Universidade de São Paulo
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Departamento de Computação e Matemática

Comparing the LMC Complexity of Neural Networks with their Inference Capability

Author: Lucas Miranda Mendonça Rezende
Supervisor: Ph.D. Luiz Otavio Murta Junior

November 7, 2025

Abstract

Current scaling laws suggest that model performance improves with increased parameters, dataset size, and compute, but these improvements follow power laws requiring exponential resources for constant gains. Understanding the learning process through alternative metrics could enable architectural improvements and optimization during training. This work investigates whether LMC statistical complexity, a measure combining Shannon entropy and disequilibrium of weight distributions, correlates with neural network inference capability. We analyzed 35 open-weight transformer-based language models from Meta, Google, Microsoft, and OpenAI, ranging from 117 million to 150 billion parameters, computing LMC complexity across different weight-type combinations (bias, norm, embedding, other) and filtering settings. Model performance was evaluated using four benchmarks: MMLU, MMLU-Pro, OpenLLM, and LMArena. Results showed no statistically significant general correlation between LMC complexity and benchmark performance while individual benchmarks exhibited mixed results.

Contents

1 Introduction

Since the creation of Transformers in 2017 [1] and the subsequent usage of this new technique in the training of Large Language Models (LLMs), there has been a gold rush within the machine learning world. GPT-3.5, the original ChatGPT model, rapidly gained widespread adoption, becoming the fastest-growing consumer application in history after its launch in 2022 [2], sparking further interest in researchers, investors, and the general public for more powerful and cost-efficient models.

Since then, multiple new models have been developed by the biggest technology companies in the world such as Google, Microsoft, NVIDIA, Amazon, and amazingly also by some smaller ones such as DeepSeek. As better models in almost every metric emerged, it also became increasingly obvious that all this improvement was not for free: billions were spent in larger datacenters, predictions that we soon would not have enough data on the internet to keep building bigger models, increasingly expensive for marginal performance gains. But why?

In 2020, before the launch of the original ChatGPT, researchers at OpenAI [3] found that "performance improves smoothly as we increase the model size N (the number of parameters excluding embeddings), dataset size D, and amount of compute C" and that "performance depends strongly on scale and weakly on model shape, such as depth vs. width". Second, it was also observed that different architectures could impact training performance, e.g., Transformers would have lower test loss than LSTMs. Those statements became the basis for the scaling "laws" we currently accept.

It is evident that the most straightforward way to get a better model, at least considering performance in terms of lower test loss, is to increase the scale (N, C, and D sizes). However, there is a huge drawback: the scaling laws described in this case are power laws, meaning we would need an exponential amount of resources to achieve a constant gain in performance. The second approach would be creating better architectures or improving the existing ones, which demands research and is generally not as simple as increasing a number.

The first approach, being the easiest, was explored by the companies creating the new models, pushing those three variables to ludicrous amounts. The GPT family, mentioned above, serves as a good example: the original GPT had 117 million parameters, GPT-2 had 1.5 billion (a 12x increase from GPT), and GPT-3 had 175 billion (a 117x increase from GPT-2) [4] [5] [6]. It is noticeable that, just a few years later, the industry is already showing signs of exhaustion: new models do not exhibit performance improvements as dramatic as those observed in the recent past, although investment in training infrastructure has never been higher.

The second approach is what we intend to contribute in this work: an improve-

ment in the architecture itself. Of course, before creating a new revolutionary architecture it would be useful to understand how machines learn. In a typical analysis setting, knowing how the process works is a requirement to engineer a better version of it. In Machine Learning, however, the process of learning, which is somehow connected to the well-known process of training, is still far from being fully understood.

Discussing the understanding of the learning process is out of the scope of this work. Yet, we are going to analyze what may be a piece of the puzzle: the LMC statistical complexity [7], a metric that appears to be related to the model's performance and might help in understanding the behavior of such by providing insights about the weights distribution.

1.1 Work thesis

The main hypothesis of this work comes from Professor Luiz Otavio Murta Junior [8] and can be summarized as: **"There exists a relationship between model complexity and its inference capability"**.

1.2 Objective

Validating the work thesis means we would have a new way of indirectly assessing a model's performance by just looking at the distribution of its weights, opening the door to new optimization processes during training, potentially reducing the amount of compute necessary to reach the best performance or even discovering new maxima. Also, having a validated weights distribution vs. performance comparison should improve the data richness when studying a model's learning process in future studies. Thus, we can establish the following objectives for this work:

- Validate the existence of a meaningful relationship between complexity of neural network weights and their inference performance.
- If possible, find the mathematical relation between those measures.
- Explore other dimensions of the problem that can affect complexity and performance measures such as parameter count.

2 Methodology

2.1 Testing Environment

We used the following computational environment for our experiments:

- OS: Linux 6.8.0-64-generic
- CPU: 2x Intel Xeon Gold 6130 @ 2.1GHz - 32 cores / 64 threads per CPU, 22M cache each
- RAM: 512GB DIMM 2400 MHz (DDR4)
- GPU: NVIDIA Quadro P5000 (GP104GL) - 16GB VRAM
- Storage: RAID5 (2x4TB) 8TB SAS with Broadcom MR9260-8i controller + 512GB RAID0 SSD

The hardware was provided for research purposes by Universidade de São Paulo. The choice of this environment was motivated by the availability of a high-performance GPU and large quantities of RAM necessary for handling large language models.

We used the following tools and libraries:

- Python 3.12.3
- PyTorch 2.8.0 [9]
- Transformers 4.56.2 [10]
- Hugging Face Hub 0.35.0 [11]
- Numpy 2.2.6
- Pandas 2.3.2
- Pysr 1.5.9
- SciPy 1.16.1 [12]
- Matplotlib 3.10.6
- Seaborn 0.13.2 [13]
- CUDA Toolkit 12.6

The choice of versions was motivated by compatibility with our computational environment's GPU.

2.2 Model Selection

Due to the wide availability of open models combined with the ease of accessing them through the Transformers library [10] and Hub library [11], **Hugging Face** was chosen as the source for model selection in this study.

Hugging Face is a platform that hosts a variety of machine learning models, datasets, and tools. It is widely used in the AI research community for sharing and collaborating on machine learning projects [14, 15]. Other platforms such as Ollama [16] were considered, but ultimately not chosen due to the limited number of models available and fewer contributing companies.

Model selection proceeded in two stages:

1. First, we identified major technology companies by market capitalization that publish openly released language models. The companies considered were **OpenAI**, **Google**, **Meta**, and **Microsoft**.

2. Second, for each company we compiled a candidate set consisting of every model that satisfied the following criteria:

    - Must be available on the Hugging Face platform on the official company account.
    - The model is a transformer-based language model.
    - Model weights are publicly accessible (open weights), including models released behind gated access.
    - The model is text-only (no multimodal image/audio inputs).
    - The model is an original base model rather than a task-specific fine-tuned variant.
    - The total parameter count is below 150 billion. This upper bound was imposed due to hardware and inference limitations in the computational environment used for our experiments.
    - The model is supported by the Hugging Face Transformers library [10] (i.e., it can be instantiated via the AutoModel utility), which ensures consistent loading and preprocessing across the candidate set.
    - There is at least one publicly available benchmark result for the model among the selected benchmarks (section 2.4.1).

The final selection of models used in this study is listed in Table 1.

Meta	Google	Microsoft	OpenAI
meta-llama/Llama-4-Scout-17B-16E	google/gemma-3-27b-pt	microsoft/Phi-4-mini-reasoning	openai/gpt-oss-120b
meta-llama/Llama-3.2-3B	google/gemma-3-12b-pt	microsoft/Phi-4-reasoning	openai/gpt-oss-20b
meta-llama/Llama-3.2-1B	google/gemma-3-4b-pt	microsoft/Phi-4-reasoning-plus	openai-community/gpt2-xl
meta-llama/Llama-3.1-70B	google/gemma-3-1b-pt	microsoft/phi-4	openai-community/gpt2-large
meta-llama/Llama-3.1-8B	google/gemma-3-270m	microsoft/phi-2	openai-community/gpt2-medium
meta-llama/Meta-Llama-3-70B	google/gemma-2-27b	microsoft/phi-1_5	openai-community/gpt2
meta-llama/Meta-Llama-3-8B	google/gemma-2-9b	microsoft/phi-1	
meta-llama/Llama-2-70b-hf	google/gemma-2-2b		
meta-llama/Llama-2-13b-hf	google/gemma-7b		
meta-llama/Llama-2-7b-hf	google/gemma-2b		
	google/recurrentgemma-9b		
	google/recurrentgemma-2b		

Table 1: Selected language models included in this study.

Meta Models: [17] [18] [19] [20] [21] [22] [23] [24] [25] [26]
Google Models: [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37]
Microsoft Models: [38] [39] [40] [41] [42] [43] [44]
OpenAI Models: [45] [46] [47] [48] [49] [50]

2.3 LMC Complexity

According to [7], LMC Statistical Complexity is the product of two other measures: Disequilibrium and Shannon entropy. It captures both the structured and

unstructured aspects of the distribution:

$$C_{LMC} = H \times D$$

Disequilibrium measures how far a probability distribution is from being uniform, quantifying the "order" or structure in the data. It is calculated as:

$$D = \sum_{i=1}^{n} \left(p_i - \frac{1}{n} \right)^2$$

Shannon entropy measures the amount of uncertainty or randomness in a probability distribution. The normalized Shannon entropy is given by:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

The values p_i represent the probabilities associated with each state i in the distribution, and n is the total number of states. K is a positive constant and, in our case, is set to 1 for simplicity. K can be changed later since $C_{LMC} = (-K \sum_{i=1}^{n} p_i \log p_i) \times D$ is equivalent to $C_{LMC} = K \times (-\sum_{i=1}^{n} p_i \log p_i) \times D$.

2.3.1 Reading Model Weights

Model weights are fetched using the Hugging Face Hub library [11] and loaded using the Transformers library [10]. For each selected model, we instantiate the model using the **AutoModel** utility, which automatically handles model architecture loading and weight initialization.

The models are loaded entirely into the main memory instead of a GPU since the amount of VRAM available in the computational environment is insufficient for larger models ($>$ 70 billion parameters). Other problems such as handling symbolic tensors also motivated this decision.

Using the main memory is considerably slower than using a GPU, but allows us to work with larger models without running into memory limitations. Since we had a large amount of RAM, we decided to cast all the model weights to float32 during the AutoModel instantiation as CPUs generally handle this format natively and, as a consequence, make calculations faster.

Once loaded, we extract all the parameters from the model using the `named_parameter()` method provided by the Transformers library. This method returns an iterator over all model parameters, each parameter being represented as a tensor.

