Vectors from Larger Language Models Predict Human Sentence Processing More Poorly when Dimensionality is Controlled

Yi-Chien Lin (lin.4434@osu.edu) and William Schuler (The Ohio State University)

There have been some studies conjecturing a 'quality-power' relationship in which LLMs' ability to predict human psychometric data improves as word accuracy increases [18]. But results are mixed: on the one hand, studies using (log) word probabilities as predictors have shown scaling inverts at a point, after which fit declines as models grow larger [11, 12]; on the other hand, studies using entire vectors from differently sized LLMs seem to continue to show positive scaling, but do not strictly control for the larger number of predictors in vectors from larger LMs [15]. A larger number of predictors gives a regression more degrees of freedom, or more opportunities to find strong predictors. This is a basis of reservoir computing [7, 10] in which a single neural layer is trained as a classifier using a larger untrained neural network as input. This study attempts to reconcile this disparity by evaluating the scaling of LLM vectors when vector length is controlled using untrained LLMs as a reservoir computer baseline.

One self-paced reading (Natural Stories SPR) [5], two eye-tracking (Dundee ET and Provo ET) [8, 9], and two functional magnetic resonance imaging (Natural Stories fMRI and Pereira fMRI) [16, 13] English datasets were analyzed in this study. We first replicated the positive scaling observed by Schrimpf et al. [15]. A set of linear regression models were fit using vector elements, collected from the final layers of LLMs from three LM families, GPT-2 [14], GPT-Neo [2, 1, 17], and OPT [19], and the response data in the fit partitions. Motivated by Feghhi et al. [4], we also included sentence position and sentence length (for Pereira fMRI) and word position (for other corpora) as baseline predictors. For Natural Stories fMRI, we applied a hemodynamic response function convolution [3] to all predictors so that the convolved predictors align with the response data. For Pereira fMRI, since its responses are per-sentence BOLD signals, we followed recent work on this dataset [6] by collecting the vector corresponding to the final word of each sentence. Following [15], we examined the predictive power of the vector elements by measuring the Pearson correlation between the regression model predictions and the human response in held-out partitions. This experiment replicated the significant positive scaling for the majority of datasets.[1]

We then examine how much the full training process of LMs contributes beyond the effect of effect of reservoir computing, we subtracted the correlation scores of an untrained Pythia variant (Figure 1A) from that of its trained counterpart (Figure 1B). Figure 2 shows that instead of increasing, the correlation score differences of the majority of the corpora decrease significantly as the model size increases. Inverse scaling on Natural Stories SPR and Pereira fMRI was not significant but the relationship between parameter count and the contribution of fully trained models beyond the effect of reservoir computing still shows a numerically negative correlation. These results suggest that after (full) training, LLMs are not contributing more as they get larger beyond the effect of reservoir computing. Moreover, the contribution of LLM training to model fit over equivalent untrained models drops to zero on the Provo ET, Natural Stories fMRI and Pereira fMRI datasets, which constitute a majority of the datasets used in this study. This may also partially explain the lack of significance of the numerically negative slope observed on Pereira fMRI.

These results suggest that the better predictive power of vector elements from larger LMs is mainly due to the effect of simple reservoir computing, which is increased with the increasing number of predictors from larger LMs, and the word-prediction training of LMs does not contribute more as models gets larger. This observation suggests that there may be a substantial misalignment between LLMs and human sentence processing, which worsens when larger models are used.

[1] Due to the space limit, the figure for this experiment is not presented; however, the results of this experiment are generally consistent with the results shown in Figure 1B.

