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
Department of Computing and Mathematics

2025

Outline

Introduction

Context: The Era of Large Language Models

- **Transformers (2017)**:
  - Introduced by Vaswani et al. [12].
  - Enabled massive parallelization, sparking a "gold rush" in ML.
- **Rapid Adoption**:
  - GPT-3.5 (ChatGPT) became the fastest-growing consumer app in history (2022) [11].
  - Triggered massive investment from Tech Giants (Google, Microsoft, Meta).

The Scaling Paradigm

- **Scaling Laws by Kaplan et al., 2020** [7]:
  - Performance depends strongly on scale:
    - N: Number of Parameters.
    - D: Dataset Size.
    - C: Amount of Compute.
  - Performance depends weakly on shape (depth vs width).
- **Power Laws**:
  - $L(N) \approx (N_c/N)^{\alpha}$
  - **Implication**: Exponential increase in resources is required for constant linear gains in performance.

Problem Statement

- **Diminishing Returns**:
  - Recent models show marginal gains despite massive cost increases.
  - "Data Wall": Running out of high-quality internet data.
- **The Challenge**:
  - Relying solely on scaling (N, D, C) is becoming unsustainable.
  - Need for alternative approaches to improve efficiency.

Motivation & Proposed Solution

- **Two Approaches**:
  1. **Brute Force**: Continue scaling (Current Industry Standard).
  2. **Understanding**: Analyze the learning process to engineer better architectures.
- **Our Focus**:
  - Investigate **LMC Statistical Complexity** [8].
  - A metric combining **Disequilibrium** (Order) and **Entropy** (Randomness).
  - Hypothesis: It might help creating a better model by relating inference performance and its distribution.

Thesis and Objectives

Work Thesis

"There exists a relationship between model complexity and its inference capability." (Murta Junior, 2025)

Main Objective:

- Validate the existence of a meaningful relationship between neural network weight complexity and inference performance.

Secondary Objectives:

- Explore the influence of **Parameter Count**.
- Analyze the impact of **Weight Types** (Bias, Norm, Embedding).
- Determine the effect of **Filtering** outliers.

Methodology

Experimental Setup

- **Hardware Constraints**:
  - **RAM**: 512GB DDR4 (Critical for loading large models).
  - **GPU**: NVIDIA Quadro P5000 (16GB).
  - **CPU**: 2x Intel Xeon Gold 6130 (64 threads).
- **Software Stack**:
  - Python 3.12, PyTorch 2.8, Transformers 4.56.
  - Models loaded in **Main Memory (CPU)** cast to `float32`.

Model Selection Strategy

- **Source**: Hugging Face (Open Weights) [14, 2].
- **Criteria**:
  - Must be available on Hugging Face (official company account).
  - Transformer-based language model.
  - Open weights (including gated access).
  - Text-only (no multimodal inputs).
  - Base model (no fine-tunes).
  - Parameter count $<$ 150 Billion (Hardware limit).
  - Supported by `AutoModel` utility.
  - Has benchmark results available.
- **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).

Inference Capability: Benchmarks

- **Why Benchmarks?**:
  - Proxies for Test Loss (Performance) [10].
  - Training from scratch/Test loss unavailable for all models.
- **Selection Criteria**:
  - **Relevance**: Widely recognized (e.g., MMLU).
  - **Generality**: Covers range of tasks.
  - **Availability**: Results publicly available.
- **Selected Benchmarks**:
  - **MMLU**: 57 tasks, STEM/Humanities. Standard for LLMs [6].
  - **MMLU-Pro**: Enhanced MMLU, harder reasoning [13].
  - **OpenLLM**: Aggregated score of multiple datasets [9].
  - **LMArena**: Crowdsourced Elo ratings based on human preference [1].

