

# Comparing the LMC Complexity of Neural Networks with their Inference Capability

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

University of São Paulo (USP)  
Faculty of Philosophy, Sciences and Letters of Ribeirão Preto

2025

# Outline

- 1 Introduction
- 2 Methodology
- 3 Results
- 4 Conclusion

# Context: The Era of Large Language Models

- **Transformers (2017)**: Revolutionized NLP, enabling massive parallelization.
- **Rapid Adoption**: GPT-3.5 (ChatGPT) became the fastest-growing consumer app.
- **Scaling Laws [?]**:
  - Performance depends strongly on scale ( $N$ ,  $D$ ,  $C$ ) and weakly on shape.
  - Follows **Power Laws**:  $L(N) \approx (N_c/N)^\alpha$ .
- **The Cost**: Exponential increase in resources (Compute, Data, Parameters) required for constant linear gains in performance.

# Problem Statement and Thesis

## The Problem:

- Current scaling is resource-intensive and showing signs of diminishing returns.
- Understanding the "learning" process is crucial for architectural improvements.

## Work Thesis

"There exists a relationship between model complexity and its inference capability." [?]

## Objectives:

- 1 Validate if LMC statistical complexity of weights correlates with inference performance.
- 2 Analyze the influence of other dimensions: Parameter count, Weight types, and Filtering.

## Hardware Constraints:

- **RAM:** 512GB DDR4 (Crucial for loading large models).
- **GPU:** NVIDIA Quadro P5000 (16GB) - Insufficient for inference of  $\geq 70$ B models.
- **CPU:** 2x Intel Xeon Gold 6130 (64 threads).

## Implication:

- Models loaded in **Main Memory (CPU)** cast to `float32`.
- Inference not possible locally; reliance on reported benchmarks.

# Model Selection Strategy

**Source:** Hugging Face (Open Weights).

**Selection Criteria:**

- Transformer-based, Text-only, Base models (no fine-tunes).
- Parameter count  $< 150$  Billion.
- Supported by AutoModel utility.

**Selected Models (35 Total):**

- **Meta:** Llama 2, 3, 3.1, 3.2, 4 (Scout).
- **Google:** Gemma 1, 2, 3, RecurrentGemma.
- **Microsoft:** Phi-1, 1.5, 2, 4 (Mini/Reasoning).
- **OpenAI:** GPT-2 (Small to XL), GPT-OSS (120B, 20B).

Defined by Lopez-Ruiz, Mancini, and Calbet (1995) [?]:

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

## 1. Disequilibrium ( $D$ ):

- Measures distance from uniform distribution (Order).
- $D = \sum_{i=1}^n (p_i - \frac{1}{n})^2$

## 2. Shannon Entropy ( $H$ ):

- Measures uncertainty or randomness.
- $H = -K \sum_{i=1}^n p_i \log p_i$

*Interpretation:* High complexity requires both structure (high  $D$ ) and information content (high  $H$ ).

- ① **Weight Extraction:** Flatten tensors from `named_parameters()`.
- ② **Filtering:** Remove outliers caused by float32 casting.
  - Range:  $\mu \pm \sigma_{\text{filter}} \cdot \sigma$ .
  - Tested  $\sigma_{\text{filter}} \in \{0.125, \dots, 20, 40(\text{unfiltered})\}$ .
- ③ **Discretization (Histogram):**
  - **Freedman-Diaconis Rule:**  $h = \frac{2 \times IQR}{N^{1/3}}$ .
  - Adapts to distribution spread and sample size ( $N$ ).
  - Crucial for stable probability ( $p_i$ ) calculation.
- ④ **Calculation:** Compute  $C_{LMC}$  from histogram probabilities.



# Inference Capability: Benchmarks

Used as proxies for Test Loss (Performance).

Benchmark	Description
<b>MMLU</b>	57 tasks, STEM/Humanities. Standard for LLMs.
<b>MMLU-Pro</b>	Enhanced MMLU, harder reasoning.
<b>OpenLLM</b>	Aggregated score of multiple datasets.
<b>LMarena</b>	Crowdsourced Elo ratings based on human preference.

