MODEL CONNECTOMES: A GENERATIONAL APPROACH TO DATA-EFFICIENT LANGUAGE MODELS

Klemen Kotar
Computer Science, Stanford
klemenk@stanford.edu

Greta Tuckute
Brain and Cognitive Sciences, MIT
gretatu@mit.edu

ABSTRACT

Biological neural networks are shaped both by evolution across generations and by individual learning within an organism's lifetime, whereas standard artificial neural networks undergo a single, large training procedure without inherited constraints. In this preliminary work, we propose a framework that incorporates this crucial generational dimension—an "outer loop" of evolution that shapes the "inner loop" of learning—so that artificial networks better mirror the effects of evolution and individual learning in biological organisms. Focusing on language, we train a model that inherits a "model connectome" from the outer evolution loop before exposing it to a developmental-scale corpus of 100M tokens. Compared with two closely matched control models, we show that the connectome model performs better or on par on natural language processing tasks as well as alignment to human behavior and brain data. These findings suggest that a model connectome serves as an efficient prior for learning in low-data regimes – narrowing the gap between single-generation artificial models and biologically evolved neural networks.

1 INTRODUCTION

How does the brain quickly and robustly learn to perform a wide array of tasks? A lot of research compares the representations from artificial and biological neural networks to broadly answer this question (1; 2; 3). However, artificial and biological networks differ fundamentally across several dimensions, including architecture, input-output constraints, and–critically–their learning processes (e.g., (4)). A key distinction lies in how learning takes place over time. In biological systems, learning consists of two nested "training loops": The evolutionary outer loop takes place across generations, where information needs to be transmitted in a low-bandwidth, compressed form–such as wiring rules and circuit priors (5). The inner learning loop takes place during the lifetime of a single individual, where relatively little data is used to flexibly learn complex behaviors and generalize to novel settings (6; 7). In contrast, standard artificial neural networks rely on a single large-scale training phase with a randomly initialized model using a generic architecture (Transformers; (8)), requiring vast amounts of data (9; 10). In this early stage work, we take a step toward bridging this gap between artificial and biological systems by proposing a generational learning framework for neural networks, followed by evaluation of task performance and alignment to human behavior and brain responses.

Our paper focuses on the domain of language, motivated by the massive engineering success of language-trained models (11; 12; 13; 14), as well as research that demonstrates alignment of language model representations with human behavior and neural responses during language processing (15; 16; 17; 18; 19).

Our problem formulation consists of two phases: an evolutionary outer loop, where the model has access to a large dataset of super-human scale, and a lifetime learning phase, where the model has to learn efficiently on a smaller, developmental-scale dataset (Figure 1A). Crucially, any learning acquired from the large dataset must be compressed into a connectome, a sparse binary mask, that is transmitted across "model generations". Specifically, across six generations of the evolutionary large-dataset loop–where 20% of the weights are pruned (i.e., zeroed out) in each generation–we

[1]Code available at: https://github.com/TuKoResearch/GenerationalConnectomes

derive a connectome that ultimately retains only 25% of the initial model weights. Additionally, the remaining weights are set to a fixed positive or negative constant after each generation, further reducing the information capacity of the connectome. This final connectome is then used to initialize a model, denoted as **Connectome**, which is trained on the smaller, developmental-scale dataset. To evaluate its performance, we compare it to two control models trained on the small dataset: i) **RandomConnectome**, a model with a randomly sampled connectome that also retains 25% of the initial weights, and ii) **NoConnectome**, an unpruned, fully-connected model with uniform initialization (the "standard choice" in machine learning) (Figure 1A).

Our core contribution is demonstrating that distilling information from a large dataset into a sparse connectome for initialization enables a model to generalize more effectively in low-data regimes. The resulting connectome model performs better or on par with control models in natural language processing (NLP) tasks and aligns more closely with human behavior and brain responses.