Benchmark Availability

Models	BENCH-OPEN_LLM_AVERAGE	BENCH-LMARENA_SCORE	BENCH-MMLU_5	BENCH-MMLU_PRO_5
meta-llama/Llama-4-Scout-17B-16E		1319.0	79.6	58.2
meta-llama/Llama-3.2-3B	8.7	1165.0	58.0	22.2
meta-llama/Llama-3.2-1B	4.2	1111.0	32.2	11.9
meta-llama/Llama-3.1-70B	26.2	1291.0	79.3	53.8
meta-llama/Llama-3.1-8B	14.42	1210.0	66.7	37.1
meta-llama/Meta-Llama-3-70B	26.71	1273.0	79.5	55.0
meta-llama/Meta-Llama-3-8B	13.63	1221.0	66.6	36.2
meta-llama/Llama-2-70b-hf	18.37	1169.0	69.7	
meta-llama/Llama-2-13b-hf	11.07	1140.0	53.8	
meta-llama/Llama-2-7b-hf	8.81	1107.0	45.7	
google/gemma-3-27b-pt		1363.0	78.6	52.2
google/gemma-3-12b-pt		1339.0	74.5	45.3
google/gemma-3-4b-pt		1301.0	59.6	29.2
google/gemma-3-1b-pt			26.5	
google/gemma-2-27b	23.93	1284.0	75.2	56.5
google/gemma-2-9b	21.21	1261.0	71.3	52.1
google/gemma-2-2b	10.36	1195.0	51.3	
google/gemma-7b	15.44	1131.0	64.3	
google/gemma-2b	7.32	1087.0	42.3	
google/recurrentgemma-9b	13.71		60.5	
google/recurrentgemma-2b	7.02		38.4	
microsoft/Phi-4-reasoning				74.3
microsoft/Phi-4-reasoning-plus				76.0
microsoft/phi-4	30.36	1253.0	84.8	71.5
microsoft/phi-2	15.53		56.7	
microsoft/phi-1_5	7.17		37.6	
microsoft/phi-1	5.57			
openai/gpt-oss-120b		1350.0	90.0	
openai/gpt-oss-20b		1325.0	85.3	
openai-community/gpt2-xl	5.09			
openai-community/gpt2-large	5.57			
openai-community/gpt2-medium	5.9			
openai-community/gpt2	6.51			

Benchmarks

LMC Statistical Complexity

Definition

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

- **Disequilibrium (D)**:
  - Measures distance from uniform distribution ("Order").
  - $D = \sum_{i=1}^{n}(p_i - \frac{1}{n})^2$
- **Shannon Entropy (H)**:
  - Measures uncertainty or randomness.
  - $H = -K \sum_{i=1}^{n} p_i \log p_i$

LMC Statistical Complexity Extraction

1. **Weight Extraction**:
   - Flatten tensors from `named_parameters()`.
   - Categorize: Bias, Norm, Embedding, Other.
   - Tested combinations: Power set of categories (15 combinations).
2. **Filtering**:
   - Remove outliers.
   - Range: $\mu \pm \sigma_{\text{filter}} \cdot \sigma$.
   - Tested $\sigma_{\text{filter}} \in \{0.125, \ldots, 20, 40(\text{unfiltered})\}$.
3. **Discretization (Histogram)**:
   - **Freedman-Diaconis Rule** [4]: $h = \frac{2 \times IQR}{N^{1/3}}$.
   - Adapts to distribution spread and sample size (N).
4. **Calculate LMC**:
   - Compute $C_{LMC} = H \times D$ using the histogram probabilities.

Analysis Dimensions: Dataset

- **Dataset Construction**:
  - **Models**: 35 Selected Models.
  - **Weight Combinations**: 15 (Power set of Bias, Norm, Embedding, Other).
  - **Filtering Settings**: 11 (σ_{filter} values).
  - **Total Data Points**: $35 \times 15 \times 11 \approx 5775$.
- **Tuple Structure**:
  - (Model, Params, Weight-Type, Filter, Complexity, Bins, Benchmarks).

Analysis Dimensions: Statistical Tools

- **Correlation Analysis**:
  - **Pearson Correlation (r)**: Measure linear relationship.
  - **T-tests**: Determine statistical significance ($p < 0.05$).
- **Regression Analysis**:
  - **Linear Regression**: $y = ax + b$.
  - **Free Regression**: Curve fitting (Linear, Quadratic, Exponential, Logarithmic, Power).
  - R^2 **Score**: Measure goodness of fit.

Results

Data Extraction Statistics

- **Scale**:
  - Total Parameters Processed: **652.8 Billion**.
  - Compute Time: **228 hours** ($\approx$ 9.5 days).
- **Dataset**:
  - Expected: 5775 data points.
  - Actual: **5511 data points**.
  - **Exclusions**: Models exceeding 1 billion bins (unfiltered) or containing NaN/Infinite values (numerical errors).

