LLMs Information Flow Diagnostic: Memory-based Evidence from Random Matrix Theory

Sami Diaf
Department of Socioeconomics
Universität Hamburg
sami.diaf@uni-hamburg.de

Abstract

The diagnostic of neural networks, particularly Large Language Models (LLMs), remains a critical aspect of today's AI-powered solutions, whose training data are not available to users for testing purposes. Practitioners usually aim to fine-tune their models to maximize the accuracy, by leveraging the traditional test metrics, whose application on large models remains expensive. Recent advances considered layer-based norms and power-law metrics for a robust meta-analysis, without the need to access training and test data. Inherently, elements from Random Matrix Theory were used to reveal inner correlation patterns and size scales within each layer, so to detect bottlenecks in pre-trained models. This article extends the use of such schemes by analyzing memory dynamics and the probabilistic properties of power-law metrics to study the information flow within specific LLMs. Taken on a pre-tained German LLM (*LLaMmlein*) and its original English model (*TinyLlama*), this approach confirmed embedded self-similar, fractal properties of power-law metrics, hinting heavy tails and long-range correlations in the training process with a substantial amount of undertrained layers. This variability was found to be slightly persistent in the original English TinyLlama model and its German version, however the latter's chat version exhibits a pure randomness in its metrics. Findings stress out the role of attention mechanism as the main driver of LLMs training issues, while language-specific structures may cause metrics' distortions, hence altering the inter-layer information transmission as a component of the training process.

1 Introduction

The advent of neural networks, coupled with intensive computational innovations, popularized the use of deep learning as a modeling standard, outperforming other existing machine learning algorithms. Although the widespread use of such capabilities opened new research areas, deep neural networks (DNNs) remain black box models, whose effectiveness depends on complex hyperparameter optimization (Wu et al., 2019) to achieve a robust training. This forced practitioners to adopt expensive feature engineering schemes, without clearly setting up a strong theoretical background for users (Martin et al., 2021).

Large Language Models (LLMs) have been extensively designed, as large scale models, to accomplish several complex tasks in Natural Language Processing (NLP). Tuning and testing such models require extensive learning time (Burns et al., 2025), while training and test data are not always publicly available. Moreover, such DNNs are based on transformers (Vaswani et al., 2017) and require a special attention because they feature memory mechanisms, as for multihead attention and BiLSTM (Graves and Schmidhuber, 2005). Although these memory-based architectures are complex to handle, they became the default choice for many NLP architectures, as for the popular BERT model (Devlin et al., 2019).

The term *memory* refers, for the particular case of DNNs, to any mechanism by which a model or agent stores, retrieves and uses historical information (Zhang et al., 2024b), whether internally or externally. This paper considers the memory stemming from the information exchanged between layers, that is the output flow of each layer in the architecture, given by its weight matrix.

Random Matrix Theory (RMT) (Tulino and Verdú, 2004) is considered as the central limit theorem for matrix analysis and was used to study the overall performance of DNNs (Martin and Mahoney, 2021), on the basis of extracted eigenvalues of each weight matrix in the architecture. While earlier approaches considered mapping neural networks to a Gaussian process (Jacot et al., 2018), Martin et al. (2021) set up a practical background to identify similarities in the learning process of multiple DNNs, particularly fitting issues and the

bona fide of different regularization schemes to reduce correlations inside each layer. This extended the concept of Self-Regularization theory (Malevergne and Sornette, 2004), which assumes the generic existence of a self-organized macroscopic state in any large multivariate system. Martin et al. (2021) came to the conclusion that an implicit self-regularization at DNNs was prevailing, at the contrast of explicit regularization (L1 and L2) constraining the norm of weight matrices.

This new field of research set up effective generalization metrics detailing the inner functioning of DNNs, especially the learning process, the inter-layer information flow and the intra-layer asymptotic convergence (Martin et al., 2021). It borrows elements from statistical mechanics and was used for many applications as for cyber threat detection (Ferrag et al., 2024) and the description of feature learning applications (Seroussi et al., 2023).