**Data Collection:** Manually aggregated from Hugging Face, Papers, and Leaderboards.

We constructed a dataset of  $\approx 5500$  tuples to analyze:

- ① **Filtering Setting:** How  $\sigma$  affects bin count and complexity.
- ② **Weight Types:**
  - Categories: Bias, Norm, Embedding, Other.
  - Combinations: Power set (15 combinations).
- ③ **Parameter Count:** Relation to complexity.
- ④ **Performance:** Correlation between  $C_{LMC}$  and Benchmarks.

## Statistical Tools:

- Pearson Correlation ( $r$ ).
- Linear and Free Regression (Curve fitting).
- T-tests for statistical significance ( $p < 0.05$ ).

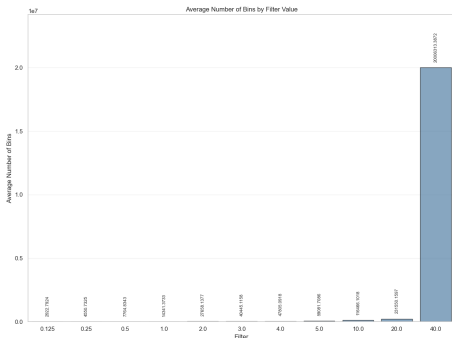
- **Total Parameters Processed:** 652,802,782,352 ( $\approx$  652 Billion).
- **Compute Time:** 228 hours ( $\approx$  9.5 days).
- **Dataset Size:** 5511 valid data points.

*Note: Some models excluded due to exceeding 1 billion histogram bins.*

# Filter Dimension Analysis

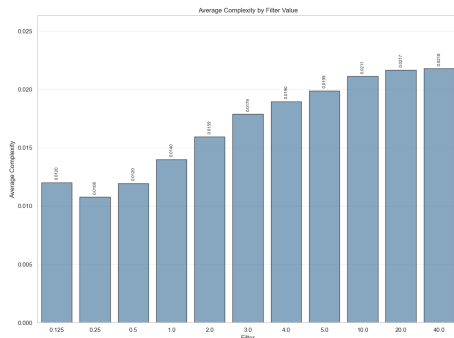
## Histogram Bins:

- Follows **Exponential Decay**.
- Max bins explode without filtering.



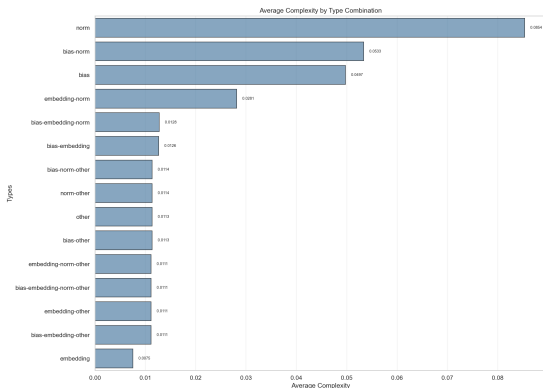
## Complexity:

- Follows **Logarithmic Trend**.
- Spike at 0.125  $\sigma$  (Global Min at 0.25).



**Decision:** 20  $\sigma$  chosen (Significant bin reduction, minimal complexity loss).

# Weight-Type Analysis



## Ranking:

- 1 **Norm:** Highest complexity.
- 2 **Bias/Other:** Medium.
- 3 **Embedding:** Near zero.

**Decision:** Use **Bias + Norm + Other** (No Embeddings).

- Embeddings dilute complexity.
- Aligns with Kaplan et al. (2020) methodology.