We then flatten each tensor into a one-dimensional array and store it in a list. Each array is labeled as one of the following **weight-type categories**:

- Bias: if 'bias' is in the parameter name
- Norm: if 'norm' is in the parameter name
- Embedding: if 'embed' is in the parameter name
- Other: all other weights

These categories are the most common types of weights found in this architecture and will be used later as another dimension of analysis.

The choice of also analyzing by weight type is motivated by the fact that different weight types may exhibit different statistical properties and, as a consequence, different complexity characteristics. Studies such as the already cited OpenAI's [3] take into account different weight types when analyzing scaling laws.

2.3.2 Filtering

Casting different encoding formats to float32 may introduce rounding errors that manifest as extreme outliers in the weight distribution. This is rare (approximately 10 to 30 values in 100 billion) but can be problematic for our next step: histogram construction (section 2.3.3), since the number of bins will be affected by the range of values in the data. The reasons for these rounding errors could not be fully investigated within the scope of this work, but they are likely related to floating point rounding errors.

To mitigate the impact of outliers, we apply a value removal approach. This process filters the data to retain only values within a configurable range centered around the mean:

$$\text{lower bound} = \mu - \sigma_{\text{filter}} \cdot \sigma$$
$$\text{upper bound} = \mu + \sigma_{\text{filter}} \cdot \sigma$$

where μ is the data mean, σ is the data standard deviation, and σ_{filter} is a configurable parameter that controls the filtering strength. Values falling outside this range are excluded from further analysis.

It is not trivial to choose the best value for σ_{filter}. A very low value may remove important parts of the distribution, while a very high value may not effectively mitigate the outlier problem. For this reason, we experiment with different values of σ_{filter} and analyze how they affect the final complexity results. As a consequence, this becomes yet another dimension of analysis in our study.

The values of σ_{filter} tested will be: **0.125**, **0.25**, **0.5**, **1**, **2**, **3**, **4**, **5**, **10**, **20**. The choice of σ_{filter} that will be used as the filtering option for the next steps is the one that:

1. most reduces the bin count compared to the unfiltered data;

2. has the least amount of filtering or, in other words, the highest value of σ_{filter} among the ones chosen in the first criterion.

2.3.3 Data Discretization and Histogram

To compute the LMC complexity of a finite array of floating-point numbers, we first construct a histogram to discretize the data into a probability distribution. This is justified as the chance of finding two identical numbers is extremely low and, as a consequence, it is hard to determine the probability of each value.

We will call the set of all data points as S, the total amount of data points as N and the number of bins they will be distributed into as n. The probabilities p_i are then calculated as $p_i = \frac{f_i}{N}$ where f_i is the frequency count of data points in bin i.

As expected, this approach revisits a classic issue in histogram-based analysis: the choice of the number of bins n will impact the resulting probability distribution, which is particularly problematic for the LMC complexity measure; variations in n can cause significant fluctuations in the final result. Selecting an inappropriate number of bins may produce misleading values, either by oversimplifying the distribution (too few bins) or by introducing noise (too many bins).

There are a variety of methods to determine n in a histogram, most of them with their own advantages and disadvantages. Commonly used methods such as **Sturges' formula** [51] and **Rice Rule** [52] rely only on the number of data points, while others like **Scott's normal reference rule** and **Freedman-Diaconis' choice** also take into account the data distribution by using standard deviation and interquartile range, respectively [53].

We chose to use the **Freedman-Diaconis' choice** as it adapts better to our needs. This is justified since N often consists of billions of numbers, and the distribution, although mostly concentrated between -1 and 1, can become sparse due to outliers and, as a consequence, requires a larger number of bins to capture its characteristics accurately. The Freedman-Diaconis rule helps mitigate the influence of outliers by using the interquartile range IQR to determine bin width h. The rule is defined as follows [54]:

$$h = \frac{2 \times IQR}{N^{1/3}}$$

where IQR is the interquartile range of the data which is calculated as $Q3 - Q1$, $Q3$ and $Q1$ are the values at the 75th and 25th percentiles in the data, respectively. Since N is very large, the percentiles are computed based on a random sample of

100000 data points to reduce computational cost. The sample size was determined empirically to be large enough to provide stable estimates of the percentiles; the final number of bins showed no variance between tests.

The number of bins n can then be calculated as:

$$n = \frac{\max(S) - \min(S)}{h}$$

Finally, we can use the already filtered data and the calculated n to build a histogram using a simple function provided by PyTorch [9]. Then, compute the probabilities p_i for each bin i.

2.3.4 Complexity Calculation

With the probabilities p_i computed from the histogram, we can now calculate the LMC complexity C_{LMC} using the formulas provided in the beginning of section 2.3. This involves calculating the Disequilibrium D and the Shannon entropy H using the probabilities, and then multiplying them to obtain the final complexity measure C_{LMC}.

2.4 Inference Capability

Often called model performance, inference capability refers to how well a trained neural network performs on unseen data. This is typically measured using various metrics depending on the specific task the model is designed for.

In the previously cited research paper by OpenAI [3], performance was associated with test cross-entropy loss. This metric quantifies the difference between the predicted probability distribution output by the model and the true distribution of the target labels. Lower cross-entropy loss values indicate better model performance, as they reflect a closer alignment between predictions and actual outcomes.

Unfortunately, not all models we intend to analyze have publicly available training and test data. It is also not possible to run models in a training setting due to performance limitations of the computational environment available. Therefore, we have to rely on imperfect proxies for performance such as **benchmarks**.

2.4.1 Benchmark Selection

Benchmarks are standardized tests designed to evaluate the performance of machine learning models across various tasks. They provide a common ground for

comparison by measuring how well different models perform on the same datasets using predefined metrics.

According to [55], benchmarks such as the famous MMLU correlate fairly well with the predicted test loss determined by scaling laws, making them suitable proxies.

They are, however, not perfect. Some benchmarks may fail to capture all aspects of a model's capabilities, leading to an incomplete assessment of performance. Other problems such as data contamination [56] can influence the validity of results and make it difficult to fairly compare models created at different times. It is also hard to find benchmarks that are widely reported for all models we intend to analyze.

A nice proposal for future research to avoid the problems cited above would involve training a model from scratch twice: optimizing it initially for minimal loss and then for minimal loss plus maximum LMC complexity, then comparing the results. This is, however, out of the scope of this work.

The benchmark selection followed three main heuristics:

- **Relevance**: The benchmark should be widely recognized and accepted in the machine learning community, e.g., used in major research papers.
- **Generality**: It should cover a range of tasks and data types to provide an assessment of model performance across different scenarios, that is, not being specialized in one task.
- **Availability**: The benchmark results should be publicly available. Benchmarks that cover more of the selected models were preferred.

Based on those heuristics, the benchmarks selected were:

- **MMLU (5-shot)**: Massive Multitask Language Understanding, a benchmark that tests models across 57 tasks spanning various subjects and difficulty levels. Widely used to evaluate LLMs [57].
- **MMLU-Pro (5-shot)**: An enhanced version of MMLU that includes additional tasks and updated datasets to provide a better evaluation [58].
- **OpenLLM (Average)**: A benchmark suite that evaluates models on a variety of tasks, including language understanding, generation, and reasoning. It aggregates results from multiple datasets to provide an overall performance score [59].
- **LMArena (Score)**: An online platform where users can submit questions and receive responses from two anonymous large language models (LLMs),

then vote on which answer they prefer, helping to crowdsource human preferences for evaluating and ranking LLMs. Its evaluation works by collecting thousands of pairwise comparisons from users, using the Bradley-Terry system to estimate win rates and compute rankings/scores [60].

2.4.2 Benchmark Collection Procedure

Benchmark values were manually collected from multiple sources, in the following order of priority:

1. Official *Hugging Face* model pages
2. Original research papers
3. Official websites
4. Third-party websites

Meta Models: [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]
Google Models: [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37]
Microsoft Models: [38], [39], [40], [41], [42], [43], [44], [61], [62]
OpenAI Models: [45], [46], [47], [48], [49], [50], [63],
Additional References: [15], [16], [64], [65],

If more than one source was available for the same model and benchmark, the one with higher priority was chosen. Not all models had results available for all benchmarks. Figure 1 shows the availability matrix.

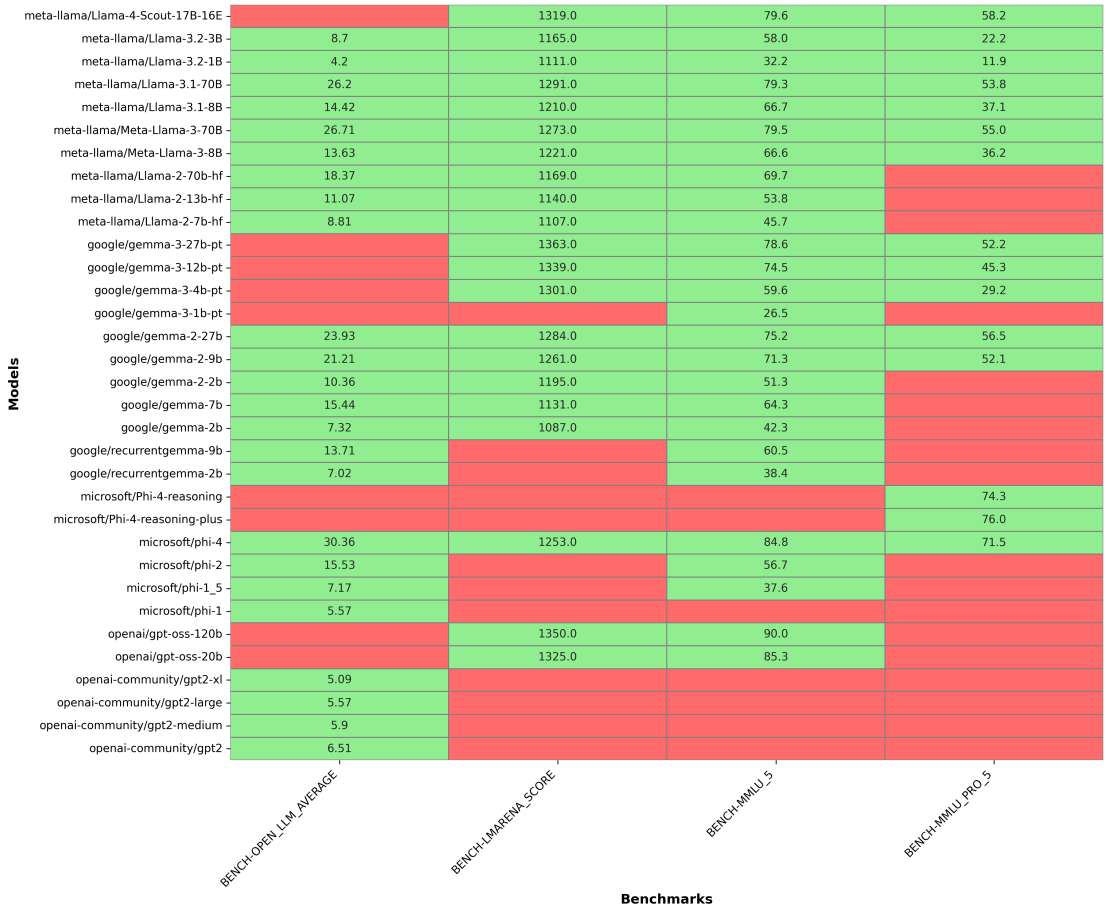

Figure 1: Availability of benchmark results for the selected models.

