Token-token correlations predict the scaling of the test loss with the number of input tokens

Anonymous Author(s)
Affiliation
Address
email

Abstract

The success of Large Language Models (LLMs) establishes that machines trained for next-token prediction can acquire language proficiency. What are the mechanisms behind this acquisition and how much data do they require? We show that these questions can be partially answered by studying the correlations between the input tokens. Specifically, using scaling concepts of physics, we formulate a conjecture on the relationship between correlations, size of the training set and effective context window, i.e. the input tokens that are actually used by the model when predicting the next. Interestingly, when the correlations decay as a power of the distance between tokens, our conjecture connects to neural scaling laws and predicts how the scaling of test loss with dataset size should depend on the length of the context window. We confirm our conjecture and predictions on two datasets, consisting of Wikipedia articles and Shakespeare's lines.

1 Introduction

The question of how language is acquired is central for linguistics as well as machine learning. For instance, the success of LLMs trained for next-token prediction [1, 2] establishes that a language can be acquired from examples alone—albeit with a training set much larger than what humans are exposed to. Furthermore, empirical studies of LLMs' representations showed that they learn a hierarchy of contextual information, including notions of linguistics such as word classes and syntactic structure [3, 4, 5]. Recent studies have begun revealing the inner workings of LLMs by using synthetic data generated via context-free grammars [6, 7, 8], determining, in particular, the algorithm that these models follow when predicting the next token. However, there is no consensus on the mechanisms behind language acquisition [9, 10]. As a result, empirical phenomena such as the *scaling* of the test loss with dataset size and number of parameters [11] and the *emergence* of specific skills at certain scales [12, 13] remain unexplained.

In this work, we explore the idea that LLMs leverage data correlations to learn. This idea was introduced in the context of deep vision models trained for image classification [14]. In particular, assuming non-zero correlations between the class label and the input features leads to learnability with an iterative clustering algorithm that mimics the structure of deep convolutional networks [15, 16]. Further developments led to a demonstration of how deep networks leverage these correlations to efficiently learn hierarchical and compositional data, both in the classification [17] and the next-token prediction[18] settings. Here we test this idea empirically, by focusing on the pretraining phase of language models and consider two datasets, consisting of English Wikipedia articles and Shakespeare's lines, respectively. We find that:

- Token-token correlations C decay as a power of the distance between tokens t, $C(t) \sim t^{-\beta}$;

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

- A finite training set size P induces a sampling noise of order $P^{-1/2}$, thus limiting the resolution of correlations to an effective context window of size $t(P) \sim P^{1/(2\beta)}$;
- The relationship between t and P predicts the training set size at which the performance of language models trained on a finite context window saturates.

2 Notation and setup

Data and Correlations. We define a text datum, or sentence, as a sequence $\boldsymbol{x} = (x_1, \dots, x_d)$ of d tokens belonging to a finite vocabulary $\mathcal{V}$. Denoting with v the vocabulary size, each token x_i is represented as a v-dimensional one-hot vector $(x_{i,\mu})_{\mu=1,\dots,v}$ [1]:

$$x_{i,\mu} = \begin{cases} 1, \text{ if } x_i \equiv \mu\text{-th element of } \mathcal{V}, \\ 0, \text{ otherwise.} \end{cases} \tag{1}$$

A dataset, or *corpus*, consists of a probability distribution over sequences, which measures the frequency at which a given combination of tokens appears within the text. Assuming that all sequences have length d, the data distribution is a joint probability over d-dimensional sequences with elements in $\mathcal{V}$, $P_{\boldsymbol{X}}(\boldsymbol{x}) := \mathbb{P}\{X_1 = x_1, \dots, X_d = x_d\}$. We measure correlations between tokens via the token co-occurrences matrix, [2]

$$C_{i,j}(\mu, \nu) := \mathbb{P}\{X_i = \mu, X_j = \nu\} - \mathbb{P}\{X_i = \mu\}\mathbb{P}\{X_j = \nu\}, \tag{2}$$

where μ and ν are arbitrary elements of the vocabulary $\mathcal{V}$ and $\mathbb{P}$ refers to the data distribution $P_{\mathbf{X}}$.