Filter Dimension: Histogram Bins (Average)

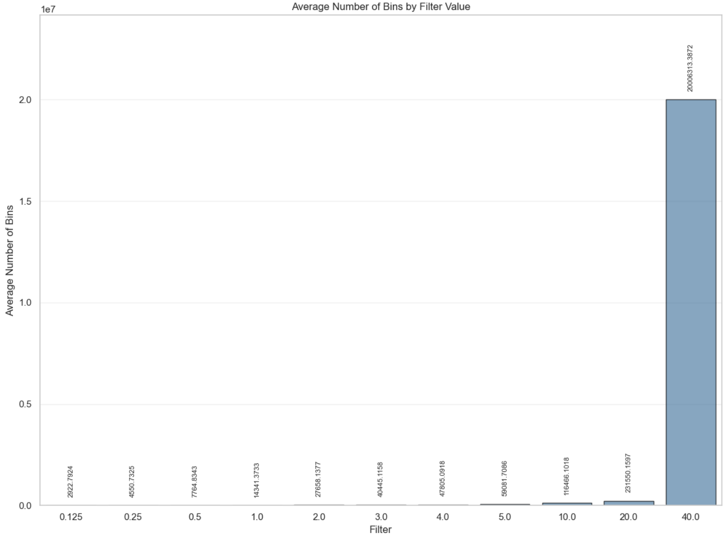

Figure: Average number of histogram bins per filtering setting.

Filter Dimension: Histogram Bins (Maximum)

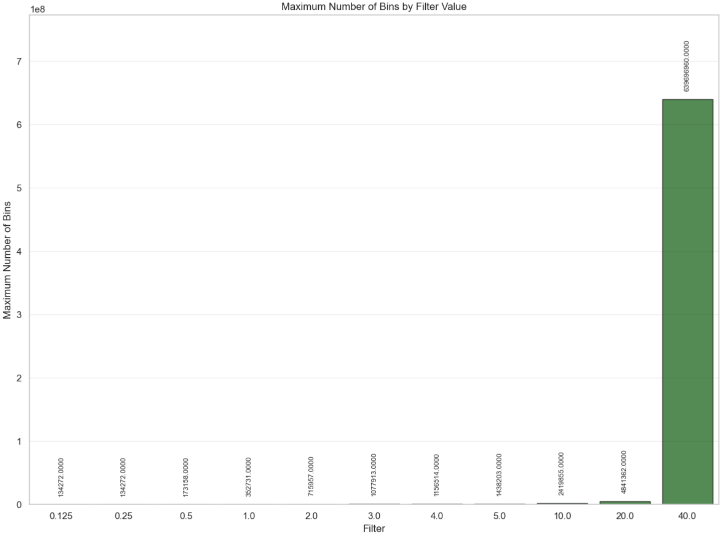

Figure: Maximum number of histogram bins per filtering setting.

Filter Dimension: Histogram Bins Regression

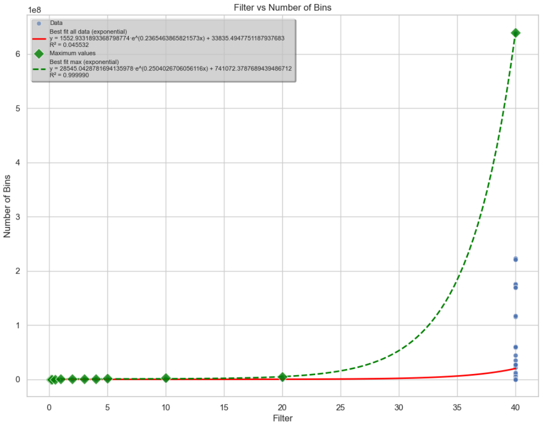

Figure: Regression of histogram bins per filtering setting.

- Maximum bins follow strict exponential trend ($R^2 = 0.999$).
- Average bins show higher variability ($R^2 = 0.045$).

Filter Dimension: Complexity (Average)

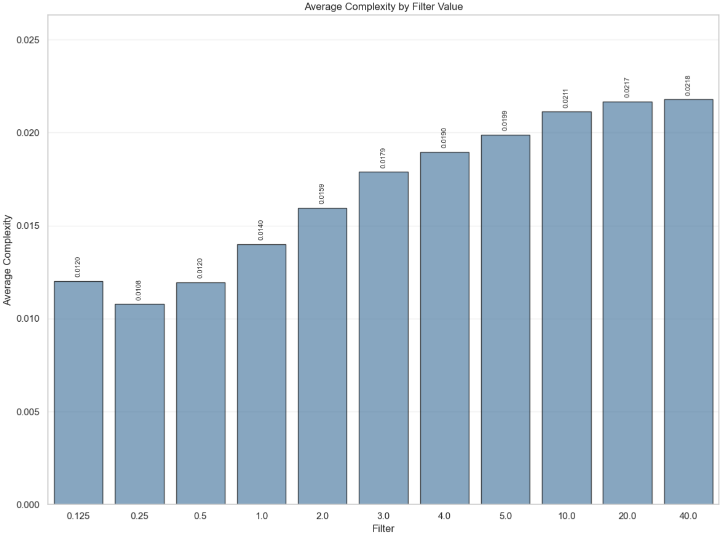

Figure: Average complexity per filtering setting.