2.5 Comparing LMC Complexity and Inference Capability

2.5.1 Building the testing dataset

We will build a dataset by creating tuples that encode all dimensions of our analysis: model information, weight-type combination, filtering setting, the resulting LMC complexity value, the number of histogram bins used, and the available benchmark results for the model.

The dataset construction follows a nested loop structure:

1. **For each model** in our selected set (e.g., Llama-4-Scout-17B-16E, gpt-oss-120b, etc.), we generate all non-empty subsets of the four **weight-type categories** defined in section 2.3.1 using the power set [66]. This produces 15 unique weight-type combinations per model:

- Single types: {bias}, {norm}, {embedding}, {other}
- Dual combinations: {bias, norm}, {bias, embedding}, {bias, other}, {norm, embedding}, {norm, other}, {embedding, other}
- Triple combinations: {bias, norm, embedding}, {bias, norm, other}, {bias, embedding, other}, {norm, embedding, other}
- All four: {bias, norm, embedding, other}

2. **For each weight-type combination**, we apply every value of σ_{filter} defined in section 2.3.2. The filtering values tested are: 0.125, 0.25, 0.5, 1, 2, 3, 4, 5, 10, 20, plus one unfiltered configuration, totaling 11 configurations per combination.

3. **For each configuration**, we compute the LMC complexity following the procedure outlined in section 2.3. We also record the number of histogram bins used in the calculation for further analysis. Lastly, we append the available benchmark results for the model from section 2.4.2.

At the end, we will have 35 models $\times$ 15 weight-type combinations $\times$ 11 filtering settings (including unfiltered) = 5775 datapoints. The resulting dataset $\mathbf{R}$ consists of tuples structured as follows:

1. Model name
2. Number of parameters
3. Weight-type combination
4. Filtering setting
5. LMC complexity value
6. Number of histogram bins
7. Available benchmark results

e.g. a hypothetical tuple in the dataset:

- `(Llama-4-Scout-17B-16E, 17B params, {bias}, 0.125, 0.324, 8592, [MMLU: 71.2, MMLU-Pro: 68.5, ...])`

2.5.2 Regression procedures

We will also use two important concepts defined as follows:

- **Linear regression**: a conventional statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation [67]. The data may not follow a linear trend, but linear regression can still be used to identify overall correlations and make it easier to compare results.

  - Linear: $y = ax + b$

  For comparisons, we will use the **Pearson correlation** [67] (linear correlation coefficient). This can be computed as:

  $$r = \frac{\text{cov}(x, y)}{SD(x) \times SD(y)}$$

  where $\text{cov}(x, y) = (\text{average of products } xy) - (\text{average of } x) \times (\text{average of } y)$.

  Or alternatively, using the linear regression slope a [68]:

  $$r = a \cdot \frac{SD(x)}{SD(y)}$$

- **Free regression**: a curve fitting method with greater flexibility but less comparability than linear regression. It was configured to test for the following functions:

  - Linear: $y = ax + b$
  - Quadratic: $y = ax^2 + bx + c$
  - Exponential: $y = a \cdot e^{bx} + c$
  - Logarithmic: $y = a \cdot \ln(x + 1) + bx + c$
  - Power: $y = a \cdot (x + 1)^b + c$

  The best fit was chosen based on the highest R^2 value.

  For both regressions, the tests were made using `curve_fit` from SciPy [12].

2.5.3 Influence of filtering

In more detail, the criteria defined in section 2.3.2 to choose the best filtering setting work as follows:

1. We will group R in relation to each value of σ_{filter}. Then, calculate the **average** and the **maximum histogram bin count** for each group, and plot them as a bar chart.

2. We will group R in relation to each value of σ_{filter}. Then, calculate the **average** and the **maximum complexity** for each group, and plot them as a bar chart.

3. Since steps 1 and 2 represent the two criteria defined in section 2.3.2, we will first filter the groups obtained in 1 to keep only the ones that have a lower average bin count than the unfiltered data. Then, among those groups, we will choose the one with the closest complexity value to the unfiltered data (least amount of filtering).

The filtering dimension will be used to determine the best general filtering setting following the criteria defined in section 2.3.2. After defining the best filtering setting, we will discard all the runs where σ_{filter} is different from the chosen one, creating a new set $\mathbf{R_f}$. There will be a total of: 35 models $\times$ 15 weight-type combinations $\times$ 1 filtering setting $=$ 525 datapoints.

We will also make a free regression on a scatter plot for the data obtained in steps 1 and 2 to analyze how filtering strength affects histogram bin count and complexity, respectively.

2.5.4 Influence of weight types

After defining the best filtering setting, we will analyze how the choice of weight types influences the final complexity results. This will help us understand if certain weight types contribute more significantly to the overall complexity measure.

This will be done by grouping R_f in relation to each weight-type combination and calculating the **average** and the **maximum complexity** for each group, then plotting a bar chart for each metric.

It is necessary to select a best weight-type combination, like it is done in 2.5.3, because keeping all the different combinations will add bias to the analysis: a good way to visualize this problem is to imagine that a single benchmark value might be associated with 15 different complexity values for the same model, one for each weight-type combination. It also might wrongfully influence the statistical significance by adding artificial datapoints.

We are heavily inclined to choose the combination that uses all weight types except for the **embedding** weights, since this approach is commonly used in research such as in the scaling laws one by OpenAI [3]. However, the final decision will be made by removing weight types with the lowest average complexity, as they are likely to contribute less to the overall complexity measure.

After choosing the best weight-type combination, we will discard all the runs in R_f where the weight-type combination is different from the chosen one, creating a new set $\mathbf{R_{best}}$. There will be a total of: 35 models $\times 1$ weight-type combination $\times$ 1 filtering setting $= 35$ datapoints.

2.5.5 Complexity vs Number of parameters

We will analyze how LMC complexity varies with the number of model parameters. This will be done by calculating the **average complexity** for each range of parameters and plotting it as a histogram. The number of bins of this plot will be defined using the Freedman-Diaconis rule [54] to adapt to the distribution.

We will look for trends or patterns that may indicate a relationship between model size and complexity.

2.5.6 Complexity vs Number of bins

Similarly to 2.5.5, we will analyze how LMC complexity varies with the number of histogram bins used in its calculation. This will be done by calculating the **average complexity** for each range of bins and plotting it as a histogram. The number of bins of this plot will be defined using the Freedman-Diaconis rule [54] to adapt to the distribution.

We will look for trends or patterns that may indicate a relationship between histogram granularity and complexity.

2.5.7 Complexity vs Benchmark performance

Finally, we will analyze the relationship between LMC complexity and benchmark performance. This analysis aims to determine whether higher LMC complexity is associated with better model performance across different benchmarks, in order to validate or refute the hypothesis that complexity correlates with inference capability.

For each benchmark available in R_{best}, we will create scatter plots with benchmark scores on the x-axis and complexity on the y-axis.

We will perform both linear and free regression analyses (section 2.5.2) to identify trends between complexity and performance. The strength and nature of

these relationships will be assessed using Pearson correlation and R^2 values.

As a control, the experiment will be rerun using the number of model parameters instead of LMC complexity. The number of parameters is a known variable that is expected to correlate with benchmark performance according to scaling laws [3]. This will help validate our methodology and provide a baseline for comparison.

We will provide a bar chart summarizing the Pearson correlation coefficient for each benchmark analyzed for both LMC complexity vs benchmark performance and number of parameters vs benchmark performance, so that they can be easily compared. In addition, for each of these plots we will check the result of the concatenation of all benchmarks available to see if a stronger correlation arises from the larger dataset, presented as another bar in the chart.

Using the complexity's Pearson correlation results, we will check for the statistical significance of the correlations found using a two-tailed t-test [69]. The t-statistic can be computed as:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

where r is the Pearson correlation coefficient and n is the number of datapoints used in the correlation. The p-value can then be obtained from the cumulative distribution function of the t-distribution with $n-2$ degrees of freedom:

$$p = 2 \cdot (1 - F_t(|t|; n-2))$$

where F_t is the cumulative distribution function of the t-distribution with $n-2$ degrees of freedom. A significance level of 0.05 will be used to determine if the correlations are statistically significant. An approximation using the standard normal distribution was avoided due to the small sample size of some benchmarks ($n < 30$).

As a bonus, we will also provide a top 20 list sorted by the absolute value of Pearson correlation coefficients found in the analysis, indicating which (benchmark, model, weight-type) combinations produced the strongest correlations.

3 Results

3.1 Extraction process

The dataset in section 2.5.1 was built successfully through the process described.

A model with 10 billion parameters took approximately 3.5 hours to process all weight-type combinations and filtering settings using the workstation described in section 2.1. The total parameter count of all models is **652.802.782.352**, taking around **228 hours** (9.5 days) of computation time in total.

The resulting dataset contains **5511 rows**, each representing a unique combination of model, weight-type combination, filtering setting, LMC complexity value, number of histogram bins, and available benchmark results. That was 264 rows short of the expected 5775 rows due to:

- Some unfiltered data models exceeded the maximum number of bins (set to 1 billion) during histogram creation, leading to their exclusion. Increasing the maximum bin count is not possible due to memory constraints and keeping the rows in the dataset capped at 1 billion bins would have introduced bias to the analysis.
- Some numerical values such as count, min, max, mean, std, bin_count, shannon_entropy, disequilibrium, and complexity were reported as NaN or infinite values in certain model and weight-type combinations, which required their exclusion from the dataset. The reasons are unknown, but the main suspects are floating-point precision errors and divisions by zero in some edge cases.

3.2 Filter dimension

Following the methodology described in section 2.5.3, we obtained the plots for average and maximum number of histogram bins per filtering setting across all models and weight-type combinations:

> For visualization purposes, unfiltered data is represented as a **40** σ filtering setting in the plots.