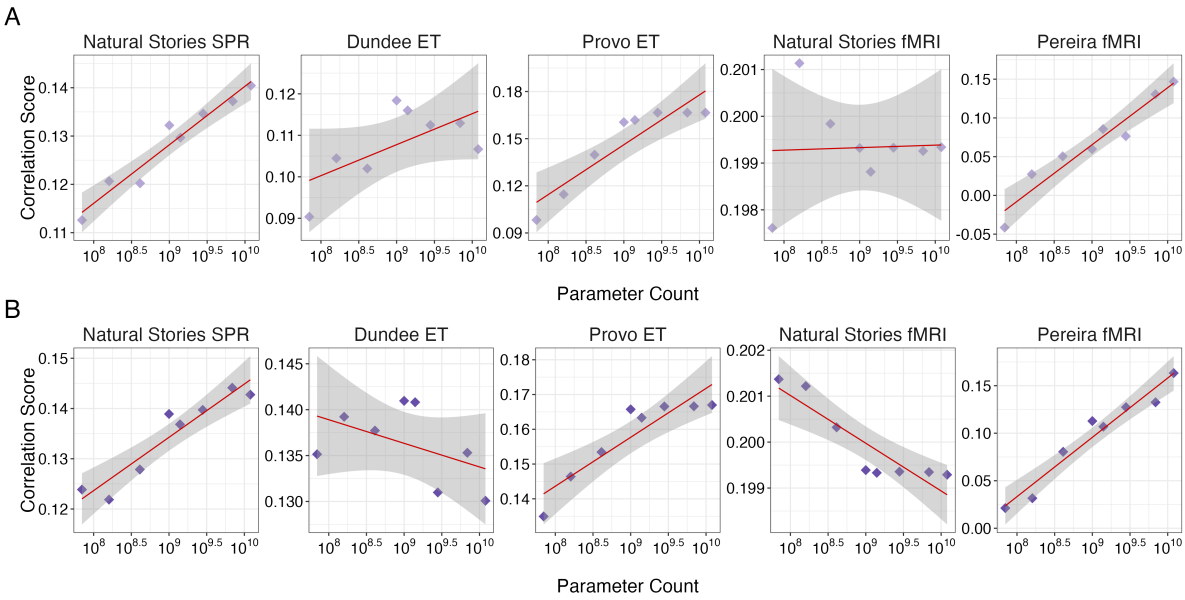

Figure 1: (A) Predictive power of vector elements from untrained Pythia models. All datasets other than Natural Stories fMRI show a significant positive scaling. (B) Predictive power of vector elements from fully trained Pythia models. Results of Natural Stories SPR, Provo ET, and Pereira fMRI show a significant positive relationship between predictive power and log-transformed parameter count.

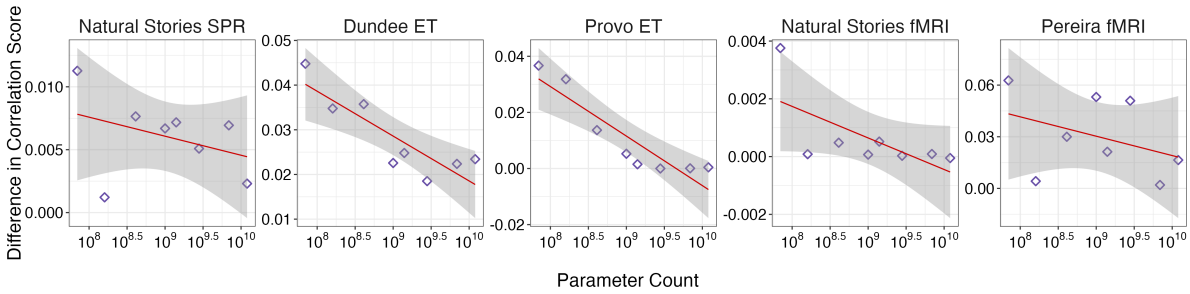

Figure 2: Contribution of full training to the predictive power of vector elements from fully trained Pythia models beyond the effect of reservoir computing.

[1] S. Black et al. GPT-NeoX-20B: An open-source autoregressive language model. In A. Fan et al., editors, *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, May 2022.

[2] S. Black et al. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. *Zenodo*, Mar. 2021.

[3] G. M. Boynton et al. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–4221, 1996.

[4] E. Feghhi et al. What are large language models mapping to in the brain? A case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*, 2024.

[5] R. Futrell et al. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77, 2021.

[6] E. A. Hosseini et al. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63, 2024.

[7] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks–with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13, 2001.

[8] A. Kennedy et al. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.

[9] S. G. Luke and K. Christianson. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833, 2018.

[10] W. Maass et al. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, Nov 2002.

[11] B.-D. Oh et al. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5, 2022.

[12] B.-D. Oh and W. Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350, 03 2023.

[13] F. Pereira et al. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, Mar. 2018.

[14] A. Radford et al. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

[15] M. Schrimpf et al. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[16] C. Shain et al. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307, 2020.

[17] B. Wang and A. Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, May 2021.

[18] E. Wilcox et al. Language model quality correlates with psychometric predictive power in multiple languages. In H. Bouamor et al., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore, Dec. 2023. Association for Computational Linguistics.

[19] S. Zhang et al. OPT: Open Pre-trained Transformer language models. *arXiv preprint*, arXiv:2205.01068v4, 2022.