Filter Dimension: Complexity (Maximum)

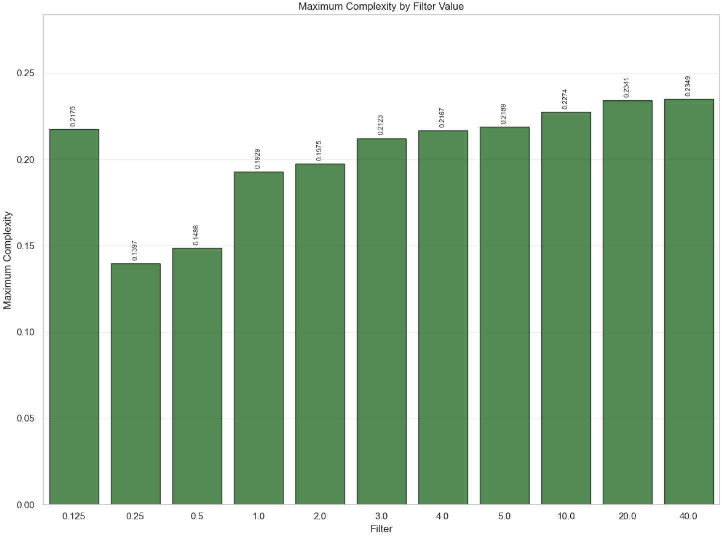

Figure: Maximum complexity per filtering setting.

Filter Dimension: Complexity Regression

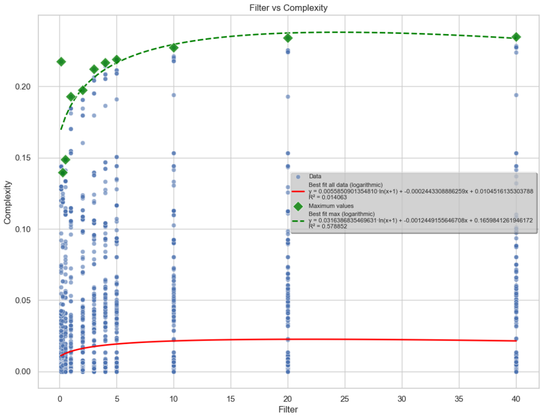

Figure: Regression of complexity per filtering setting.

- Both follow logarithmic trend.
- Maximum fit ($R^2 = 0.578$) is better than average fit ($R^2 = 0.014$).

Filter Selection

Decision: **20** σ chosen.

- Significant bin reduction.
- Complexity values almost identical to unfiltered data.

Weight-Type Dimension: Average Complexity

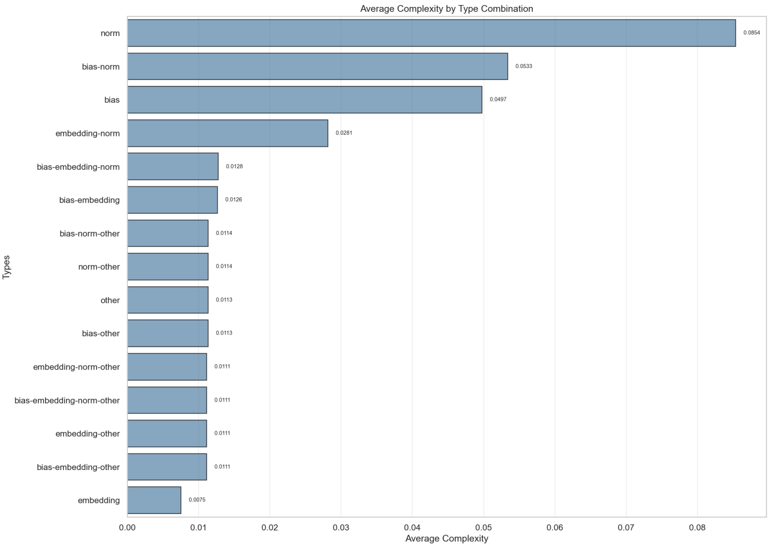

Figure: Average complexity per weight-type combination.