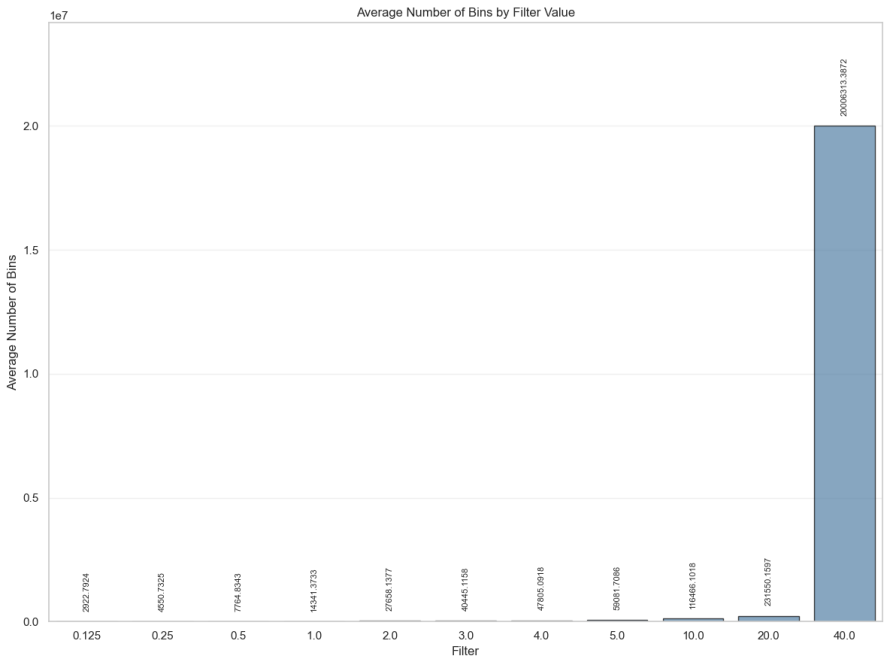

Figure 2: Average number of histogram bins per filtering setting.

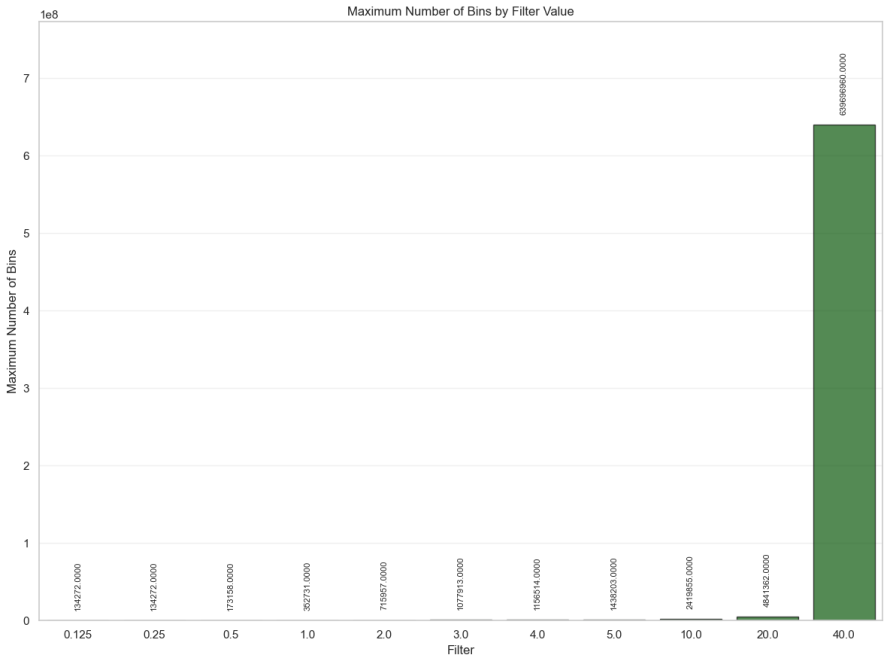

Figure 3: Maximum number of histogram bins per filtering setting.

It is clear that both plots are pretty similar in shape and proportions. There is a big jump downwards from the unfiltered setting (40 σ) to the first filtered one (20 σ), then a slow and gradual decrease in the number of bins as the filtering becomes more aggressive.

We can understand this behavior more clearly by looking at the regression lines:

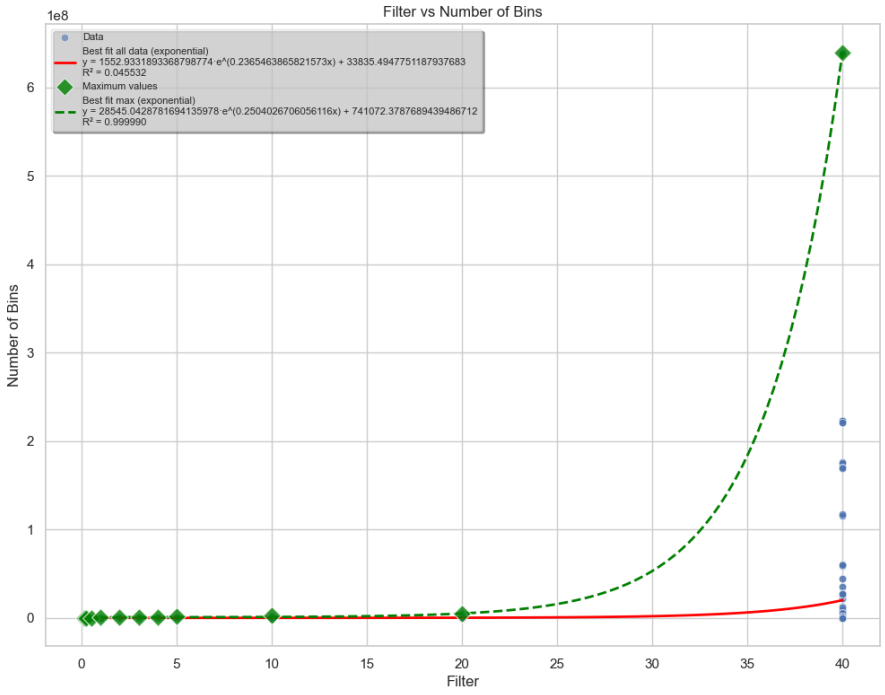

Figure 4: Average and maximum number of histogram bins per filtering setting.

The average bins show a weak fit with $R^2 = 0.045$, indicating high variability. In contrast, the maximum bins demonstrate an exceptional fit with $R^2 = 0.999$. This is expected, as the maximum bins are only one data point per filtering setting, making it easier to fit a curve through them. The fact that both curves follow an **exponential trend**, however, is noteworthy.

The maximum bins (green dashed line) explodes faster than the average bins (red line) as the filtering becomes less aggressive, meaning that a worst-case scenario grows more rapidly than the average case. Mathematically, while maximum bins grow at a rate of $e^{0.250x}$, average bins only grow $e^{0.236x}$.

We can also analyze the filtering setting using our second criterion: complexity values. We obtained the following plots for average and maximum LMC complexity values per filtering setting across all models and weight-type combinations:

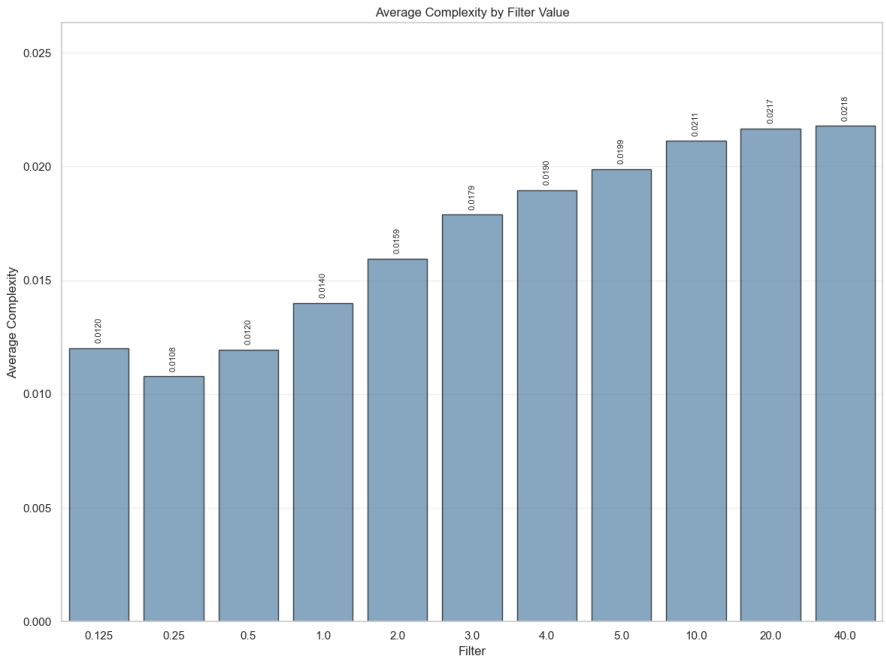

Figure 5: Average complexity per filtering setting.

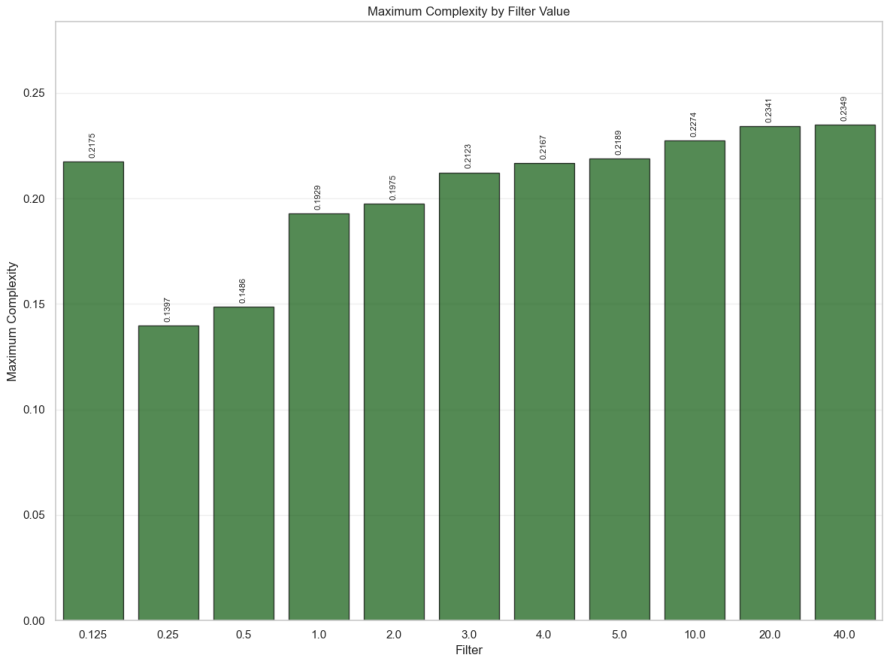

Figure 6: Maximum complexity per filtering setting.

We can see that the shapes of the plots are similar to each other but follow a different trend compared to the previous figures 2 and 3. Except for the spike at 0.125 σ, they appear to follow a **logarithmic trend**.