In parallel to the use of power laws (PL) in various scientific fields, pattern similarities were studied under the name of *fractal analysis*, defining the behavior of self-similar patterns whose occurrence is not purely random, but follows a power-law behavior (Mandelbrot, 1982). The *fractality* is an essential feature in language theory, denoting the complexity stemming from word usage (Hiver et al., 2022), and was recently used in information processing (Wang et al., 2024). It fits the study of the information correlation proposed by Martin et al. (2021) which relies on a power-law fit over heavy-tailed distributions.

While the training quality of popular NLP and Computer Vision models came to scrutiny via norms and PL-based metrics (Yang et al., 2023), it ignored their inter-layer information exchange as a component of the training process. This concern is particularly determinant for LLMs, whose complex architecture features two distinct types of attention mechanisms (Vaswani et al., 2017; Martin et al., 2021), as a key component a transformer.

Thus, this paper enriches the existing DNNs empirical methodology by investigating the existence of pattern similarity in the information transmission on selected LLMs trained over English and German corpora. It extends the layer-based meta-analysis on such big architectures and details inter-layer persistence behavior. The latter reveals short/long term variations in the training process, whose non-linearity is linked to underfitted layers.

For this aim, two German LLMs, namely *LLaMmlein_1B* model[1] and a lightweight, small-scale version *LLaMmlein_120M* model[2], were used in this paper to conduct a transfer learning experiment, along the English *TinyLlama*, who served in training the *LLaMmlein*.

Aside from a meta-analysis on each selected LLM following Martin et al. (2021), an additional memory check was conducted to dissect hidden trends in the PL-based metrics. It revealed mild persistency and underfitting of metrics featuring information correlation and the size scale. Metrics based solely on information correlation were found to indicate heavy-tailed distribution of the eigenvalues and a high persistence, denoting the importance of the size scale in the information flow analysis.

Findings indicate layers exhibit substantial underfitting properties in both languages, mainly due to attention mechanisms. Original TinyLlama (Zhang et al., 2024a), both the full and the chat versions, have a mild persistent flow of information, compared to the German LLaMmlein whose lightweight version is though slightly anti-persistent. The size scale, measured by the maximum eigenvalue, proved to be important in harmonizing the per-layer metrics. Differences in results obtained from English and German LLMs could be explained by the morphologically-rich characteristic of the German language, known to be a SOV (Subject-Object-Verb), while English language exhibits a less complex SVO structure (Vikner, 2019).

The paper outlines the use of Random Matrix Theory in DNNs analysis (Section 2), then details the Rescaled Range Analysis (Hurst, 1951), as a method to study fractal properties and persistency measurement (Section 3). Section 4 features two language-based applications on English and German LLMs and compares their metrics and persistency measurements.

2 Random Matrix Theory

Train and test data have been the de facto tools to assess machine learning models in general, and neural networks in particular. In the absence of such data, elements from Random Matrix Theory were applied on final weight matrices of neural networks (Martin and Mahoney, 2021) to check their asymptotic convergence. It resulted several norms and metrics, whose statistical properties were found to

[1] https://huggingface.co/LSX-UniWue/LLaMmlein_1B

[2] https://huggingface.co/LSX-UniWue/LLaMmlein_120M

match DNNs accuracy, without accessing data used to train the models (Martin et al., 2021). In other terms, this strategy permits to discover whether a layer learned too much from the noise (overfitting) or alternatively has not learned enough from the signal (underfitting), assuming data stem from two components: signal and noise.

The *WeightWatcher* open source tool (Martin et al., 2021) investigates the weight matrix W of a given DNN layer, by analyzing its spectral properties. While every element of the weight matrix[3] W_{ij} is assumed to follow a normal distribution $\mathcal{N}(0, \sigma^2)$, the empirical correlation (Wishart) matrix $\boldsymbol{X} = \frac{1}{N} W^\intercal W$ is taken as the basis for quality assessment, by extracting its eigenvalues spectrum.