Last-token prediction. We consider a simplified language modelling setup where the last token of the sequence is masked and a machine-learning model is trained to predict it. In other words, the model takes the *context window* $(x_1, \dots, x_{d-1})$ as input and outputs a parametric approximation p_θ of the conditional probability of the last token,

$$p_\theta(x_d | x_1, \dots, x_{d-1}) \approx \mathbb{P}\{X_d = x_d | X_1 = x_1, \dots, X_{d-1} = x_{d-1}\}. \tag{3}$$

p_θ is obtained by updating the parameters θ via gradient descent on the empirical cross-entropy loss, computed from a set of P training examples drawn from $P_{\boldsymbol{X}}$. The architectures we consider have the same structure as BERT [1]: They consists of multiple blocks, where each block includes a standard Multi-Head Attention layer [19], a token-wise two-layer perception (MLPs), layer normalization operations before the attention layer and the MLP and residual connections. Transformers are trained with the Adam optimizer, with a warmup scheduler bringing the learning rate to 10^{-2} within the first 10 training epochs. The batch size is set to the minimal size allowing convergence, where we define convergence as the training cross-entropy loss reaching a threshold value of 10^{-2}. We use a validation set of size 2^{15} to select the model with the best validation loss over the training trajectory.

3 Correlations, training set size and effective context window

Since the masked token is always the last in our setup, we define a *correlation function* as follows. Take the left-hand side of Eq. 2, set $j = d$ and define the distance $t = |i - d|$ between the i-th and the masked token, then compute the root mean square over the vocabulary:

$$\tilde{C}(t) := \left(v^{-2} \sum_{\mu,\nu \in \mathcal{V}} \left(C_{d-t,d}(\mu, \nu) \right)^2 \right)^{1/2}. \tag{4}$$

$\tilde{C}(t)$ measures the typical dependency between tokens as a function of their distance t. We denote the *empirical* correlation function, where correlations are measured from P samples of the data distribution, with $\tilde{C}_P(t)$. Examples of $\tilde{C}_P(t)$ are shown in the top-left panels of Fig. 1 (Wikipedia) and Fig. 2 (Shakespeare). The power-law decay is ubiquitous for text-like data [20], and observed empirically for different choices of tokenisation, including syllables [21], words [22] and part-of-speech tags [23]. This behaviour can be derived from the hierarchical and compositional structure of grammar [20], as show in [18] for a specific example of context-free grammar.

[1]throughout the paper, Latin indices indicate the token position and Greek indices the vocabulary entry.

[2]$C_{i,j}(\mu, \nu)$ is also equivalent to the covariance matrix of the one-hot representation, $\mathbb{E}\left[(X_{i,\mu} - \mathbb{E}[X_{i,\mu}])(X_{j,\nu} - \mathbb{E}[X_{j,\nu}])\right]$

Saturation due to finite sample size. Notice that the empirical correlation functions of Fig. 1 and Fig. 2 saturate for large t. This saturation is caused by the sampling error: For large P, $\tilde{C}_P(t)$ converges to a Gaussian random variable having mean equal to the infinite-P correlation function $\tilde{C}(t)$ and variance of order $1/(v^2 P)$. [3] This characteristic size is highlighted by horizontal, coloured dashed lines in the figures. As t increases, the mean $\tilde{C}(t)$ decreases and the sampling noise size emerges, resulting in an effective context window size $t^*(P)$, given by the value of t where $\tilde{C}(t) \sim t^{-\beta}$ intersects the sampling noise scale $\sim P^{-1/2}$,

$$(t^*)^{-\beta} \sim P^{-1/2} \Rightarrow t^*(P) \sim P^{1/z}, \quad \text{with } z = 2\beta. \tag{5}$$

As shown in the top right panels of Fig. 1 and Fig. 2, the relationship between t and P can also be represented by the following *scaling hypothesis* for the empirical correlations,

$$\tilde{C}_P(t) = P^{-1/2} c(t/P^{1/z}), \tag{6}$$

with $c(x) \sim x^{\beta}$ for $x \ll 1$ and $c(x) \sim$ cost. for $x \gg 1$.