Overall, the maximum values tend to be a bit more flat and unstable, while the average values show a smoother curve. The shapes are not as simple as the previous figures 2 and 3 since it does not follow a monotonic downward trend. For reasons unknown, the 0.25 σ bar is the global minimum as we can observe an upward trend from 0.25 to 0.125 σ.

Complexity values in 10 and 20 σ are almost identical to the unfiltered setting (40 σ), followed by a gradual decrease that accelerates as the filtering becomes more aggressive until reaching the global minimum.

We can take a look at the regression lines to understand this behavior better:

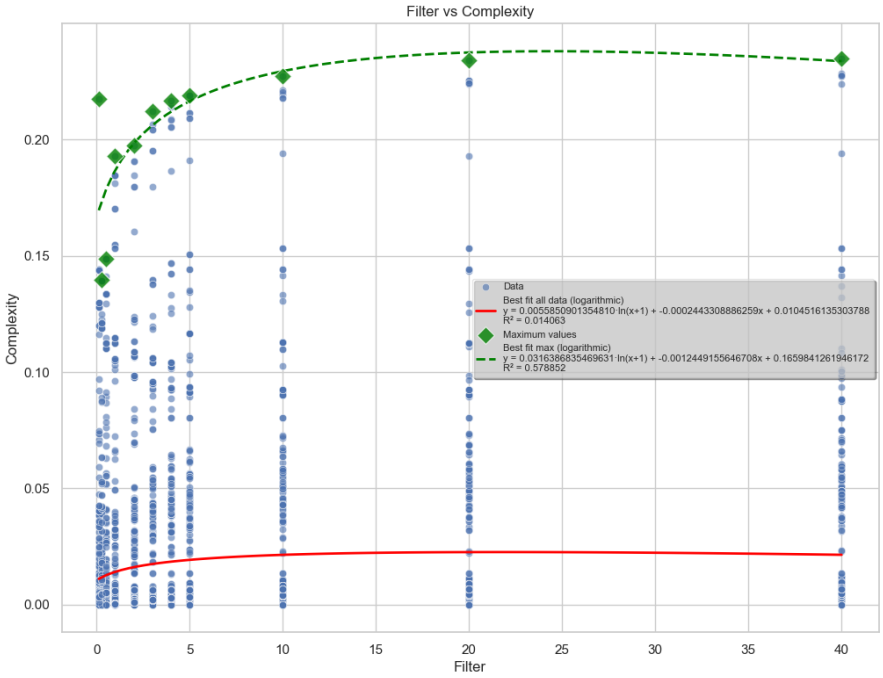

Figure 7: Average and maximum number of complexity per filtering setting.

As expected, both average (red line) and maximum (green dashed line) complexity values follow a **logarithmic trend**, with $R^2 = 0.014$ and $R^2 = 0.578$ respectively. The maximum fit is much better than the average fit due to the same reasons as explained in figure 4. The maximum fit here is, however, much worse than the maximum fit in figure 4, probably due to the unusual spike at 0.125 σ.

Both curves show a downward trend as the filtering becomes more aggressive

that accelerates the more we approach 0 σ, where all the values are removed. This acceleration is faster in the maximum complexity fit.

We also observe a very slight downward concavity in both curves approaching the unfiltered setting (40 σ) since the x terms are slightly negative. For now, we are considering this an artifact of the regression that happens as a consequence of 40 σ being an approximation of the unfiltered data even though it is very close.

Also, it looks like using even less aggressive filtering settings (such as 30 σ) could be a better choice, but concluding this would require further experiments with more filtering settings between 20 σ and the unfiltered data. Further experiments could also be conducted to understand the intriguing spike in complexity at 0.125 σ and the behavior between it and totally filtered data (0 σ). Both of those points are left as future work.

Analyzing the criteria together, we can conclude that the best filtering setting is **20** σ, as it provides a significant reduction in the number of histogram bins while maintaining an almost identical complexity value compared to the unfiltered data.

3.3 Weight-type dimension

Following the methodology described in section 2.5.4, we obtained the plots for average and maximum LMC complexity values per weight-type combination across all models using the best filtering setting (20 σ):

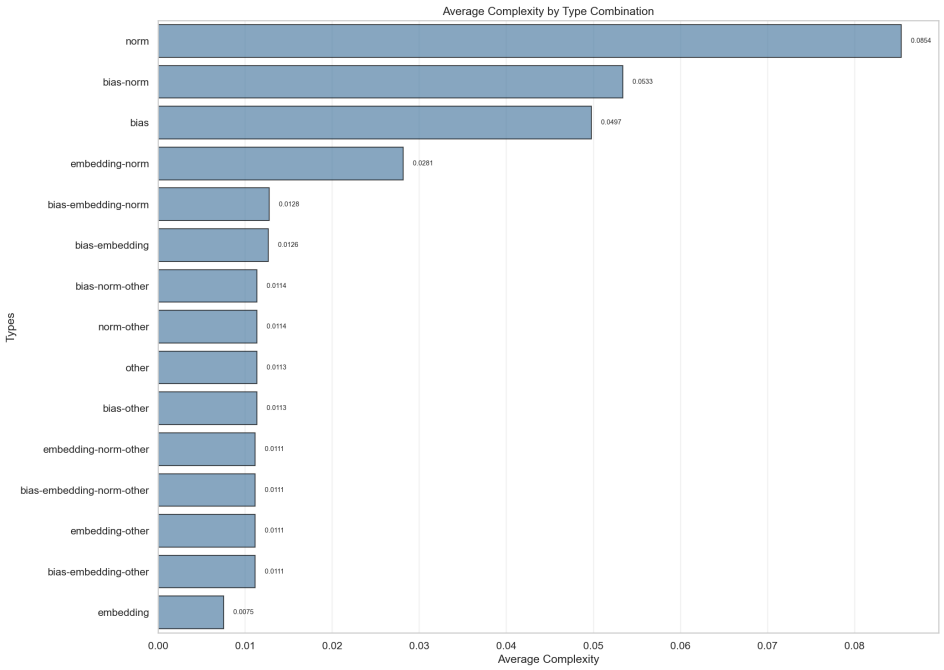

Figure 8: Average complexity per weight-type combination.

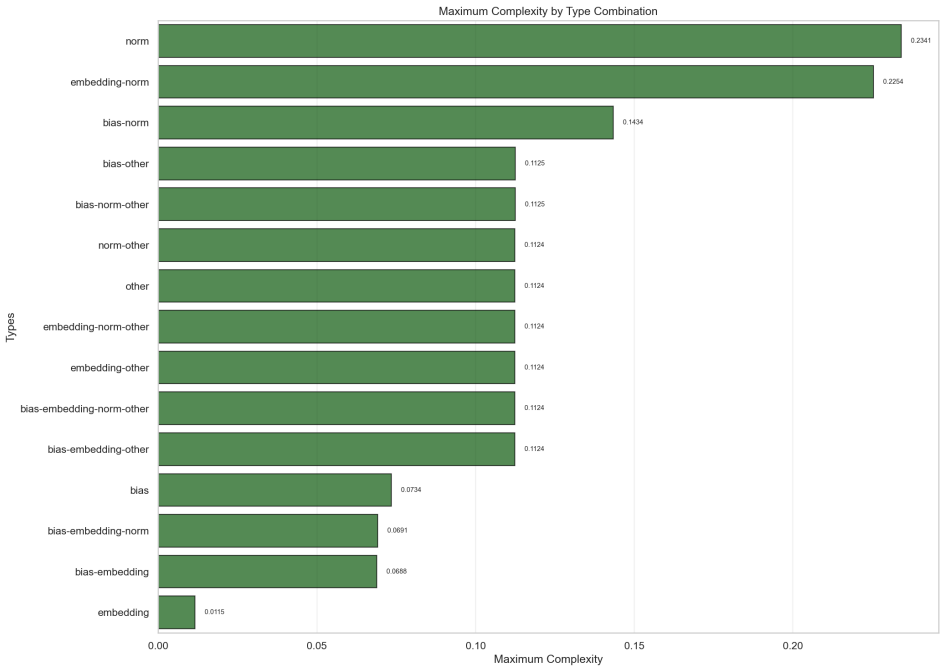

Figure 9: Maximum complexity per weight-type combination.

Looking at both plots, we can see that the shapes appear to be similar to each other. For a more careful analysis we will divide the ranking in multiple weight types (e.g. bias-embedding; norm-other) and single weight types (e.g. bias; other).

We could not find a consistent pattern while analyzing the multiple weight types. It is obvious, however, that classes that include **other** have similar values since **other** weights are the majority of parameters in a model.

Analyzing the single weight types, it is possible to see that while **norm** is responsible for the highest complexity values in any case and **embedding** is responsible for the lowest in any case, near zero. **Bias** and **other** weights vary between second and third place depending on the metric (average or maximum).

It is unclear why this ranking occurs. As a consequence, it would be interesting to further investigate if training a model while optimizing the complexity value of certain weight types (e.g. norm) could lead to better performance or generalization. However, this is also left as future work.

We will proceed with the analysis using all weight types combined except **embedding** (the **bias-norm-other** combination). This is justified as **embedding** has the lowest complexity values and is not likely to contribute significantly to the overall complexity of the model. Also because removing them allows us to follow the same methodology as the scaling laws study [3].

3.4 Complexity vs Number of parameters

Following the methodology described in section 2.5.5, we obtained the following plot for average and maximum LMC complexity values against the number of parameters across the bias-norm-other weight-type combination using the best filtering setting (20 σ):

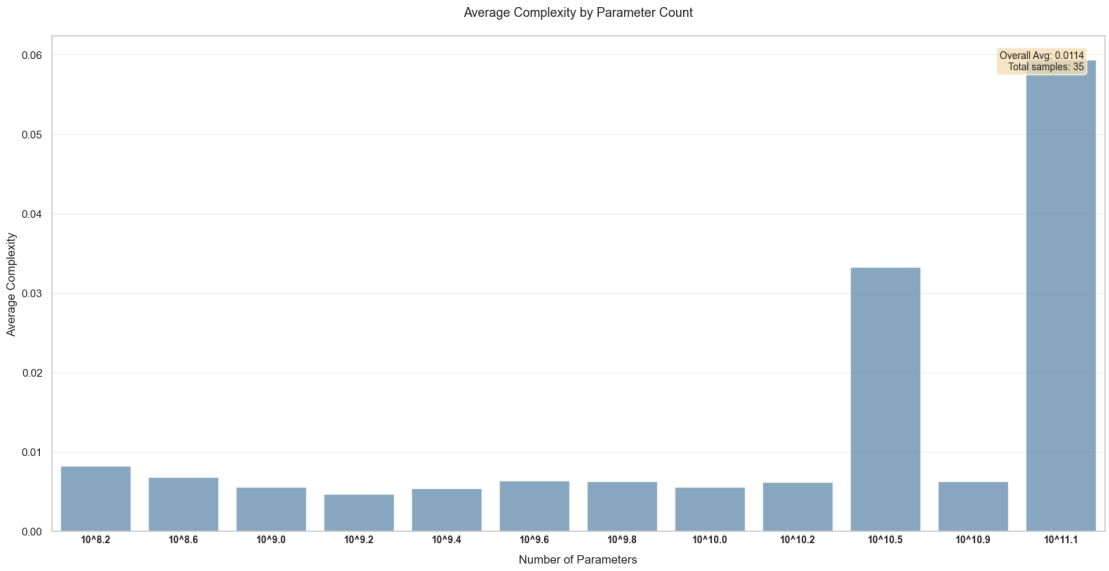

Figure 10: Average complexity vs number of parameters.