The Marchenko-Pastur (MP) distribution (Marchenko and Pastur, 1967) considers the spectrum of eigenvalues bounded between λ_- and λ_+ as relevant to the noise randomness. Its probability density $f(\lambda)$ is given for a $T \times N$ matrix and a noise level σ^2 as:

$$f(\lambda) = \begin{cases} \frac{N}{T} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\sigma^2} & \text{if } \lambda \in [\lambda_-, \lambda_+] \,, \\ 0 & \text{if } \lambda \notin [\lambda_-, \lambda_+] \,. \end{cases}$$

where $\lambda_- = \sigma^2(1 - \sqrt{\frac{T}{N}})^2$ and $\lambda_+ = \sigma^2(1 + \sqrt{\frac{T}{N}})^2$

The eigenvalues distribution, plotted as a histogram using the Empirical Spectral Density (ESD), is an informative feature of the randomness prevailing in every layer constituting the DNN, in addition to reveal inter-layer differences.

Because many matrices hold strongly correlated elements, the MP distribution is used to empirically evaluate a noisy spectrum of eigenvalues, that could be separated from other eigenvalues representing the signal.

Martin and Mahoney (2021) found most weight matrices in DNNs exhibit heavy-tailed distributions of eigenvalues as they become increasingly correlated, suggesting rather drawing elements from power-law generated data, as for Pareto distribution. This concept, known as Heavy-Tailed Self-Regularization (HT-SR) theory, is linked to situations where separating the noise from the signal becomes difficult to achieve, as eigenvalues are in this case better modeled via heavy-tailed distributions (Malevergne and Sornette, 2004), rather than a simple MP distribution.

For this aim, Martin and Mahoney (2021) estimated a truncated power-law fit (Clauset et al., 2009) over the MP curve, yielding the exponent α from the equation ESD–eigenvalues: $\rho(\lambda) \sim \lambda^{-\alpha}$ for $\lambda \in [\lambda_-, \lambda_+]$. The amplitude of the PL-exponent α is considered as the *information correlation* index within each weight matrix, denoting the strength of the existing element-wise correlations. Moreover, the α exponent is indeed a power-law fit that can be considered as a complexity index or a *fractal dimension* (Mandelbrot, 1982).

Based on the eigenvalues spectrum λ_i of each correlation matrix $\boldsymbol{X}$, several metrics were used as for:

- Frobenius norm : $\|W\|_F^2 = \|X\|_F = \sum_{i=1}^{M} \lambda_i^2$
- Spectral norm : $\|W\|_\infty = \|X\|_\infty = \lambda_{max}$
- Weighted α : $\hat{\alpha} = \alpha Log \lambda_{max}$
- α norm (Shatten-norm) : $\|W\|_{2\alpha}^{2\alpha} = \|X\|_\alpha^\alpha = \sum_{i=1}^{M} \lambda_i^\alpha$

where λ_i is the i^{th} eigenvalue of $\boldsymbol{X}$, λ_{max} is the maximum eigenvalue and α is the fitted power-law exponent, usually truncated because it needs defining specific lower and upper bounds, respectively λ_- and λ_+. For instance, Figure 1 reports simulations yielding random-like eigenvalues fitted with a scale-invariant Marchenko-Pastur curve between λ_- $\simeq$0.31 and λ_+ $\simeq$1.17 and spikes (signal) associated with $\lambda_i > \lambda_+$. The PL-fit yields a value of 0.571 for α.

The plain α metric is a scale-invariant, weak estimation of the information correlation, as it ignores the size scale (λ_{max}) within each layer. The latter remains an important determinant of HT-SR because DNNs are known to be non-linear, while LLMs particularly feature attention layers with large matrices. For small values of α, the size scale λ_{max} was found to be a good proxy for estimating the difference between the noise and the signal, however, for higher values of α (HT-SR), the signal gets mixed with the noise and λ_{max} is non-informative.