Weight-Type Dimension: Maximum Complexity

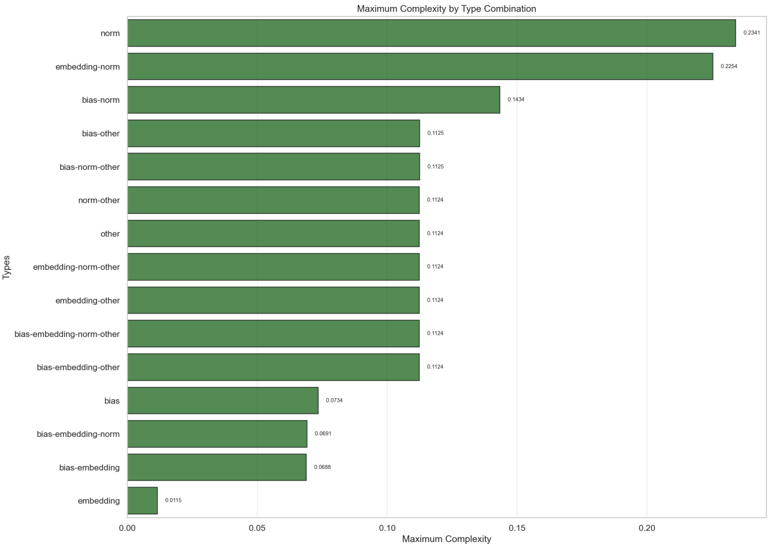

Figure: Maximum complexity per weight-type combination.

Weight-Type Selection

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

- Embeddings have very low complexity (near zero).

Complexity vs. Number of Parameters

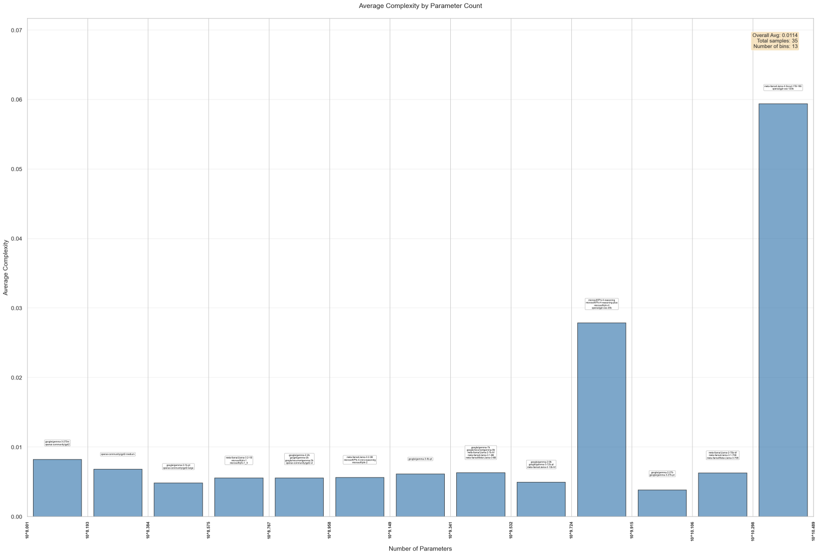

Figure: Average complexity vs number of parameters.

Complexity vs. Number of Bins

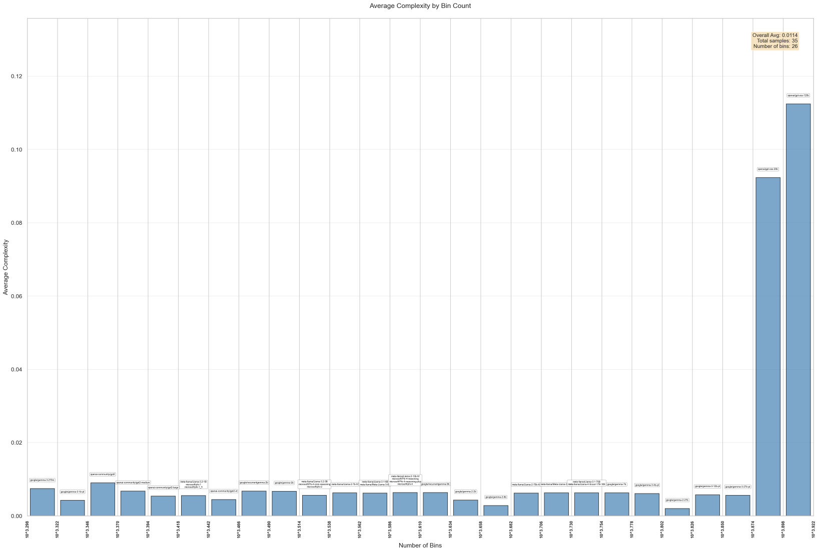

Figure: Average complexity vs number of histogram bins.