A clear trend emerges from the histogram: as the number of parameters increases, the complexity values tend to stay mostly flat, with a sudden spike at the end (around $10^{10.5}$ to $10^{11.1}$ parameters).

This indicates that complexity values tend to remain stable with the amount of parameters. The spike at the end is, however, unexpected and its causes are unknown. Further investigation is required to understand if this is an artifact of the dataset or a genuine phenomenon.

Since number of parameters is a known predictor of model performance [3], we can conclude right away that the high average complexity values found in the spike region, genuine or not, are likely to induce a bias towards better performance in the next analysis.

3.5 Complexity vs Number of bins

Following the methodology described in section 2.5.6, we obtained the following plot for average LMC complexity values against the number of histogram bins across the bias-norm-other weight-type combination using the best filtering setting (20 σ):

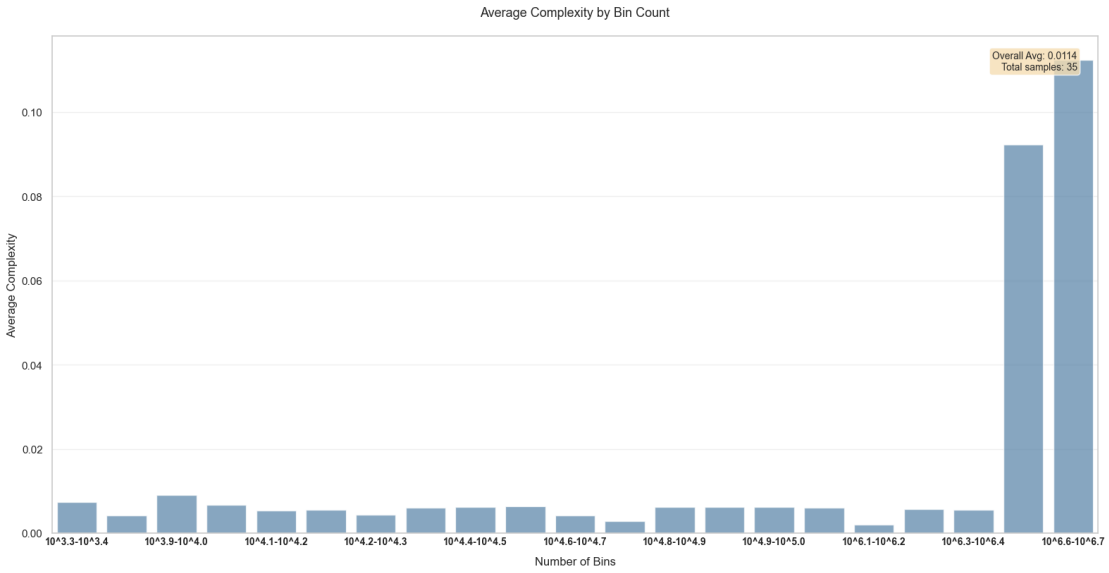

Figure 11: Average complexity vs number of histogram bins.

Again, a clear trend emerges, almost identical to the previous section (3.4). That indicates that there may be a strong correlation between the number of parameters and the number of histogram bins. A possible explanation is that larger models tend to have a wider range of weight values, leading to more histogram bins when using the Freedman-Diaconis choice [54].

3.6 Complexity vs Benchmarks

Following the methodology described in section 2.5.7, using the bias-norm-other combination and the best filtering setting (20 σ) dataset, we made a control plot using a known correlation: parameter count vs benchmark performance. The results are as follows:

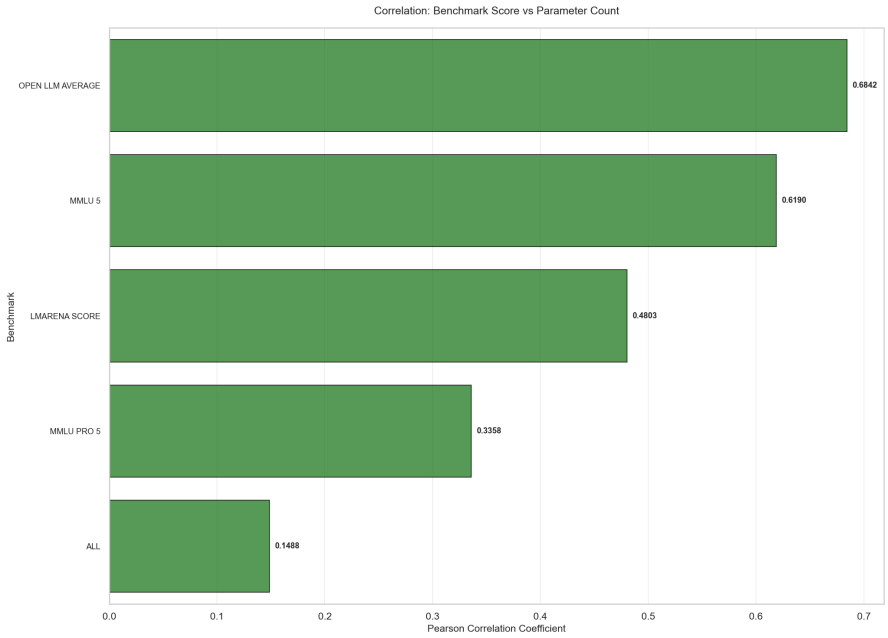

Figure 12: Pearson correlation for parameter count vs benchmark performance.

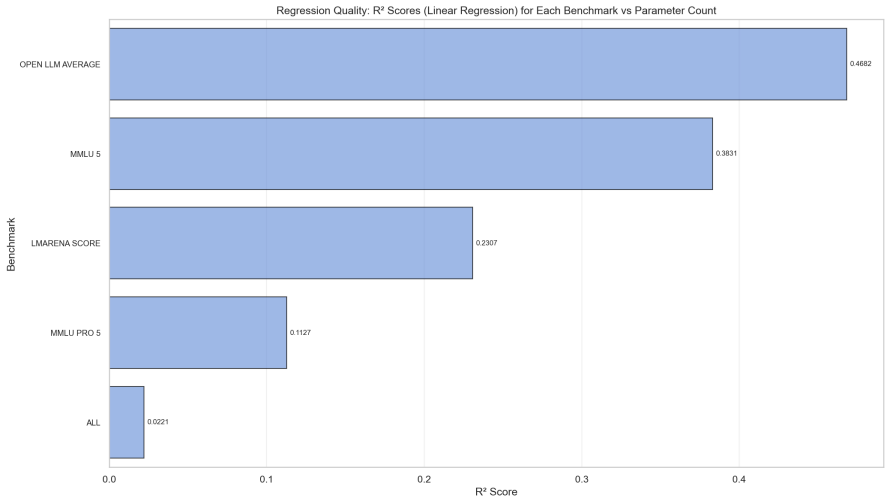

Figure 13: R^2 values for parameter count vs benchmark performance.

As expected, all the benchmarks show a positive correlation with parameter count, validating the control set. It is also noticeable that the linear regression R^2 values are relatively low for the concatenation of all benchmarks (**all**), indicating a non-linear relationship between parameter count and benchmark performance.

Now, we can analyze the actual relationship of interest: LMC complexity vs benchmark performance. The results are as follows:

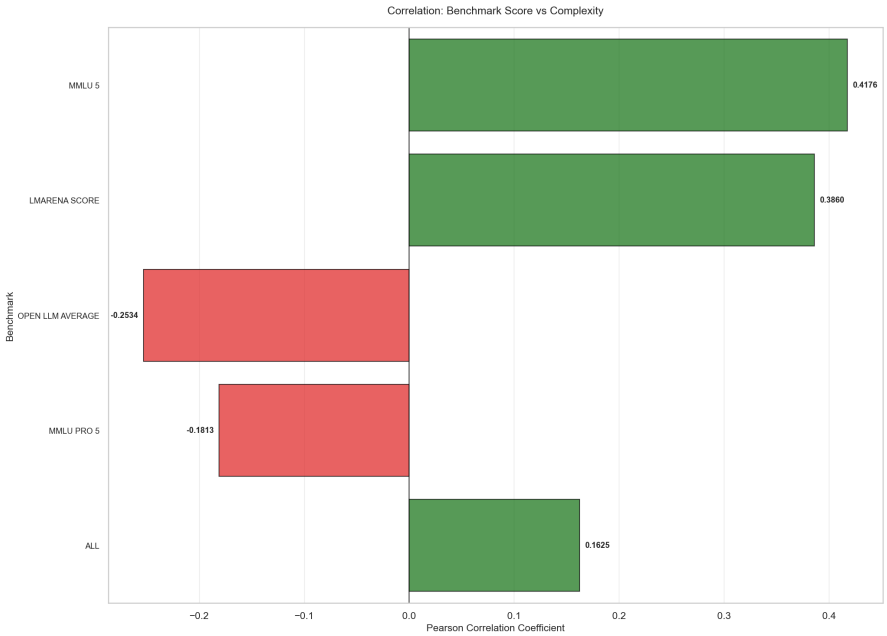

Figure 14: Pearson correlation for Complexity vs benchmark performance.

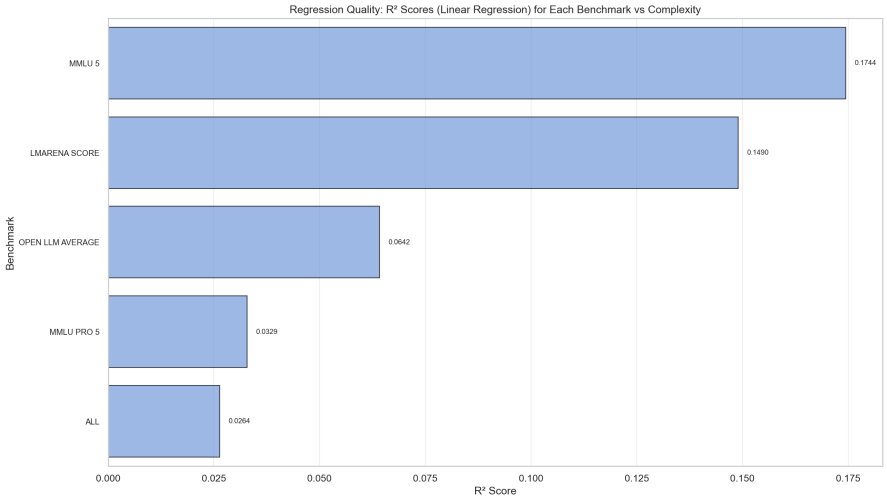

Figure 15: R^2 values for complexity vs benchmark performance.