A clear distinction between norm-metrics and PL-based metric was given when studying the performance of several DNNs models (Martin et al., 2021; Yang et al., 2023). They concluded that

[3] A layer with multiple weight matrices will have a single concatenated weight matrix (Martin et al., 2021).

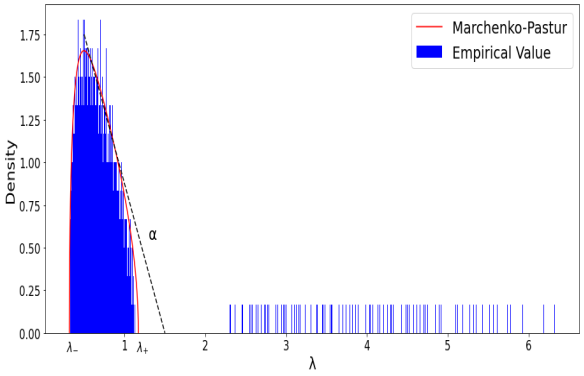

Figure 1: Marchenko-Pastur distribution simulated 1,000 times on the correlation matrix of an initial random matrix with $\frac{T}{N}$=10 and $\sigma^2 = \frac{2}{3}$. α is the PL-exponent of the Marchenko-Pastur fit over the interval [λ_-,λ_+].

PL-based metrics, aside from being good proxies for overall accuracy measurements, remain robust in detecting potential bottlenecks and training issues than norm-based metrics. Hence, PL exponent remains a robust empirical metric to asses well-trained DNNs and quantify the layer-wise correlation flow (Martin et al., 2021).

In practice, α was found to match an ideal DNN fit when approaching 2. This means the DNN model performs well as it facilitates the propagation of information/features across layers, because it learns from both data signal and noise. Values in the interval [4,6] are proxies of underfitting situations (not learning enough from the signal), while lower values equaling 1.5 are synonyms of overfitting (learning too much from the noise) (Martin et al., 2021). Large values of $\alpha > 6$ are associated with a pure randomness, which requires the aspect ratio $\frac{T}{N}$ to differentiate layers.

Because the size of DNNs layers changes according to adopted architectures, Martin et al. (2021) proposed to weight the α with the size scale to produce the weighted α metric. It was found that for small values, the weighted α approximates well the α Shatten-norm; the latter weighs the α exponent for all eigenvalues within the layer.

Martin et al. (2021) reported that weighted α and log α norm correlate at a higher level for well trained models. The size scale, given by λ_{max}, could be informally linked to situations where input clusters are at a greater distance. This means the size scale is related, in the case of LLMs, to the language morphologic aspects (sentence structures).

Particularly in LLMs, distortions in the series of PL exponents is called *scale collapse*, mostly linked to transformers (Vaswani et al., 2017; Lefaudeux et al., 2022). As memory-based blocks of layers, transformers feature a complex inner structure usually yielding larger weight matrices.

The study of such variations and the training process requires detailing the information flow throughout the whole network. The adoption of advanced tool for self-similar patterns, known as fractals (Mandelbrot, 1982) is clearly indicated to test the persistency hypothesis on trained DNNs. Persistent behavior of the aforementioned metrics reinforces the hypothesis of a strong, correlated inter-layer linkage propping up the information flow. One can assert that anti-persistency of PL-metrics may indicate colliding trends that alter the training process and the inter-layer dynamics, while persistency may reinforce the hypothesis of a harmonized network design that better captures long-range dependencies via attention layers.

3 Fractal Analysis

Mandelbrot tried first to uncover repeated patterns able to explain the randomness of irregular shapes (Mandelbrot, 1982), as exemplified by Koch's snowflake. This led to the concept of self-similar patterns, which stands for scale-dependent shapes with a known geometry. Hence, the *fractal* analysis was first established as a research field in geometry having a wide range of applications, from physics to hydrology. The fractal theory relies on the definition of a fractal dimension, a hidden variable that quantifies the irregularity of shapes found in many objects.

