\pi - \arccos(\rho))\right]$ Cho and Saul [2009]. Taking these results together, we arrive that the procedure described in algorithm 1 to track the evolution of the average pairwise cosine similarity between tokens in a Transformer at initialisation. We close the paper by discussing several applications of the algorithm.

B.6.1 Action of residual connections

Recall the action of the skip connection in self-attention:

$$\text{RES}_{\text{SA}}(X) = \mathcal{S}(X) + \alpha_{\text{SA}}X = AXW_V + \alpha_{\text{SA}}X.$$

Consider the quantity

$$\mathbb{E}_{Q,K,V}\left[\text{RES}_{\text{SA}}(X)_t \cdot \text{RES}_{\text{SA}}(X)_s\right].$$

The expectation over the value matrix vanishes in the mixed terms, leading to

$$\mathbb{E}_{Q,K,V}\left[\mathcal{S}(X)_t \cdot \mathcal{S}(X)_s\right] + \alpha_{\text{SA}}^2 X_t \cdot X_s.$$

Overall, considering

$$\rho = \frac{\mathbb{E}\langle q_{ts}\rangle}{\mathbb{E}\langle q_{tt}\rangle} = \frac{\sigma_v^2 p + \alpha_{\text{SA}}^2 p}{\sigma_v^2\big(p + (q-p)Y^{(2)}(\beta)\big) + \alpha_{\text{SA}}^2 q},$$

where we used $\mathbb{E}[W_V^\top W_V] = \sigma_v^2 I_d$ and the updates for p and q described in section B.2.3.

B.6.2 Action of MLPs

Finally, we can use the theory developed by Poole et al. [2016], Schoenholz et al. [2016] to include the effect of the ReLU MLP on signal propagation. For a two-layer ReLU MLP, the propagation of squared norm $q^{(l)}$ and pairwise inner product $p^{(l)}$ across layers is governed by:

$$q^{(l)} = \sigma_w^2 \int \mathcal{D}z\, \phi\left(\sqrt{q^{(l-1)}}z\right)^2 + \sigma_b^2, \quad l = 2, \ldots, L \tag{35}$$

$$p^{(l)} = \sigma_w^2 \int \mathcal{D}\mathbf{z}\, \phi\left(\sqrt{q^{(l-1)}}z_1\right) \phi\left(\sqrt{q^{(l-1)}}z_2\right) + \sigma_b^2 \tag{36}$$

where ϕ is the ReLU activation function, and $\mathbf{z} = (z_1, z_2)^\top$ is a pair of standard Gaussian variables with variance 1 and covariance $\rho^{(l-1)} = {}^{p^{(l-1)}}/_{q^{(l-1)}}$.

The initial conditions after the first linear layer are $q^{(1)} = \sigma_w^2 + \sigma_b^2$ and $p^{(1)} = \sigma_w^2 \rho^{(0)} + \sigma_b^2$. For a two-layer MLP with ReLU activations, the second layer outputs simplify to:

$$q^{(2)} = \frac{\sigma_w^2}{2} q^{(1)} + \sigma_b^2, \qquad p^{(2)} = \frac{\sigma_w^2}{2} q^{(1)} f(\rho^{(1)}) + \sigma_b^2 \tag{37}$$

where $f(\rho)$ captures the correlation structure after the ReLU nonlinearity Cho and Saul [2009],

$$f(\rho) = \frac{1}{\pi}\left(\sqrt{1-\rho^2} + \rho\,(\pi - \arccos(\rho))\right)$$

After adding the final residual connection after the MLP, we find the updated cosine similarity of tokens after a full transformer block:

$$\rho^{\text{block}} = \frac{p^{(2)} + \alpha_{\text{MLP}}^2 \rho^{(0)}}{q^{(2)} + \alpha_{\text{MLP}}^2} \tag{38}$$

B.6.3 Full Block cosine similarity update algorithms

Algorithm 2 Pre-norm Block Update

Inputs: β, q, p, α_{SA}, α_{MLP}, σ_w^2, σ_b^2, σ_v^2
▷ Pre-norm LN before attention
$p_{\text{LN}} \leftarrow p/q; \quad q_{\text{LN}} \leftarrow 1$
▷ Attention layer (normed input) + residual
$\beta_c \leftarrow \sqrt{\frac{2}{1(1-p_{\text{LN}})}}$
$Y^{(2)}(\beta) \leftarrow \max(0, 1 - \beta_c/\beta)$
$q \leftarrow \sigma_v^2 \left(p_{\text{LN}} + (q_{\text{LN}} - p_{\text{LN}}) \cdot Y^{(2)}(\beta)\right) + q \cdot \alpha_{\text{SA}}^2$
$p \leftarrow \sigma_v^2 p_{\text{LN}} + \alpha_{\text{SA}}^2 p$
▷ Pre-norm LN before MLP
$p_{\text{LN}} \leftarrow p/q; \quad q_{\text{LN}} \leftarrow 1$
▷ MLP layer (normed input) + residual
$q_1 \leftarrow \sigma_w^2 q_{\text{LN}} + \sigma_b^2; \quad p_1 \leftarrow \sigma_w^2 p_{\text{LN}} + \sigma_b^2$
$q_2 \leftarrow \frac{\sigma_w^2}{2} q_1 + \sigma_b^2; \quad p_2 \leftarrow \frac{\sigma_w^2}{2} f(p_1/q_1) q_1 + \sigma_b^2$
$q \leftarrow q_2 + \alpha_{\text{MLP}}^2 q; \quad p \leftarrow p_2 + \alpha_{\text{MLP}}^2 p$
return (q, p)

