Larger Language Models Better Predict Neural Activity During Natural Language Processing

Zhuoqiao Hong* **(hongzhuoqiao@gmail.com)**
Princeton Neuroscience Institute, Washington Road
Princeton, NJ 08544 USA

Haocheng Wang* **(kw1166@princeton.edu)**
Princeton Neuroscience Institute, Washington Road
Princeton, NJ 08544 USA

Zaid Zada (zaid@princeton.edu)
Department of Psychology, Princeton University, Washington Road
Princeton, NJ 08544 USA

Harshvardhan Gazula (hgazula@mit.edu)
McGovern Institute for Brain Research, MIT, 43 Vassar St
Cambridge, MA 02139 USA

Bobbi Aubrey (baubrey@princeton.edu)
Princeton Neuroscience Institute, Washington Road
Princeton, NJ 08544 USA

Werner Doyle (Werner.Doyle@nyulangone.org)
New York University Grossman School of Medicine, 550 1st Avenue
New York, NY, 10016 USA

Sasha Devore (Sasha.Devore@nyulangone.org)
New York University Grossman School of Medicine, 550 1st Avenue
New York, NY, 10016 USA

Patricia Dugan (Patricia.Dugan@nyulangone.org)
New York University Grossman School of Medicine, 550 1st Avenue
New York, NY, 10016 USA

Daniel Friedman (Daniel.Friedman@nyulangone.org)
New York University Grossman School of Medicine, 550 1st Avenue
New York, NY, 10016 USA

Orrin Devinsky (Orrin.Devinsky@nyulangone.org)
New York University Grossman School of Medicine, 550 1st Avenue
New York, NY, 10016 USA

Adeen Flinker (Adeen.Flinker@nyulangone.org)
New York University Grossman School of Medicine, 550 1st Avenue
New York, NY, 10016 USA

Uri Hasson** **(hasson@princeton.edu)**
Princeton Neuroscience Institute, Washington Road
Princeton, NJ 08544 USA

Samuel A. Nastase** **(snastase@princeton.edu)**
Princeton Neuroscience Institute, Washington Road
Princeton, NJ 08544 USA

Ariel Goldstein** **(ariel.goldstien@mail.huji.ac.il)**
Department of Cognitive and Brain Sciences, The Hebrew University Jerusalem
Jerusalem, 9190401 Israel

* Equal first author, alphabetical order
** Equal senior author

Abstract

Recent research has used large language models (LLMs) to study the neural basis of naturalistic language processing in the human brain. LLMs have grown in complexity, leading to improved language processing capabilities. Here, we utilized several families of transformer-based LLMs to investigate the relationship between model size and their ability to capture linguistic information in the brain. Crucially, a subset of LLMs were trained on a fixed training set, enabling us to dissociate model size from architecture and training set size. We used electrocorticography (ECoG) to measure neural activity in epilepsy patients while they listened to a 30-minute naturalistic audio story. We fit electrode-wise encoding models using contextual embeddings extracted from each hidden layer of the LLMs to predict word-level neural signals. In line with prior work, we found that larger LLMs better capture the structure of natural language and better predict neural activity. We also found a log-linear relationship where the encoding performance peaks in relatively earlier layers as model size increases.

Keywords: language comprehension; LLM; ECoG; speech

Introduction

Research on the neural basis of natural language processing has shifted from a modular perspective linking distinct neural regions with specific language features (Friederici, 2011; Saxe, Brett, & Kanwisher, 2006) toward a more unified approach driven by the emergence of large language models (LLMs). In this new paradigm, artificial neural networks serve as explicit models of neural computations and representations supporting high-level cognitive functions (Hasson, Nastase, & Goldstein, 2020; Richards et al., 2019). The internal representations of LLMs better predict human brain activity during natural language processing than any prior generation of models (Caucheteux & King, 2022; Goldstein et al., 2022; Kumar et al., 2022; Schrimpf et al., 2021).

LLMs rely on vast numbers of parameters and extensive diet training data, allowing them to encode diverse linguistic structures in high-dimensional embedding spaces (Linzen & Baroni, 2021; Manning, Clark, Hewitt, Khandelwal, & Levy, 2020; Piantadosi, 2023). Recent work suggests that model size—the number of learnable parameters—is critical: specific linguistic competencies emerge only in larger LLMs (Bommasani et al., 2021; Kaplan et al., 2020; Manning et al., 2020; C. Zhang, Bengio, Hardt, Recht, & Vinyals, 2021). This observation that simply scaling up LLMs yields more human-like language behavior led us to assess the relationship between the size of LLMs and their ability to predict human brain activity during natural language comprehension. In keeping with prior work using fMRI (Antonello, Vaidya, & Huth, 2024), we hypothesized that larger LLMs that more accurately capture linguistic structure would better capture neural activity.