In time series analysis, the fractal approach was first featured when studying the Nile river flooding history. Hurst (1951) designed the *Rescaled Range (R/S) Analysis* and reckoned the Hurst exponent as a measure of a time series memory, later corrected by Mandelbrot and extended to the fractional Brownian motion (Mandelbrot and van Ness, 1968) when studying cotton prices in the United States.

The R/S algorithm takes the variations of a given time series of length T and divides them into N adjacent intervals of length τ, where $T = N\tau$. For each interval, the average value is computed and a new time series is created as accumulated deviations from the arithmetic mean values (hereafter named profile). The difference (range) between the

maximum and the minimum value of the profile, and the standard deviation of the original time series for each interval, are calculated. Each range is standardized by the corresponding standard deviation and forms a rescaled range so that the average *rescaled range* for a given interval of length $(R/S)_\tau$ is calculated.

The rescaled range scales are given by $(R/S)_\tau \approx c\tau^H$, where c is a finite constant independent of τ (Taqqu et al., 1995). To estimate the power law relationship, a simple log-log ordinary least squares regression is used for: $\log(R/S)_\tau \approx \log c + H \times \log \tau$, where H is the estimated Hurst exponent (Barunik and Kristoufek, 2010). R/S analysis was shown to be biased for small τ (Couillard and Davison, 2005), and empirical application considered rather the expected Hurst exponent (Weron, 2011). Values of H exceeding 0.5 are proxies of a persistent behavior resulting from long-range correlations, while values less than 0.5 are anti-persistent. A Hurst exponent not significantly different from 0.5 is associated to the standard Brownian motion. The Hurst exponent H is also a proxy of the fractal dimension D in time series, linked by the relationship: $D = 2 - H$.

Given the relatively reduced number of layers in most DNNs, this article considers the existence of a single fractal dimension, approached by the Hurst exponent. For each layer in an LLM, PL-metrics are computed on the related weight matrix, yielding three different series across the whole LLM to run the R/S Analysis on each one of them.

4 Application

The study of memory properties of specific LLMs is conducted on the weight matrices, stored after achieving the LLMs training. PL-based metrics adopted by Martin et al. (2021) were previously found to be robust when assessing hundreds of LLMs, outperforming simple algebraic norms (Frobenius and spectral norms).

The weighted α and log α norm are compound metrics computed from a truncated PL-fit of the eigenvalues and the size scale. These two metrics will have a particular attention in this section, as they go in-line with the PL-exponent yielded by the R/S Analysis, known as the Hurst exponent. The purpose lies on investigating the inter-layer dynamic flow using above two metrics and uncover potential variability known as *scale collapse* (Martin et al., 2021), which is assumed to reveal dysfunctions in the learning process. The α series will not be considered for the R/S analysis, as it ignores the size scale.

The selected LLMs are publicly available and their PyTorch versions (Paszke et al., 2019) were used to run the *WeightWatcher* diagnostic tool. The R/S analysis was performed on the basis of estimated PL-metrics, whose relatively reduced size requires a corrected version of the Hurst exponent (Weron, 2011) reported in Table 2.

4.1 English *TinyLlama*

TinyLlama model (Zhang et al., 2024a) was trained on a complex architecture featuring flash attention 2 and various fused schemes, comprising xFormers (Lefaudeux et al., 2022) as a research tool for accelerated transformers.

Figure 2 displays the per-layer metrics for the TinyLlama 1.1B model trained over 155 layers. The weighted α and the log α norm are highly correlated and clearly separable from the simple α metric, which exhibits a pronounced variability. This denotes the importance of the size scale, absent from the α metric, but present in the two others. Similar patterns were found in the TinyLlama 1.1B chat model (Figure 3), although its first layers are less pronounced then the original model.