Finite sample size equals effective context window. Eq. 5 suggests that a machine learning method that uses P examples can only extract information from the tokens within distance $t^*(P)$ from the last, leading to the following

Conjecture: *"If the token correlation function decays with the token distance, then a language model trained to predict the next token from a training set of P examples can only extract relevant information from an effective context window of P-dependent size $t^*(P)$."*

4 Test on real language data

In this section, we report the results of the test of our conjecture in two datasets: a selection of lines from Shakespeare's plays [25] and a collection of articles from English Wikipedia [24]. For both datasets we adopt a character-level tokenisation, resulting in over 10^6 tokens. We then extract sequences of t consecutive tokens and train BERT-like deep transformers in the setup of section 2. The results are reported in the bottom panels of Fig. 1 for Wikipedia and Fig. 2 of App. A for Shakespeare. First, as P increases, the test loss follows the empirical scaling law $\mathcal{L} \sim P^{-\alpha}$ (bottom left). However, the learning curve levels off at some characteristic scale P that grows with the size t of the context window. This phenomenon is qualitatively compatible with our conjecture, as it implies that the gains in performance observed when increasing P are entirely due to the ability of the model to leverage longer-range correlations.

Furthermore, by inverting the function $t^*(P)$ of Eq. 5 we get a characteristic training set size $P^*(t)$ where the training set allows for resolving correlations at all distances $t' < t$, $P^*(t) \sim t^z$. In other words, the relationship between t and P measured from the correlation functions predicts quantitatively the training set size where learning curves level off. Paired with the empirical power-law scaling with P, this result leads to the following context-dependent scaling hypothesis for the test loss:

$$\mathcal{L}(P,t) = t^{-\alpha z} f\left(\frac{P}{t^z}\right), \tag{7}$$

with $f(x) \sim x^{-\alpha}$ for $x \ll 1$ and constant for $x \gg 1$. This scaling hypothesis could also be formulated so as to highlight the dependence on the training set size,

$$\mathcal{L}(P,t) = P^{-\alpha} g\left(\frac{P}{t^z}\right), \tag{8}$$

with $g(x)$ constant for $x \ll 1$ and $g(x) \sim x^{\alpha}$ for $x \gg 1$. The collapse observed in the bottom right panels of Fig. 1 and Fig. 2 confirms Eq. 7 and our conjecture.

[3] While this is techinically true only if the token entries appear with the same frequency, it remains approximately true as long as the freuencies are not too dissimilar.