# Complexity vs. Parameter Count

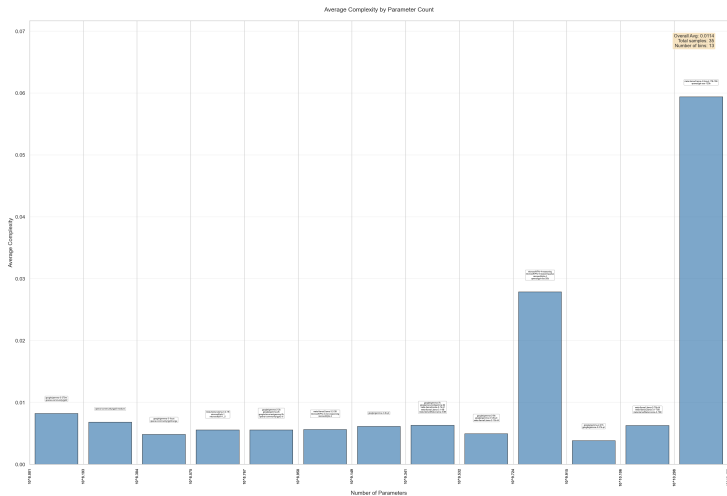
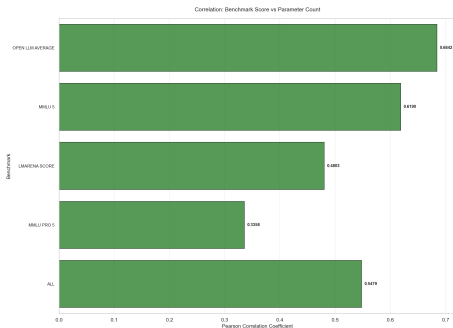


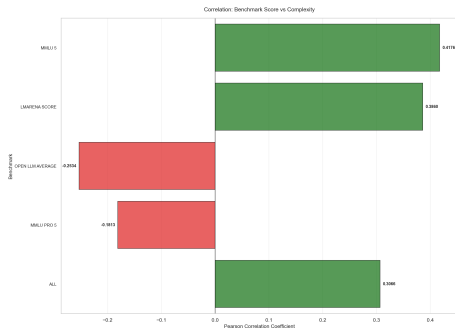
Figure: Average complexity vs number of parameters.

# Control: Parameters vs. Benchmarks



- **Validation:** All benchmarks show positive correlation with parameter count.
- **Expected:** Confirms scaling laws.
- **Baseline:**  $R^2$  values indicate non-linear relationship.

# Complexity vs. Benchmarks: Overview



## Inconsistency:

- **Positive:** MMLU, LMArena, All.
- **Negative:** MMLU-Pro, OpenLLM.

## Comparison:

- Lower correlations than Control.
- Lower  $R^2$  values.



# Statistical Significance (t-test)

Benchmark	$r$	$n$	p-value	Sig. ( $< 0.05$ )
LMarena	0.3860	21	0.0839	No
<b>MMLU</b>	<b>0.4176</b>	<b>26</b>	<b>0.0338</b>	<b>Yes</b>
MMLU-Pro	-0.1813	15	0.5179	No
OpenLLM	-0.2534	24	0.2322	No
<b>All (Aggregated)</b>	<b>0.3066</b>	<b>86</b>	<b>0.0041</b>	<b>Yes</b>

- **MMLU** and **Aggregated (All)** show statistically significant positive correlations.
- Negative correlations (MMLU-Pro, OpenLLM) are **not** significant.

# Regression Analysis: MMLU

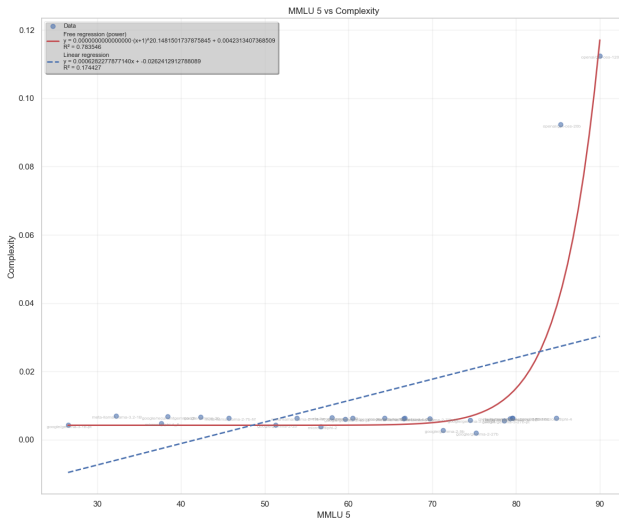


Figure: LMC complexity vs MMLU benchmark.