The variability of the above metrics is a result of heavy-tailed eigenvalues distributions associated to a *scale collapse*. This denotes implicit changes or perturbations that occurred when training the model, likely due to distillation, data augmentation or fine-tuning.

Both LLMs feature a relatively high number of layers found to be under-trained, as reported in Table 1. These demonstrate high α values and are linked to value-type (V) self attention layers (having a rank of 256). They are particularly aggregated representations of the words in context (Vaswani et al., 2017), compared to query (Q) and key (K) matrices. The relative low number of over-trained layers confirms difficulties of fine tuning LLMs who are over-trained (Springer et al., 2025).

First layers, usually associated with higher metrics due to their effective normalization (Martin et al., 2021), do not exhibit here higher values of weighted α and the log α norm, compared to what was reported in Martin et al. (2021).

Table 2 reveals a slight persistency of the weighted α and log α norm metrics for the LLM chat version (Hurst exponent respectively 0.60

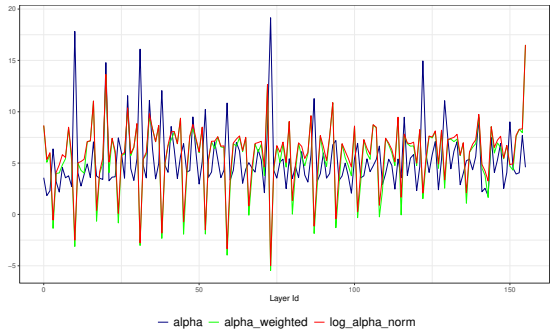

Figure 2: PL metrics estimated from TinyLlama 1.1B model

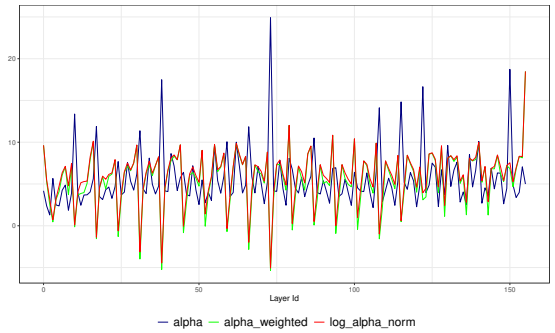

Figure 3: PL metrics estimated from TinyLlama 1.1B Chat

and 0.61), while the full model exhibits a non-persistent, Brownian-like behavior (Hurst exponent 0.51 each). The buildup of the chat version proved to have more inter-layer information than the original model, as a result of intensive fine-tuning on synthetic dialogues provided by Zephyr (Tunstall et al., 2023).

Both LLMs show similar PL-metric patterns and persistence, reinforcing the hypothesis of a strong transfer learning between the original model TinyLlama 1.1B and its chat version. The metric correlations of weighted α and log α norm are almost identical, respectively 0.879 and 0.887.

4.2 German *LLaMmlein*

The layer-to-layer information flow, as given by three metrics in Figure 4 and Figure 5, demonstrates key differences between the German LLaMmlein and its lightweight version (LLaMmlein 120M chat). The latter features 85 layers, compared to the 155 comprised in the former. Weighted α and log α norm are highly correlated in both models, however, the lightweight version displays a relatively stable α metric, not as variable as in the LLaMmlein 1B model, whose metrics have long-range correlations (Hurst exponent 0.61 in Table 2.

Higher values of α for LLaMmlein 1B are associated with V self attention layers of rank 256 (Figure 4), that carry context-based information of each sentence/word fed to the LLM. The lightweight version (LLaMmlein 120M) presents the lowest rate of under-trained layers, despite its reduced depth. This means this abridged version does not suffer from over-parametrization, relative to the amount of data. However, slight differences in the Hurst exponent values indicate a weak anti-persistency of the weighted α (Hurst exponent 0.46) compared to Brownian-like log α norm (Hurst exponent 0.52).