It is immediately clear that results are inconsistent across benchmarks. While

LMArena and **MMLU** show a positive correlation between complexity and performance, **MMLU-Pro** and **OpenLLM** show a negative correlation.

For individual benchmarks, the negative correlations have a smaller absolute value compared to the positive correlations, indicating a bias towards positive correlation. The correlation values are lower than the control's in all cases. The R^2 values are also lower than the control's in all cases, which might indicate an even less linear relationship between complexity and benchmark performance compared to control. It can also indicate that complexity is not a good predictor of benchmark performance.

However, the **all** correlation, formed by analyzing a regression aggregating all data points from all benchmarks, shows an impressive positive correlation of 0.1625 which is higher than the control's 0.1488. The R^2 scores are still low but also slightly higher than the control's (0.0264 vs 0.0221).

This is an interesting result: while individual benchmarks show inconsistent correlations and score consistently worse than control, the aggregation of all benchmarks is better than control. This suggests that there may be a general trend of increasing benchmark performance with increasing LMC complexity when considering a diverse set of benchmarks. Since benchmarks are an imperfect measure of model performance, it is possible that individual benchmark inconsistencies are smoothed out.

We also found the statistical significance of the correlations found using t-tests [69]:

Benchmark	**r**	**n**	**t**	**p-value**	**Significance**
LMArena	0.3860	21	1.8239	0.0839	No
MMLU	0.4176	26	2.2515	0.0338	Yes
MMLU-Pro	-0.1813	15	-0.6647	0.5179	No
OpenLLM	-0.2534	24	-1.2287	0.2322	No
all	0.1625	86	1.5094	0.1349	No

Table 2: Statistical significance of Pearson correlations between complexity and benchmark performance.

It is possible to observe that only the MMLU correlation is statistically significant at a 0.05 significance level. The most important correlation, however, the **all** correlation, is not.

The margins of error are quite high due to the small sample sizes (n ¡ 30), which makes it difficult to draw strong conclusions from the correlations found. Even with small sample sizes, however, we can see that the correlations found are not that weak: **LMArena** and **all** were close to significance despite having only

21 and 86 datapoints respectively.

Interestingly, the benchmarks that are the furthest from significance are the ones with negative correlation (**MMLU-Pro** and **OpenLLM**), further indicating that there is more likely a positive bias towards correlation between complexity and benchmark performance than the other way around.

To better understand the shape of the relationship, we can look at the free and linear regression lines for the complexity plots:

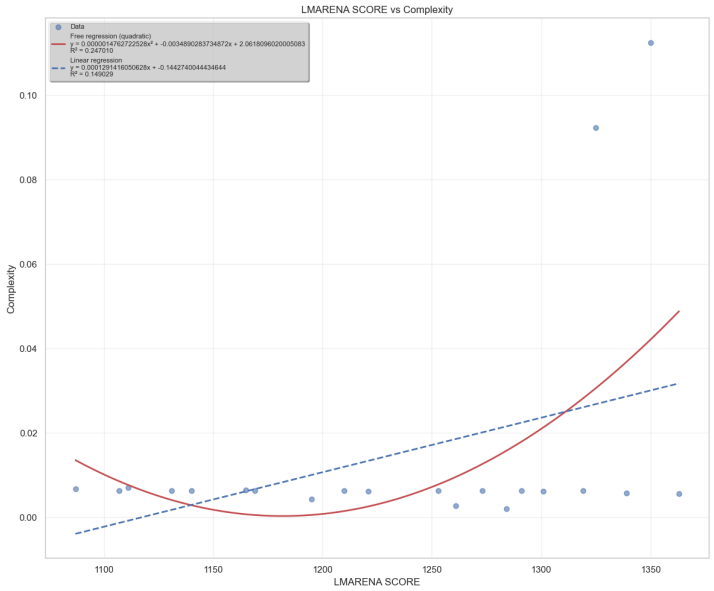

Figure 16: LMC complexity vs LMArena benchmark.

Here it is possible to see that the datapoints follow an almost perfect constant trend with two outliers that drive the positive correlation. This is likely related to the result seen in section 3.4, where complexity values remain mostly flat with a sudden spike at the end.

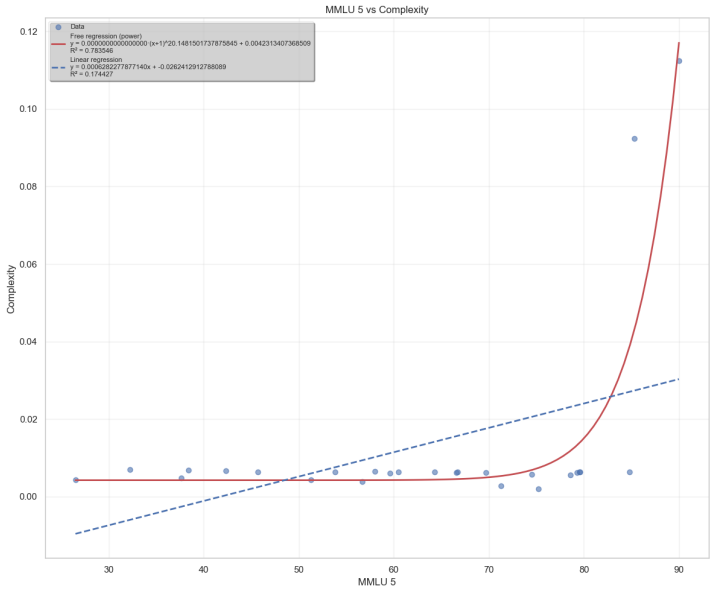

Figure 17: LMC complexity vs MMLU benchmark.

Again, a nearly constant trend is observed with a sudden spike at the end driving the positive correlation. Here we can see that the free regression fits an **exponential**, which can be the explanation for the apparently constant trend in both this and figure 16 since exponentials can appear almost constant for small x values.

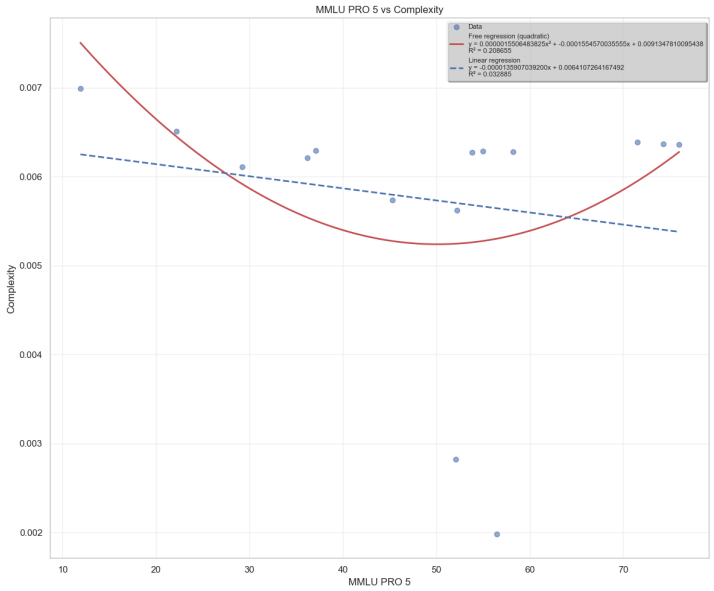

Figure 18: LMC complexity vs MMLU-Pro benchmark.

Here we see a different behavior: a slightly downward linear trend with high variability and two outliers in the middle driving the negative correlation.

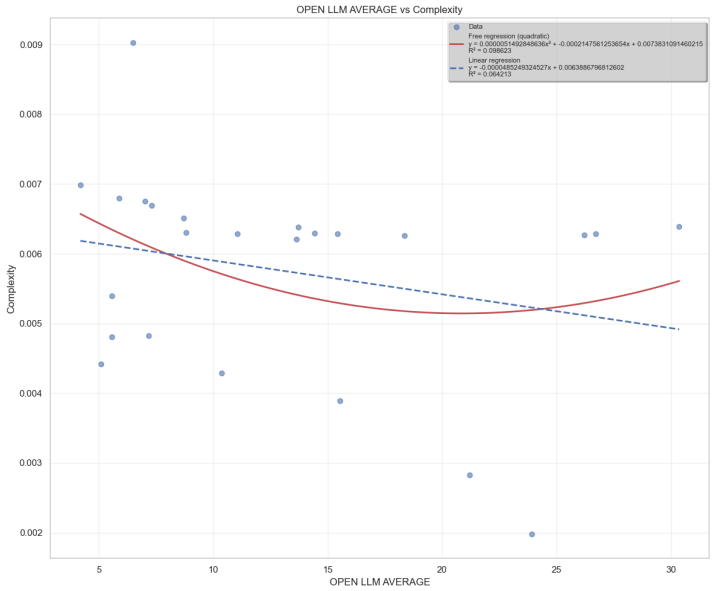

Figure 19: LMC complexity vs OpenLLM benchmark.

And here the most interesting behavior is observed: two lines are formed,

one with a slight logarithmic upward trend and another with a slight logarithmic downward trend, both regression fits trying to find a middle ground between them.

In all cases, we can see that the free regressions do not appear to be able to capture the behavior of the datapoints well due to their outliers. The linear regressions are also influenced by the same factor, leading to low R^2 values. This is a further indicator that LMC complexity appears not to be a reliable predictor of benchmark performance.