# Regression Analysis: OpenLLM

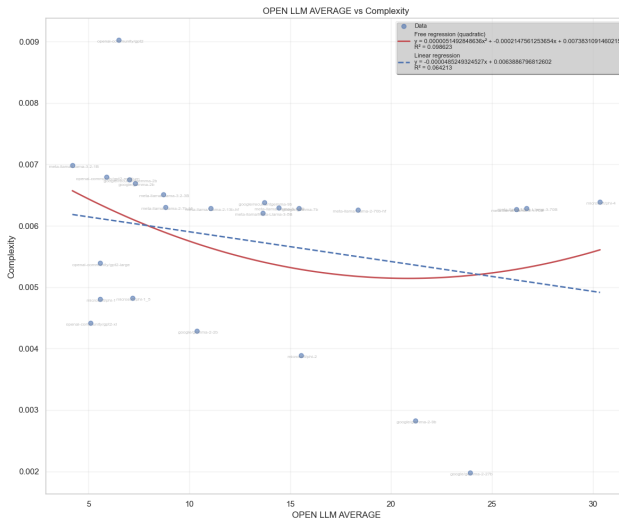


Figure: LMC complexity vs OpenLLM benchmark.

# Top 20 Correlations

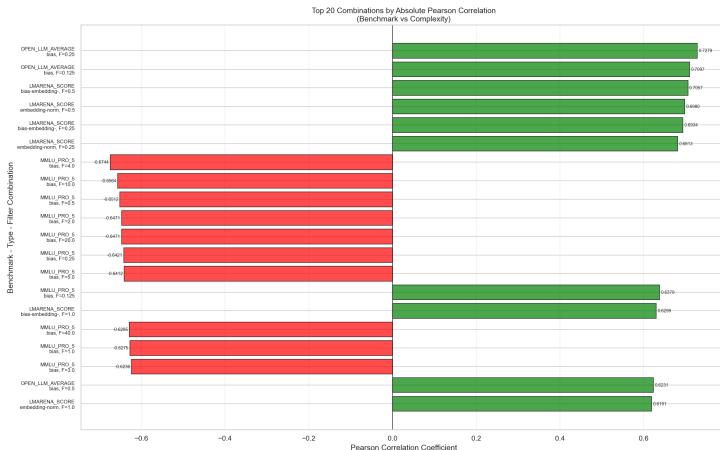


Figure: Top 20 configurations by Pearson correlation.

- Dominated by **High Filtering** ( $0.25 \sigma$ ).
- **Bias** weights appear in almost all top configurations.

## Main Finding

**A general correlation between LMC Complexity and Inference Capability cannot be confirmed.**

## Evidence For Hypothesis:

- Statistically significant positive correlation in aggregated data ( $r \approx 0.31$ ).
- MMLU shows significant positive correlation ( $p < 0.05$ ).
- Positive correlations are stronger/more significant than negative ones.

## Evidence Against Hypothesis:

- Positive trends are heavily driven by outliers (GPT-OSS).
- Inconsistent trends across benchmarks (some negative).
- Low  $R^2$  values compared to parameter count control.

## ① **Controlled Training:**

- Train models from scratch.
- Compare Test Loss vs Complexity directly (removes benchmark noise).

## ② **Self-Comparison:**

- Track complexity evolution of a *single* model during training.

## ③ **Optimization:**

- Can maximizing LMC complexity (e.g., of Norm layers) during training improve efficiency or performance?

# Thank You!

Questions?