Algorithm 3 Gain-controlled Transformer Block Update with Post-norm

Inputs: β, q, p, α_{SA}, α_{MLP}, σ_w^2, σ_b^2, σ_v^2
▷ Attention layer + residual (centered-value update)
$\beta_c \leftarrow \sqrt{\frac{2}{q(q-p)}}$
$Y^{(2)}(\beta) \leftarrow \max(0, 1 - \beta_c/\beta)$
$q \leftarrow \sigma_v^2 \left((q-p) Y^{(2)}(\beta)\right) + \alpha_{\text{SA}}^2 q$
$p \leftarrow \alpha_{\text{SA}}^2 p$
▷ Post-norm LN
$p \leftarrow p/q; \quad q \leftarrow 1$
▷ MLP + residual
$q_1 \leftarrow \sigma_w^2 q + \sigma_b^2; \quad p_1 \leftarrow \sigma_w^2 p + \sigma_b^2$
$q_2 \leftarrow \frac{\sigma_w^2}{2} q_1 + \sigma_b^2; \quad p_2 \leftarrow \frac{\sigma_w^2}{2} f(p_1/q_1) q_1 + \sigma_b^2$
$q \leftarrow q_2 + \alpha_{\text{MLP}}^2 q; \quad p \leftarrow p_2 + \alpha_{\text{MLP}}^2 p$
▷ Post-norm LN
$p \leftarrow p/q; \quad q \leftarrow 1$
return (q, p)

C Figure details

All experiments have been conducted using a single NVIDIA A100-PCIE-40GB. Training times vary from approximately one hour (one layer models) to one day (60 layer models).

C.1 Figure 1

(a, b) Attention visualizations obtained using `BertViz` [Vig, 2019] on a single-layer, single-head Transformer at initialization. The attention maps illustrate the effect of varying β (directly related to variance of queries/keys): in yellow, a model initialized with $\beta = 0.1$ (low-variance regime, resulting in approximately uniform attention distributions), and in red, a model initialized with $\beta = 1.8$ (high-variance regime, leading to sharp attention).

(c) Trainability diagram for a 60-layer Transformer. The diagram is based on the critical value $\beta_c = \sqrt{2}$, which marks the threshold at which entropy collapse occurs in the first self-attention layer, assuming a sequence of orthogonal token embeddings. The residual strength threshold is defined as the smallest value of the residual scaling factor α_{SA} below which rank collapse (loss of representation diversity) occurs across 60 layers.

(d) Evolution of cosine similarity between token embeddings across layers in a 60-layer Transformer initialized in the low-variance regime (using standard HuggingFace initialization for queries and keys, which corresponds to small β. Lines denote theoretical predictions, while dots indicate empirical averages over 10 random initializations and 10 input sequences. (Error bars are the standard deviation.) High similarity indicates representational collapse.

(e) Same as (d), but for a 12-layer Transformer initialized in the high-variance regime ($\beta = 1.8$). Again, lines show theoretical predictions and dots indicate empirical means. *Remark:* we intentionally use a shallower model in this regime to ensure that any training failure observed in panel (f) is attributable to entropy collapse rather than rank collapse, as signal propagation across 12 layers is guaranteed.

(f) Pre-training results for BERT-style encoder models using a masked language modeling task (masking probability = 0.15) on the TinyStories dataset. We compare models with 60 layers (blue and yellow: small $\beta \simeq 0.02 < \beta_c$) and 12 layers (red: $\beta = 1.8 > \beta_c$). In particular learning curves in the three phases are: *trainable* ($\beta \simeq 0.02,\ \alpha_{\text{SA}} = 1.5,\ 2$), *rank collapse* ($\beta \simeq 0.02, \alpha_{\text{SA}} = 1.0$), *entropy collapse* ($\beta = 1.8, \alpha_{\text{SA}} = 1.0,\ 1.5,\ 2.0$). Remark: $\beta \simeq 0.02$ corresponds to the standard initialization from Hugginface given the set of hyper-paramters we are using. All models use ReLU activations, 6 attention heads, embedding dimension $d = 600$, and absolute positional embeddings.

Initialization: $\sigma_w^2 = 0.2$, $\sigma_b^2 = 0.0004$, $\sigma_V^2 = \sigma_w^2$; standard HuggingFace initialization for queries/keys; no biases or affine transformations in LayerNorm.

Residual scaling: α_{SA} as shown in the figure, $\alpha_{\text{MLP}} = 1.0$.