We used electrocorticography (ECoG) to measure brain activity while participants listened to a naturalistic story stimulus. We calculated perplexity—the average level of surprise or uncertainty the model attributes to a word sequence—for multiple families of transformer-based LLMs. We extracted contextual embeddings from each hidden layer of all LLMs and fit electrode-wise encoding models to predict neural activity for each word in the stimulus. First, we replicated the finding that larger LLMs better predicted neural activity (Antonello et al., 2024). We then focused on the GPT-Neo family of models, which span a broad range of sizes and are trained on the same text corpora, to more thoroughly explore the relationships between model size and how well the embeddings predict neural activity.

Results

To investigate scaling effects between model size and alignment with brain activity, we utilized four families of transformer-based LLMs: GPT-2, GPT-Neo, OPT, and Llama 2 (Gao et al., 2020; Radford et al., 2019; Touvron et al., 2023; S. Zhang et al., 2023). These models span 82 million (M) to 70 billion (B) parameters and 12 to 80 layers. Each model family varies in architecture and training corpora. To control for these confounding variables, we also focused on the GPT-Neo family (Gao et al., 2020) with a comprehensive range of models that vary only in size, spanning from 125 M to 20 B parameters. For simplicity, we renamed the four models as “SMALL” (gpt-neo-125M), “MEDIUM” (gpt-neo-1.3B), “LARGE” (gpt-neo-2.7B), and “XL” (gpt-neox-20b).

We collected ECoG data from ten epilepsy patients listening to a 30-minute audio podcast (Chavis, 2017). We extracted the high-frequency broadband power for each electrode in 200 ms epochs at lags ranging from -2 s to +2 s relative to each word onset. Using the podcast transcript, we computed perplexity values for all LLMs. Consistent with prior research (Radford et al., 2019), we found that perplexity decreases as model size increases (Fig. 1A). Next, we extracted contextual embeddings from each hidden layer, utilizing the full context length of each LLM. Then, we used ridge regression to constructed electrode-wise encoding models that predict neural activity from the contextual embeddings for each word in the stimulus. We evaluated the encoding models using 10-fold cross-validation and calculated the Pearson correlation between predicted and actual neural signals. We repeated this analysis for each lag from -2 s to 2 s in 25 ms increments relative to word onset.

Larger LLMs better predict brain activity

To assess how well LLMs at different sizes predict neural activity, we obtained the maximum encoding performance (correlation) for each electrode across all lags and layers, then averaged these correlations across electrodes to derive the overall encoding performance for each model (Fig. 1B). We replicated previous fMRI work (Antonello et al., 2024) reporting a log-linear relationship between model size and encoding performance, indicating that larger models better predict neural activity. We also observed an earlier plateau in encoding