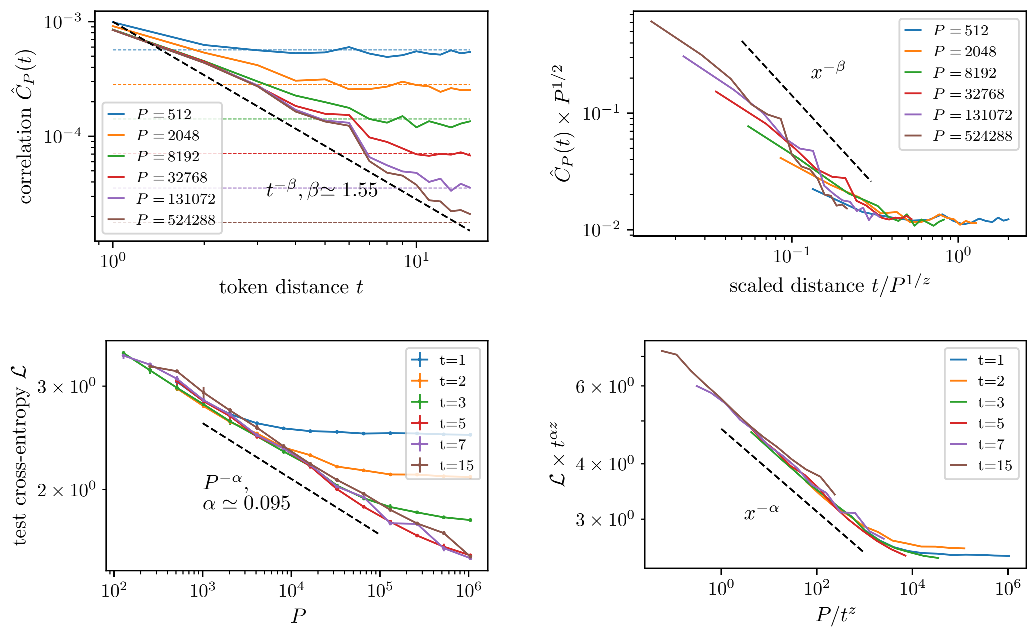

Figure 1: **Top, Left:** Empirical correlation functions $\hat{C}_P(t)$ of 16-character blocks from the WikiText-103 dataset [24], with P as in the key. All curves display an initial, approximately power-law decay, followed by saturation due to the sampling noise. The scales of the sampling noise are indicated by coloured, dashed lines. **Top, Right:** The empirical curves $\hat{C}_P(t)$ collapse when rescaling the correlations by the sampling noise size $P^{-1/2}$ and t by the characteristic distance $t^*(P) \sim P^{1/z}$, with $z = 2\beta \simeq 3.1$. **Bottom, Left:** Test losses of 6-layers transformers trained on $(t+1)$-characters blocks of the WikiText-103 [24] (t as in the key). The number of heads is set to $n_h = 8$, the embedding dimension to $d_e = 512$ and the size of the MLP hidden layer to $4d_e$. Increasing the number of layers or the number of heads does not affect the results presented in the figure. Notice the saturation of the loss to some t-dependent value after reaching a characteristic training set size. **Bottom, Right:** As predicted by our conjecture, the losses collapse when rescaled according to Eq. 7 with the same z as the correlation functions and $\alpha \simeq 0.095$.

5 Conclusions

We proposed a conceptual framework for understanding the performance-vs.-data scaling laws of language models trained for next-token prediction. In our picture, increasing the number of data allows for the resolution of a longer range of correlations. These correlations, in turn, can be exploited to improve the next-token prediction performance. This scenario is consistent with the empirical phenomenology of language models [11]. Furthermore, our analysis predicts a fundamental relationship between the effective context window captured by a language model trained with a finite training set and the decay of token-token correlations, which we confirmed empirically on two examples of text data. This finding suggests that the exponents entering scaling laws are influenced by intrinsic (and measurable) properties of the data. On the one hand, our predictions can be tested on state-of-the-art LLMs trained on larger datasets. On the other hand, our framework can be extended to shed light on other aspects of scaling laws of high practical relevance, such as the role of the number of parameters and the behaviour of performance when the model size and the number of data are optimised under a fixed compute budget.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.

[3] M. E. Peters, M. Neumann, L. Zettlemoyer, and W. Yih. Dissecting contextual word embeddings: Architecture and representation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, 2018. Association for Computational Linguistics.

[4] I. Tenney, D. Das, and E. Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, 2019. Association for Computational Linguistics.

[5] C. D Manning, K. Clark, J. Hewitt, U. Khandelwal, and O. Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

[6] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, context-free grammar. *arXiv preprint arXiv:2305.13673*, 2023.

[7] Haoyu Zhao, Abhishek Panigrahi, Rong Ge, and Sanjeev Arora. Do transformers parse while predicting the masked word? *arXiv preprint arXiv:2303.08117*, 2023.

[8] Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A Smith, Navin Goyal, and Yulia Tsvetkov. Learning syntax without planting trees: Understanding when and why transformers generalize hierarchically. *arXiv preprint arXiv:2404.16367*, 2024.

[9] Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

[10] Michael R Douglas. Large language models. *arXiv preprint arXiv:2307.05782*, 2023.

[11] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[12] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

[13] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

[14] E. Mossel. Deep learning and hierarchal generative models. *arXiv preprint arXiv:1612.09057*, 2016.

[15] Eran Malach and Shai Shalev-Shwartz. A provably correct algorithm for deep learning that actually works. *Preprint at http://arxiv.org/abs/1803.09522*, 2018.

[16] E. Malach and S. Shalev-Shwartz. The implications of local correlation on learning some deep functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 1322–1332, 2020.

[17] Francesco Cagnetta, Leonardo Petrini, Umberto M. Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Phys. Rev. X*, 14:031001, Jul 2024.

[18] Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. *arXiv preprint arXiv:2406.00048*, 2024.

[19] 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.