The impact of the size scale (λ_{max}) seems to be mild in the lightweight version, in comparison with the full model. This explains why the information correlation series α does not feature very high values in the lightweight model and exhibit a relative stability compared to the full model. The size scale has, particularly for the lightweight version, a linguistic feature embedded in the dataset[4].

The German language features a SOV structure (Vikner, 2019), at the contrary of the common SVO structures found in English and French. This considers German as a morphologically-rich language (Günther et al., 2019) whose structure is complex but rich, compared to English. Moreover, German LLMs are mostly trained on the basis of existing English and/or Multilingual LLMs, while recent attempts proposed a data curation methodology to improve LLMs training (Burns et al., 2025).

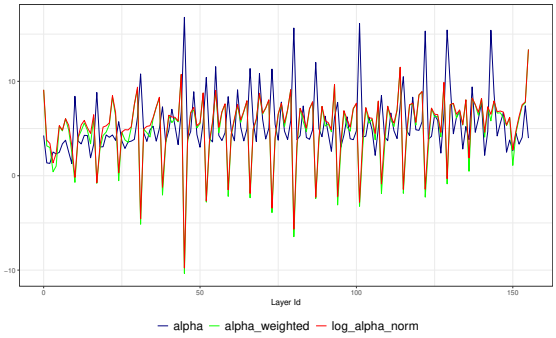

Figure 4: PL metrics estimated from LLaMmlein 1B model.

5 Conclusion

Machine learning models have long been associated with the train/test paradigm and the related metrics to perform quality control checks. For DNNs, practitioners use models without access

[4]Training data were de-duplicated on the paragraph level and filtered using a token-to-word ratio.

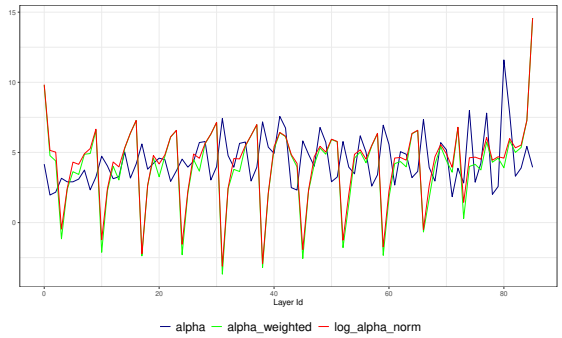

Figure 5: PL metrics estimated from LLaMmlein 120M model.

Model	Overtrained	Undertrained
TinyLlama 1.1 B	1.3%	26.3%
TinyLlama 1.1 B Chat	1.3%	29.5%
LLaMmlein 1B	2.9%	28.8%
LLaMmlein 120M	2.3%	13.9%

Table 1: Percentages of over-/under-trained layers, based on estimated α values, obtained from *WeightWatcher* tool (Martin and Mahoney, 2021)

to training data and are not able to perform independent accuracy tests. Elements from statistical mechanics were used to check the robustness of DNNs on the basis of their weight matrices, as information-carriers of the learning process. The use of Random Matrix Theory helped revealing embedded, heavy-tailed properties of eigenvalues via a truncated power-law fit, whose exponent is taken as a proxy of underfitting or overfitting presence in the related layer. Hybrid metrics combining power-law exponents and size scale proved to be accurate in estimating the between/within layer information flow, particularly in the case of LLMs who feature attention layers as memory-driver mechanisms. The inter-layer information flow, as an element of the training process, was found to exhibit a noticeable persistence in terms of long-range correlations. Such findings confirm the fractality of LLMs learning process and the importance of language-properties carried by data, whose complexity flags substantial underfitting issues affecting attention layers. The self-similarity analysis provides tools to detect potential training bottlenecks, but also a powerful way to assess transfer learning strategies when designing lightweight and task- and language-specific models. This proved particularly effective for the German language, whose morphologically-rich properties make the training difficult and require a special hyperparameter tuning and data processing.