Control: Parameters vs. Benchmarks (Correlation)

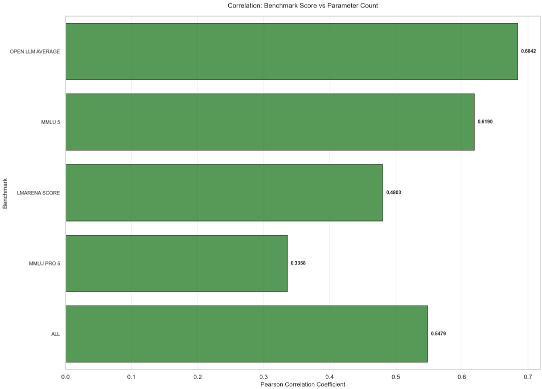

Figure: Pearson correlation for parameter count vs benchmark performance [3].

- All benchmarks show **positive correlation**.
- Validates methodology and scaling laws.

Control: Parameters vs. Benchmarks (R^2)

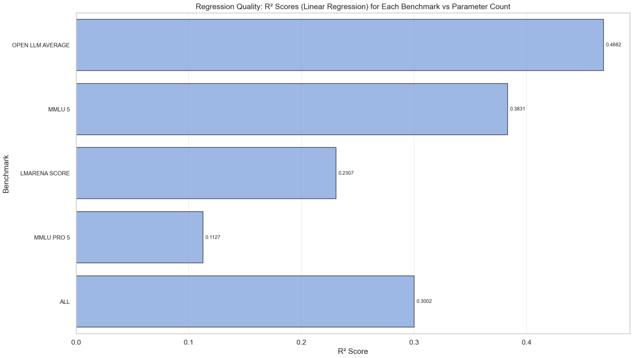

Figure: R^2 values for parameter count vs benchmark performance.

- Low R^2 values indicate non-linear relationship.
- Consistent with Power Laws.

Complexity vs. Benchmarks: Correlation

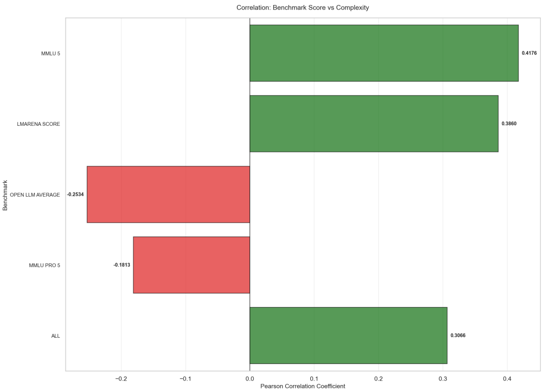

Figure: Pearson correlation for Complexity vs benchmark performance.

- **Inconsistent**: Positive (MMLU, LMArena) vs Negative (MMLU-Pro, OpenLLM).
- Lower correlations than Control.

Complexity vs. Benchmarks: R^2

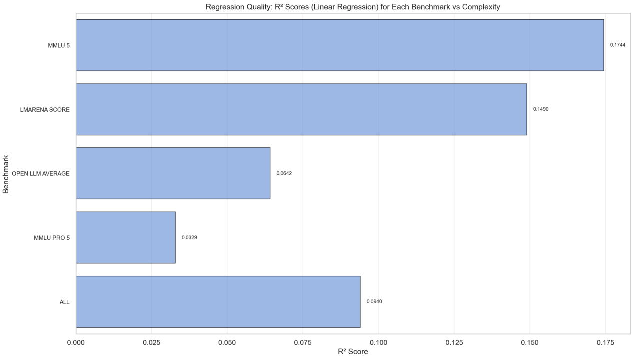

Figure: R^2 values for complexity vs benchmark performance.

- Very low R^2 values.
- Indicates weak predictive power.

Statistical Significance

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

- **MMLU** and **Aggregated (All)** are significant (T-test [5]).
- Negative correlations are not significant.