[20] Henry W. Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7), 2017.

[21] Tim Sainburg, Brad Theilman, Marvin Thielk, and Timothy Q Gentner. Parallels in the sequential organization of birdsong and human speech. *Nature communications*, 10(1):3636, 2019.

[22] Nikolay Mikhaylovskiy and Ilya Churilov. Autocorrelations decay in texts and applicability limits of language models. *arXiv preprint arXiv:2305.06615*, 2023.

[23] Kai Nakaishi, Yoshihiko Nishikawa, and Koji Hukushima. Critical phase transition in a large language model. *arXiv preprint arXiv:2406.05335*, 2024.

[24] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

[25] The unreasonable effectiveness of recurrent neural networks, 2015.

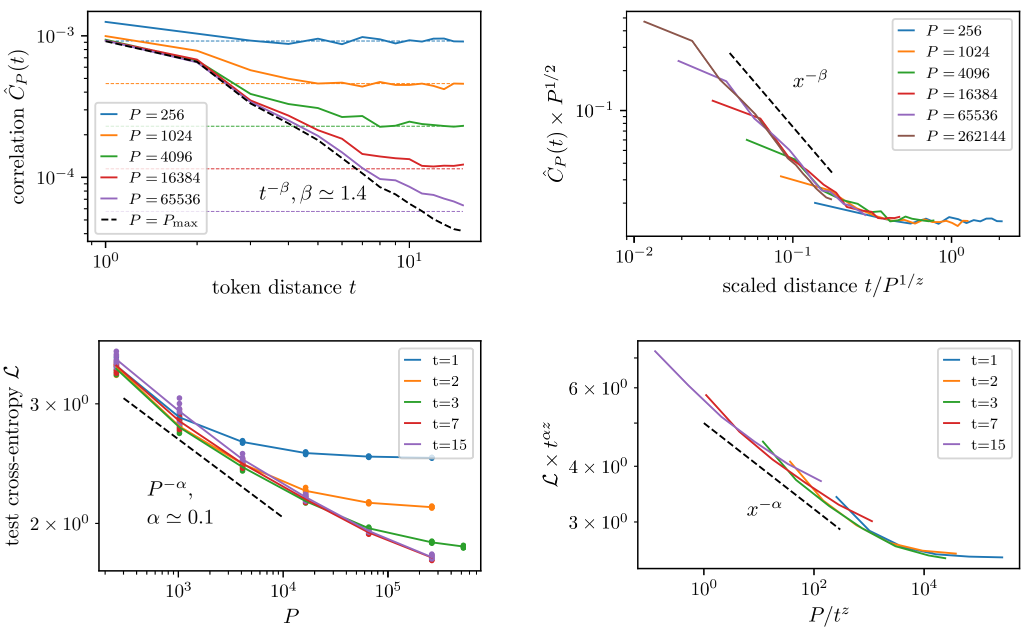

Figure 2: **Top, Left:** Empirical estimates $\hat{C}_P(t)$ for different training set sizes P as in the key. The curves initially follow the true correlation $\tilde{C}(t)$ (black dashed), but then saturate due to the sampling noise (coloured dashed). **Top, Right:** The empirical curves $\hat{C}_P(t)$ collapse when rescaling correlations by the sampling noise size $P^{-1/2}$ and t by the characteristic distance $t^*(P) \sim P^{1/z}$, with $z \simeq 2.8$. **Bottom, Left:** Test losses of 3-layers transformers trained on $(t+1)$-characters blocks of the tiny-Shakespeare dataset [25] (t as in the key). The number of heads is set to $n_h = 8$, the embedding dimension to $d_e = 256$, the size of the MLP hidden layer to $4d_e$. The saturation of the loss to some t-dependent value indicates that performance improves with P because the model can use information from a larger context window. **Bottom, Right:** As predicted by our conjecture, the losses collapse when rescaled according to Eq. 7 with the same z as the correlation functions.

A Loss saturation and correlations for tiny Shakespeare

In this section, we report the results of the test of our conjecture for the tiny Shakespeare dataset [25]. The results are summarised in Fig. 2, which displays the same measures as Fig. 1 and, as Fig. 1, confirms our conjecture.