As a bonus, for the sake of completeness, we made a bar plot ranking the top 20 configurations (model + weight-type combination) sorted by highest correlation between LMC complexity and benchmark performance (all benchmarks aggregated):

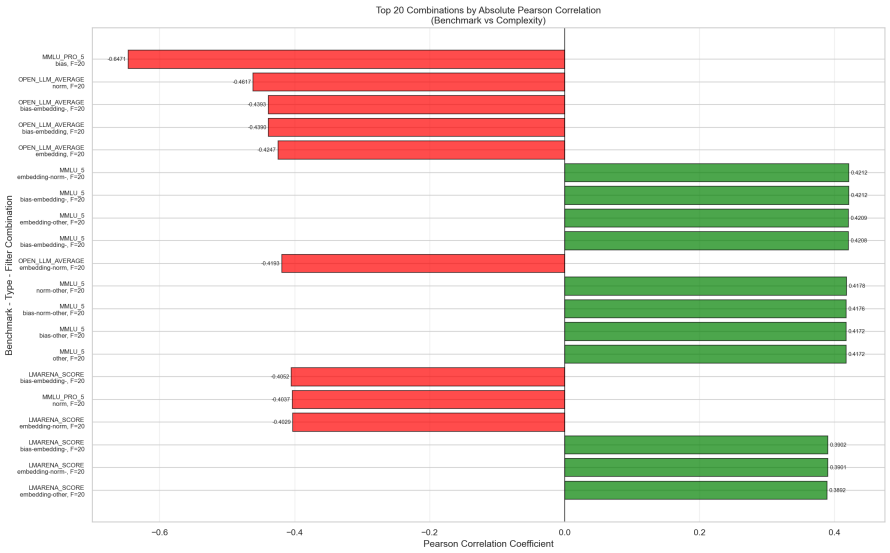

Figure 20: Top 20 configurations by Pearson correlation.

4 Conclusion

In this work, we investigated the relationship between the LMC statistical complexity of neural network weights and their inference performance across various benchmarks. Our analysis revealed that, for the amount of data we had available, there was **no consistent and significant general correlation** between LMC complexity and benchmark performance. In order to validate or falsify the initial hypothesis, a new study with a larger dataset would be necessary.

Individually, some benchmarks such as **LMArena** and **MMLU** showed a significant positive correlation with LMC complexity, while others like **MMLU-Pro** exhibited a significant negative correlation. These and other facts such as the inconsistency in the best fits between the four benchmarks analyzed indicate that LMC complexity alone, even if we were able to validate a significant small correlation, might not be a reliable predictor of neural network inference capability.

Another intriguing observation was that in section 3.4 we observed a general trend of decreasing LMC complexity with increasing number of parameters. The number of parameters has a known relationship with model performance: the more parameters, the better the performance. As a consequence, it was expected that LMC complexity would correlate negatively with performance, but in reality it showed a small non-significant bias to positive correlation.

A hypothesis that could be investigated in future work is whether complexity measures how close we are to the performance ceiling given by the scaling laws, instead of measuring the performance itself. Based on the observations made in this work, it is still possible that this is true and it might provide a reduction in the amount of compute needed for training by providing a good early stopping criterion.

Overall, while our findings appear to not support the initial hypothesis, they open avenues for further research into the nuanced relationship between complexity measures and neural network performance.

References

[1] A. Vaswani et al., "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] T. Trust, J. Whalen, and C. Mouza, "Editorial: Chatgpt: Challenges, opportunities, and implications for teacher education," *Contemporary Issues in Technology and Teacher Education*, vol. 23, no. 1, pp. 1–23, 2023. [Online]. Available: https://www.learntechlib.org/primary/p/222408/

[3] J. Kaplan et al., "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[4] T. B. Brown et al., "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[7] R. Lopez-Ruiz, H. Mancini, and X. Calbet, "A statistical measure of complexity," *Physics Letters A*, vol. 209, no. 5-6, pp. 321–326, 1995.

[8] L. O. Murta, *Relationship between model complexity and inference capability*, Personal communication and ongoing research, 2025.

[9] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

[10] T. Wolf et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019. [Online]. Available: https://arxiv.org/abs/1910.03771

[11] H. Face, *Huggingface_hub: The hugging face hub python library*, https://huggingface.co/docs/huggingface_hub, 2021.

[12] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haagerud, A. Reddy, et al., "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2 [Online]. Available: https://www.nature.com/articles/s41592-019-0686-2

[13] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021 [Online]. Available: https://doi.org/10.21105/joss.03021

[14] U. R. Pol, P. S. Vadar, and T. T. Moharekar, *Hugging face revolutionizing ai and nlp*, https://www.researchgate.net/profile/Tejashree-Moharekar/publication/383497950, 2024.

[15] *Hugging face – the ai community building the future*, https://huggingface.co/, 2024.

[16] *Ollama – get up and running with large language models locally*, https://ollama.com/, 2024.

[17] *Meta llama 4 scout 17b 16e*, https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E, 2024.

[18] *Meta llama 3.2 3b*, https://huggingface.co/meta-llama/Llama-3.2-3B, 2024.

[19] *Meta llama 3.2 1b*, https://huggingface.co/meta-llama/Llama-3.2-1B, 2024.

[20] *Meta llama 3.1 70b*, https://huggingface.co/meta-llama/Llama-3.1-70B, 2024.

[21] *Meta llama 3.1 8b*, https://huggingface.co/meta-llama/Llama-3.1-8B, 2024.

[22] *Meta llama 3 70b*, https://huggingface.co/meta-llama/Meta-Llama-3-70B, 2024.

[23] *Meta llama 3 8b*, https://huggingface.co/meta-llama/Meta-Llama-3-8B, 2024.

[24] *Meta llama 2 70b hf*, https://huggingface.co/meta-llama/Llama-2-70b-hf, 2024.

[25] *Meta llama 2 13b hf*, https://huggingface.co/meta-llama/Llama-2-13b-hf, 2024.

[26] *Meta llama 2 7b hf*, https://huggingface.co/meta-llama/Llama-2-7b-hf, 2024.

[27] *Google gemma 3 27b pt*, https://huggingface.co/google/gemma-3-27b-pt, 2024.

[28] *Google gemma 3 12b pt*, https://huggingface.co/google/gemma-3-12b-pt, 2024.

[29] *Google gemma 3 4b pt*, https://huggingface.co/google/gemma-3-4b-pt, 2024.

[30] *Google gemma 3 1b pt*, https://huggingface.co/google/gemma-3-1b-pt, 2024.

[31] *Google gemma 2 27b*, https://huggingface.co/google/gemma-2-27b, 2024.

[32] *Google gemma 2 9b*, https://huggingface.co/google/gemma-2-9b, 2024.

[33] *Google gemma 2 2b*, https://huggingface.co/google/gemma-2-2b, 2024.

[34] *Google gemma 7b*, https://huggingface.co/google/gemma-7b, 2024.

[35] *Google gemma 2b*, https://huggingface.co/google/gemma-2b, 2024.

[36] *Google recurrentgemma 9b*, https://huggingface.co/google/recurrentgemma-9b, 2024.

[37] *Google recurrentgemma 2b*, https://huggingface.co/google/recurrentgemma-2b, 2024.

[38] *Microsoft phi 4 mini reasoning*, https://huggingface.co/microsoft/Phi-4-mini-reasoning, 2024.

[39] *Microsoft phi 4 reasoning*, https://huggingface.co/microsoft/Phi-4-reasoning, 2024.

[40] *Microsoft phi 4 reasoning plus*, https://huggingface.co/microsoft/Phi-4-reasoning-plus, 2024.

[41] *Microsoft phi 4*, https://huggingface.co/microsoft/phi-4, 2024.

[42] *Microsoft phi 2*, https://huggingface.co/microsoft/phi-2, 2024.

[43] *Microsoft phi 1.5*, https://huggingface.co/microsoft/phi-1_5, 2024.

[44] *Microsoft phi 1*, https://huggingface.co/microsoft/phi-1, 2024.

[45] *Openai gpt oss 120b*, https://huggingface.co/openai/gpt-oss-120b, 2024.

[46] *Openai gpt oss 20b*, https://huggingface.co/openai/gpt-oss-20b, 2024.

[47] *Openai gpt-2 xl*, https://huggingface.co/openai-community/gpt2-xl, 2024.

[48] *Openai gpt-2 large*, https://huggingface.co/openai-community/gpt2-large, 2024.

[49] *Openai gpt-2 medium*, https://huggingface.co/openai-community/gpt2-medium, 2024.

[50] *Openai gpt-2*, https://huggingface.co/openai-community/gpt2, 2024.

[51] H. A. Sturges, "The choice of a class interval," *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, 1926. [Online]. Available: https://www.jstor.org/stable/2965501

[52] J. Moral De La Rubia, "Rice university rule to determine the number of bins," *Open Journal of Statistics*, 2024. DOI: 10.4236/ojs.2024.141006 [Online]. Available: https://www.scirp.org/pdf/ojs_2024022914191399.pdf

[53] K. H. Knuth, "Optimal data-based binning for histograms," *arXiv preprint arXiv:physics/0605197*, 2006. [Online]. Available: https://arxiv.org/abs/physics/0605197

[54] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L_2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, pp. 453–476, 1981. [Online]. Available: https://link.springer.com/content/pdf/10.1007/BF01025868.pdf

[55] D. Owen, "How predictable is language model benchmark performance?" *arXiv preprint arXiv:2401.04757*, 2024. [Online]. Available: https://arxiv.org/abs/2401.04757

[56] I. Magar and R. Schwartz, "Data contamination: From memorization to exploitation," *arXiv preprint arXiv:2203.08242*, 2022. [Online]. Available: https://arxiv.org/abs/2203.08242

[57] D. Hendrycks et al., "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020. [Online]. Available: https://arxiv.org/abs/2009.03300

[58] Y. Wang et al., "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," *arXiv preprint arXiv:2406.01574*, 2024. [Online]. Available: https://arxiv.org/abs/2406.01574

[59] A. Myrzakhan, S. M. Bsharat, and Z. Shen, "Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena," *arXiv preprint arXiv:2406.07545*, 2024. [Online]. Available: https://arxiv.org/abs/2406.07545

[60] W.-L. Chiang et al., "Chatbot arena: An open platform for evaluating llms by human preference," *arXiv preprint arXiv:2403.04132*, 2024. [Online]. Available: https://arxiv.org/abs/2403.04132

[61] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: Phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023. [Online]. Available: https://arxiv.org/abs/2309.05463

[62] M. Javaheripi and S. Bubeck, *Phi-2: The surprising power of small language models*, https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/, 2023.

[63] OpenAI, *Open models*, https://openai.com/open-models/, 2024.

[64] *Llm explorer – compare language models*, https://llm-explorer.com/, 2024.

[65] *Ranked agi*, https://rankedagi.com/, 2024.

[66] G. Cantor, "Beiträge zur begründung der transfiniten mengenlehre," *Mathematische Annalen*, vol. 46, no. 4, pp. 481–512, 1895. [Online]. Available: https://webhomes.maths.ed.ac.uk/~v1ranick/papers/cantor1.pdf

[67] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th. W. W. Norton & Company, 2007.

[68] T. H. Wonnacott and R. J. Wonnacott, *Introductory Statistics*, English, 5th. Wiley, Jan. 1990, p. 736, ISBN: 978-0471615187.

[69] R. A. Fisher, "On the probable error of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921. [Online]. Available: https://digital.library.adelaide.edu.au/items/002ad8fb-c23c-407b-8a89-f036d8da6030