Model	α	Weighted α	Log α norm
TinyLlama 1.1 B	0.63	0.51	0.51
TinyLlama 1.1 B Chat	0.49	0.60	0.61
LLaMmlein 1B	0.79	0.61	0.61
LLaMmlein 120M	0.74	0.46	0.52

Table 2: Estimates of Hurst exponents for each model, based on estimated α, weighted α and log α norm, obtained from *WeightWatcher* tool (Martin and Mahoney, 2021)

References

Jozef Barunik and Ladislav Kristoufek. 2010. On hurst exponent estimation under heavy-tailed distributions. *Physica A: Statistical Mechanics and its Applications*, 389(18):3844–3855.

Thomas F Burns, Letitia Parcalabescu, Stephan Wäldchen, Michael Barlow, Gregor Ziegltrum, Volker Stampa, Bastian Harren, and Björn Deiseroth. 2025. Aleph-alpha-germanweb: Improving german-language llm pre-training with model-based data curation and synthetic data generation.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Michel Couillard and Matt Davison. 2005. A comment on measuring the Hurst exponent of financial time series. *Physica A: Statistical Mechanics and its Applications*, 348(C):404–418.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C. Cordeiro, Merouane Debbah, Thierry Lestable, and Narinderjit Singh Thandi. 2024. Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices. *IEEE Access*, 12:23733–23750.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Fritz Günther, Eva Smolka, and Marco Marelli. 2019. 'understanding' differs between english and german: Capturing systematic language differences of complex words. *Cortex*, 116:168–175. Structure in words: the present and future of morphological processing in a multidisciplinary perspective.

Phil Hiver, Ali H. Al-Hoorie, and Reid Evans. 2022. Complex dynamic systems theory in language learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition*, 44(4):913–941.

Harold Edwin Hurst. 1951. Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers*, 116(1):770–799.

Arthur Jacot, Franck Gabriel, and Clément Hongler. 2018. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA. Curran Associates Inc.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. 2022. xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers.

Y. Malevergne and D. Sornette. 2004. Collective origin of the coexistence of apparent random matrix theory noise and of factors in large sample correlation matrices. *Physica A: Statistical Mechanics and its Applications*, 331(3):660–668.

Benoit B. Mandelbrot and John W. van Ness. 1968. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437.

Benoît B. Mandelbrot. 1982. *The fractal geometry of nature*. W. H. Freeman and Comp., New York.

V. A. Marchenko and L. A. Pastur. 1967. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, 1:422–437.

Charles H. Martin and Michael W. Mahoney. 2021. Implicit self-regularization in deep neural networks: evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(1).

Charles H. Martin, Tongsu Peng, and Michael W. Mahoney. 2021. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. 2023. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. 2025. Overtrained language models are harder to fine-tune.

Murad S. Taqqu, Vadim Teverovsky, and Walter Willinger. 1995. Estimators for long-range dependence: An empirical study. *Fractals*, 03(04):785–798.

Antonia Tulino and Sergio Verdú. 2004. *Random Matrix Theory and Wireless Communications*. Now Foundations and Trends.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Sten Vikner. 2019. *Why German is not an SVO-language but an SOV-language with V2*. AU Library Scholarly Publishing Services.

Zhenhua Wang, Fuqian Zhang, Ming Ren, and Dong Gao. 2024. A new multifractal-based deep learning model for text mining. *Information Processing and Management*, 61(1):103561.

Rafal Weron. 2011. HURST: MATLAB function to compute the Hurst exponent using R/S Analysis. HSC Software, Hugo Steinhaus Center, Wroclaw University of Science and Technology.

Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng. 2019. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, 17(1):26–40.

Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E. Gonzalez, Kannan Ramchandran, Charles H. Martin, and Michael W. Mahoney. 2023. Test accuracy vs. generalization gap: Model selection in nlp without accessing training or testing data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 3011–3021, New York, NY, USA. Association for Computing Machinery.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024a. Tinyllama: An open-source small language model.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024b. A survey on the memory mechanism of large language model based agents.