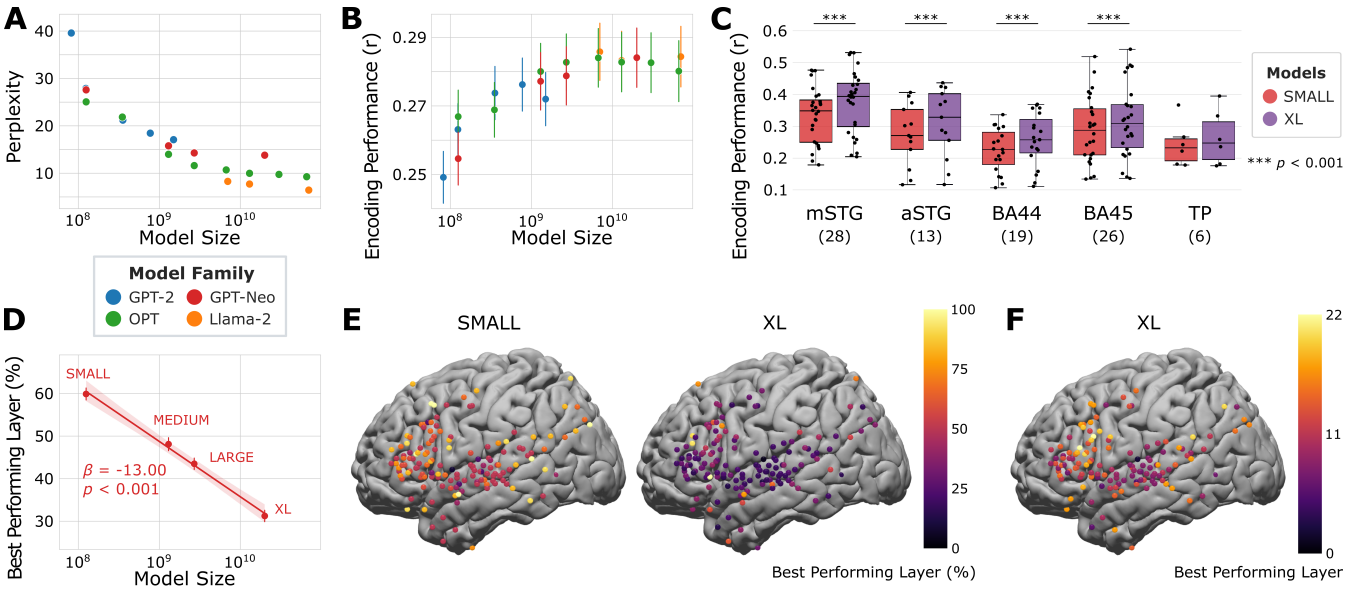

Figure 1: Encoding varies with model size. (A) Perplexity decreases with increase (log) model size. (B) Predictions of brain activity improve with increasing (log) model size. (C). The XL model better predicts brain activity than the SMALL model for most language regions. (D) Encoding performance peaks at relatively earlier layers (in percentage of overall depth) for larger/deeper models. (E) Best-performing layer (in percentage) for SMALL and XL. (F) Best-performing layer among the first half of layers in XL.

performance, occurring at smaller model sizes (13 B) compared to previous studies (30 B).

To dissociate between model size and other confounding variables, we focused on the GPT-Neo models and assessed encoding performance for each region of interest (ROI). We identified five ROIs across the language network: middle superior temporal gyrus (mSTG), anterior superior temporal gyrus (aSTG), Brodmann area 44 (BA44) and Brodmann area 45 (BA45) in inferior frontal cortex, and the temporal pole (TP). Consistent with prior studies (Goldstein et al., 2022), our encoding models achieved the highest correlations in mSTG and BA45. Furthermore, encoding performance for the XL model significantly surpassed SMALL in mSTG, aSTG, BA44, and BA45 (Fig. 1C).

Encoding peaks at earlier layers in larger LLMs

Next, we examined which layer in each model provides the best encoding performance. To that end, we identified the best layer for each electrode based on its maximum encoding performance. To account for the variation in depth across models, we computed the best layer as the percentage of each model's overall depth. We found a log-linear relationship such that as models increase in size, peak encoding performance tends to occur in relatively earlier layers, being closer to the input in larger models (Fig. 1D). This was consistent across multiple model families.

We further observed variations in how the best-performing layers mapped onto the cortical language processing hierarchy (Fig. 1E). In the SMALL model, peak encoding was observed for earlier layers in STG electrodes and for later layers in IFG; in the XL model, the majority of electrodes exhibited peak encoding in the first 25% of all layers (Fig. 1E). However, despite the XL model showing less variance in the best layer distributions across cortex, we found the same hierarchy present for the first half of the layers in the model (Fig. 1F). In this analysis, we observed that the best relative layer nominally increases from mSTG (M = 21.916, SD = 10.556) to aSTG (M = 29.720, SD = 17.979) to BA45 (M = 30.157, SD = 16.039) and TP (M = 31.061, SD = 16.305), and finally to BA44 (M = 36.962, SD = 13.140).

Discussion

In this study, we investigated how encoding models of neural activity scale with LLM model size. Corroborating prior work using fMRI (Antonello et al., 2024), we found that larger LLMs, ranging from 82 M to 70 B parameters, are better aligned with neural activity. Combined with the observation that larger LLMs produce lower perplexity, our findings suggest that larger LLMs' capacity for learning natural language structure yields better brain activity predictions. Our investigation into the best layers for encoding correlation revealed a log-linear relationship where peak encoding performance tends to occur in earlier layers as model size increases. Moreover, we observed variations in best relative layers across different brain regions, corresponding to a language processing hierarchy. These findings indicate that as LLMs increase in size, the later layers of the model may contain representations that are increasingly divergent from the brain during natural language comprehension.

Acknowledgments

This work was supported by the National Institutes of Health under award numbers DP1HD091948 (to A.G., M.H., H.W., Z.Z., B.A., A.F. and U.H.), R01NS109367 (to A.F.), and R01MH112566 (to S.A.N.), Finding a Cure for Epilepsy and Seizures (FACES), and Schmidt Futures Foundation DataX Fund.

References

Antonello, R., Vaidya, A., & Huth, A. (2024). Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, *36*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134.

Chavis, D. (2017). *This american life: So a monkey and a horse walk into a bar [audio podcast].*

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological Reviews*.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., ... others (2020). The pile: an 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... others (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*(3), 416–434.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., ... Nastase, S. A. (2022). Shared functional specialization in transformer-based language models and the human brain. *bioRxiv*, 2022–06.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Piantadosi, S. (2023). Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, *7180*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... others (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770.

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... others (2023). Opt: open pre-trained transformer language models, 2022. *URL https://arxiv.org/abs/2205.01068*, *3*, 19–0.