Regression Analysis: LMArena

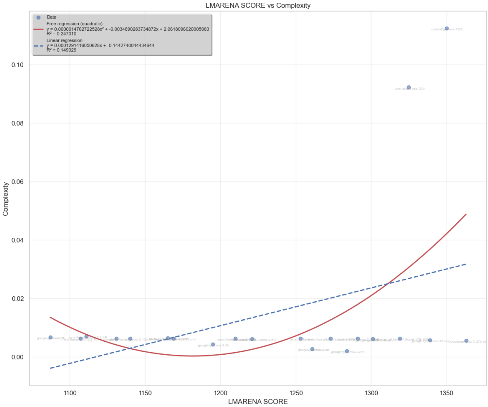

Figure: LMC complexity vs LMArena benchmark.

- Constant trend with outliers.
- Outliers (GPT-OSS) drive positive correlation.

Regression Analysis: MMLU

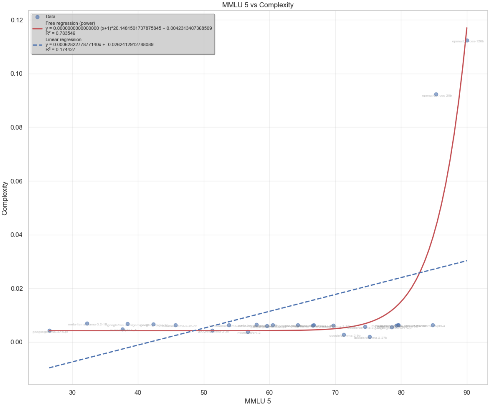

Figure: LMC complexity vs MMLU benchmark.

- Similar to LMArena.
- Exponential fit suggested, but driven by outliers.

Regression Analysis: MMLU-Pro

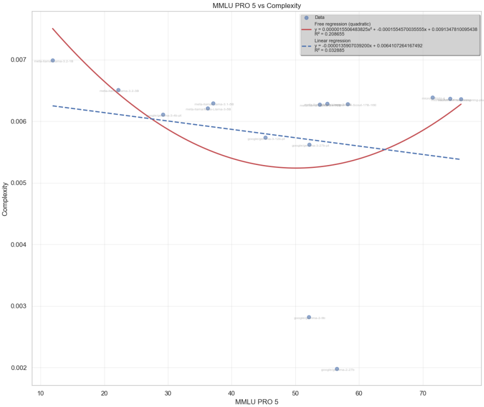

Figure: LMC complexity vs MMLU-Pro benchmark.

- Slight downward trend.
- Outliers (Gemma-2) drive negative correlation.

Regression Analysis: OpenLLM

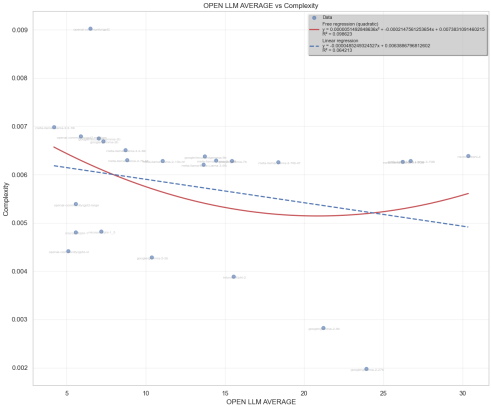

Figure: LMC complexity vs OpenLLM benchmark.

- **Dual Trend**: Upward (LLaMA, Phi-4) vs Downward (Gemma-2, Phi-1.5).

Regression Analysis: All Benchmarks

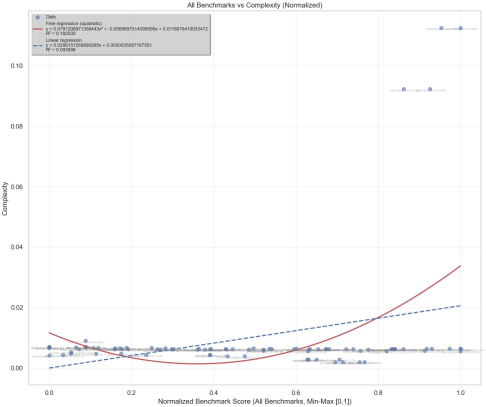

Figure: LMC complexity vs all benchmarks aggregated.

- Follows the "Constant + Outlier" pattern.
- Statistically significant positive correlation.

Top 20 Correlations

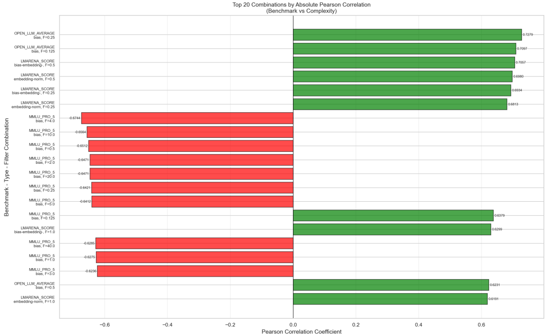

Figure: Top 20 configurations by Pearson correlation.