Optimizer:AdamW with learning_rate=1e-4, num_train_epochs=1, batch_size=64, max_grad_norm=1.0, lr_scheduler_type="linear", weight_decay=0.01, and warmup_ratio=0.05.

Note: $\beta_c = \beta_c(\rho = 0) = \sqrt{2}$ is the critical threshold for the first layer of self-attention, which at initialization takes inputs that are approximately orthogonal.

Pre-training 1-layer Transformer with 1 head using masked language modeling (masking probability = 0.15) on TinyStories. Embedding dim $d = 768$, standard residual strengths, ReLU activation, absolute position embedding. Custom init of queries/keys with βs as in figure; standard init for other weights, no biases for queries/keys, no affine in LayerNorm. Optimizer: AdamW, learning_rate=5e-4, num_train_epochs=1, batch_size=64, max_grad_norm=1.0, lr_scheduler_type="linear", weight_decay=0.01, warmup_ratio=0.05.

D Supplementary Figures

D.1 Training Gain-controlled Transformers on TinyStories.

We train gain-controlled transformers with post-LN with skip connections strength $\alpha_{\text{SA}} = \alpha_{\text{MLP}} = 1$ of one and twenty layers. The latter case would fail for a standard transformer in the same setting due to rank collapse. The training dynamics is reported in fig. 5.

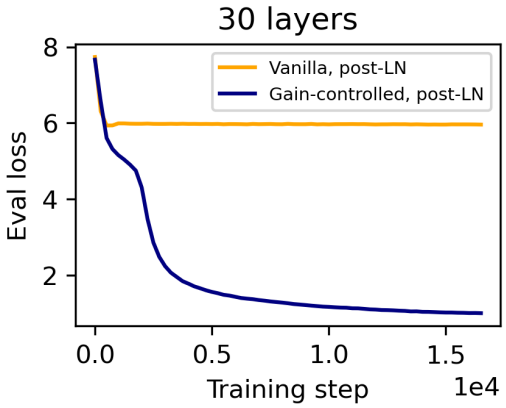

Figure 5: Training 30 layers of vanilla and Gain-controlled Transformer on TinyStories.

Details of training: 30-layer, single-head BERT-style model with embedding size 480 and ReLU activation, using masked language modeling with 15% masking probability, a learning rate of 5e-4, batch size 64, warmup ratio 0.05, weight decay 0.01, for 0.5 epochs.

D.2 Phase transition in infinitely deep transformers

A natural question is whether signal propagation can remain stable at infinite depth. While this is not possible with ReLU activations, it becomes a genuine phenomenon when using tanh. This behaviour was first observed by Poole et al. [2016] in MLPs, and more recently extended to Transformers by Cowsik et al. [2024]. In fig. 6, we confirm the phase transition in forward signal propagation, but emphasise that achieving this behaviour requires placing the MLP in the chaotic phase, which is associated with exploding gradients and unstable training.

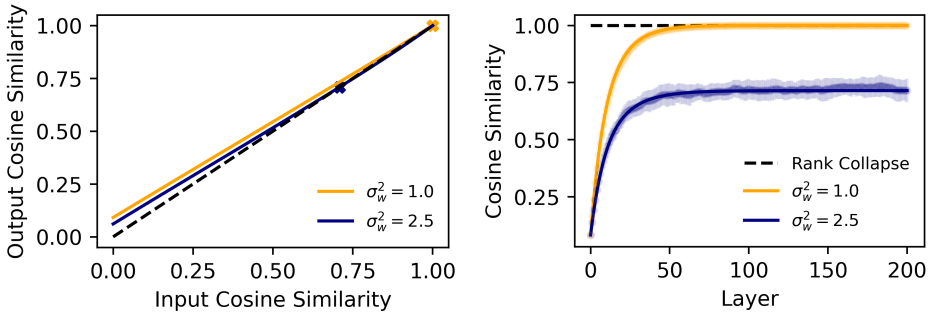

Figure 6: **Phase Transition to Infinitely Deep Signal Propagation.** *(Left)* Cosine-similarity update map of a full transformer block with `tanh` activations in the MLPs. By tuning the MLP variance to enter the chaotic regime, the collapsing effect of self-attention can be counterbalanced, resulting in a non-trivial fixed point in the similarity dynamics. *(Right)* Iterating the update map reveals the evolution of cosine similarity with depth—predictions align closely with experimental observations.

(Left) Theoretical predictions for a full Transformer block initialized in the low query/key variance regime, with MLP weights initialized as $\sigma_w = 1.0, 2.5$, $\sigma_b^2 = 0.1$. Residual connection scaling factors are set to $\alpha_{\text{SA}} = 6.0$ and $\alpha_{\text{MLP}} = 1.0$.

(Right) Evolution of the overlap by iteratively applying the map from the left panel over 200 layers, starting from an initial overlap of $\mathcal{O}(d^{-1/2})$. Solid lines denote theoretical predictions, while dots represent experimental results averaged over 25 random initializations, each evaluated on 10 input sequences (error bars are the standard deviation).