- Dominated by high filtering (0.25 σ).
- **Bias** weights are prevalent.

Conclusion

Main Conclusion

Finding

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

- While statistical significance was found in aggregated data, the relationship is inconsistent across individual benchmarks.
- The results suggest that LMC Complexity, in its current form, is not a reliable universal predictor of model performance.

Evidence Supporting the Hypothesis

- **Aggregated Significance**:
  - The aggregated dataset (**All**) showed a statistically significant positive correlation ($p < 0.05$).
- **Positive Bias**:
  - Positive correlations (e.g., MMLU) were stronger and more significant than negative ones.
- **Parameter Relation**:
  - Complexity tends to increase with parameter count, which is a known predictor of performance.

Evidence Against the Hypothesis

- **Outlier Dependence**:
  - Positive trends were heavily driven by specific outliers (e.g., GPT-OSS family).
  - Removing outliers often reduced correlations to near zero.
- **Inconsistency**:
  - Different benchmarks yielded contradictory results (Positive vs. Negative correlations).
  - Regression shapes varied widely (Constant, Linear, Dual-trend).
- **Predictive Power**:
  - Low R^2 values compared to the control (Parameter Count).

Future Work

1. **Controlled Training & Optimization**:
   - Train models from scratch to compare Test Loss vs. Complexity directly (Intra-Model Analysis).
   - Eliminates the noise and inconsistency of public benchmarks.
   - Investigate if maximizing LMC complexity improves performance.
2. **Filtering Refinement**:
   - Explore less aggressive filtering (e.g., 30 σ) and the complexity spike at 0.125 σ.
3. **Outlier Investigation**:
   - Investigate why the GPT-OSS family is an outlier for LMC complexity.
4. **Alternative Hypothesis**:
   - Complexity might measure "distance to performance ceiling" (Early Stopping Criterion).

Thank You!

Questions?

References I

[1] Wei-Lin Chiang et al. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference". In: *arXiv preprint arXiv:2403.04132* (2024). URL: https://arxiv.org/abs/2403.04132.

[2] Hugging Face. *Hugging Face – The AI community building the future*. https://huggingface.co/. 2024.

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

[4] David Freedman and Persi Diaconis. "On the Histogram as a Density Estimator: L_2 Theory". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57 (1981), pp. 453–476. URL: https://link.springer.com/content/pdf/10.1007/BF01025868.pdf.

References II

[5] David Freedman, Robert Pisani, and Roger Purves. *Statistics*. 4th. W. W. Norton & Company, 2007.

[6] Dan Hendrycks et al. "Measuring Massive Multitask Language Understanding". In: *arXiv preprint arXiv:2009.03300* (2020). URL: https://arxiv.org/abs/2009.03300.

[7] Jared Kaplan et al. "Scaling Laws for Neural Language Models". In: *arXiv preprint arXiv:2001.08361* (2020). URL: https://arxiv.org/abs/2001.08361.

[8] Ricardo Lopez-Ruiz, Hector Mancini, and Xavier Calbet. "A Statistical Measure of Complexity". In: *Physics Letters A* 209.5-6 (1995), pp. 321–326.

[9] Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. "Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena". In: *arXiv preprint arXiv:2406.07545* (2024). URL: https://arxiv.org/abs/2406.07545.

[10] David Owen. "How predictable is language model benchmark performance?" In: *arXiv preprint arXiv:2401.04757* (2024). URL: https://arxiv.org/abs/2401.04757.

[11] T. Trust, J. Whalen, and C. Mouza. "Editorial: ChatGPT: Challenges, Opportunities, and Implications for Teacher Education". In: *Contemporary Issues in Technology and Teacher Education* 23.1 (2023), pp. 1–23. URL: https://www.learntechlib.org/primary/p/222408/.

References IV

[12] Ashish Vaswani et al. "Attention Is All You Need". In: *arXiv preprint arXiv:1706.03762* (2017). URL: https://arxiv.org/abs/1706.03762.

[13] Yubo Wang et al. "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark". In: *arXiv preprint arXiv:2406.01574* (2024). URL: https://arxiv.org/abs/2406.01574.

[14] Thomas Wolf et al. "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *arXiv preprint arXiv:1910.03771* (2019). URL: https://arxiv.org/abs/1910.03771.