Related work in machine learning has explored compressing neural network weights by employing lower-bit representations (20), enforcing structural sparsity (21), and routing computations through uniform sub-modules as in mixtures of experts (22). Another line of research has shown that sparsely connected sub-networks within a larger network, when trained from their original initialization, can match the performance of a full image-trained network (23). Subsequent studies have refined this idea through iterative pruning over multiple generations (24), extended it to language models (25; 26), shown its benefits for transfer learning (27), proposed improved pruning policies (28), and demonstrated that sub-networks can be defined using binary masks alone (29). We consolidate these ideas under the notion of a **model connectome** — a sparse binary wiring diagram composed of excitatory and inhibitory connections (i.e., weights with a constant positive or negative sign) that is refined over successive generations. The connectome defines a model's initialization. While previous work has focused on achieving loss equivalence between pruned and unpruned models on the *same* training dataset (with the goal of reducing model size), our study investigates a pruned model's (i.e., initialized with a connectome) ability to learn effectively from a much smaller. Hence, in contrast to prior work, our approach highlights the potential of sparse initialization to support learning in data-limited regimes.

One line of related work within "NeuroAI" (3) has explored directly optimizing the compression of a model's weights to achieve high innate (zero-shot) performance for vision tasks (30; 31). In contrast, we do not optimize for compression–instead, our connectome evolves over generations through "standard" task optimization. In practice, our resulting models converge faster during training, rather than performing well at initialization. Another direction has pruned the embedding space of standard pre-trained models to identify model features that are most important for alignment with human behavior or neural data (32; 33). Unlike these approaches, we prune our model based solely on next-word prediction performance, assessing human alignment only after generational pruning. To our knowledge, the behavioral and brain alignment of generationally pruned models (here, denoted as "Connectome" models) has not previously been studied.

2 METHODS

2.1 MODEL DEVELOPMENT

2.1.1 MODEL ARCHITECTURE

We investigate how language models in the GPT-2 family (11) can transmit information through initialization across model generations. We utilize the standard GPT-2 configurations used in modern AI approaches (see Appendix A.1). In our main experiments, we train models with 124M weights (see Section 3.1), and, in an exploratory analysis we scale up to 417M weights (see Section 3.1.1). To ensure robustness of our results, all analyses are conducted with four seeds per model instantiation.

2.1.2 TRAINING DATASETS

Our framework leverages two datasets: A small ($\boldsymbol{S}$) dataset, approximating a child's language exposure up to age 10 (34; 9; 10) (100M tokens of FineWeb (35), $\sim$75M words), and a much larger ($\boldsymbol{L}$) dataset (4B tokens of FineWeb (35), $\sim$3B words) containing roughly the amount of tokens that are optimal for training a standard 124M-parameter GPT model (11) (see A.2 for details). We explore how the $\boldsymbol{L}$ dataset (orders of magnitude larger than $\boldsymbol{S}$) can transmit information through model initialization ("outer" evolutionary loop), denoted as the model connectome, across succes-

sive generations of large language models (LLMs), which are then ultimately trained on the smaller $\boldsymbol{S}$ dataset ("inner" learning loop).

2.1.3 GENERATIONAL OUTER LOOP: WIRING UP THE CONNECTOME

We define the model connectome as a special form of initialization: a binary mask of weights ("synapses") connecting LLM units between layers. The connectome has the following two properties: i) **sparsity**, where most synapses are zeroed out (effectively reducing the size of the final model), and ii) **binary initialization**, assigning the remaining synapses a binary initialization value, denoting whether it is excitatory (a constant positive) or inhibitory (a constant negative).

To derive the connectome, we explore the Iterative Pruning approach (23) on the $\boldsymbol{L}$ dataset. We begin with a fully dense normally initialized model with standard parameters ($\mu = 0, \sigma = 0.02$) (11), f_θ^0, and train it on $\boldsymbol{L}$ for 7,000 iterations. After training, we take the final state of this model, f_θ^{0fin}, and prune (i.e., zero out) 20% of the weights with lowest absolute magnitude, yielding the next model generation's initialization, f_θ^1 [1]. In addition to pruning, f_θ^1 retains only the sign of the unpruned weights in f_θ^{0fin}, initializing the weights of the subsequent model generation as 0.0, -0.02 or $+0.02$ (29) (where 0.02 is the standard deviation of the random initialization of the initial model parameters). We then train f_θ^1 for another $7,000$ iterations, only optimizing the remaining 80% unpruned binary initialized weights. This prune-and-retrain cycle is repeated for six generations, ultimately yielding f_θ^{5fin}, a model with 25% of the original weights retained from the original model, f_θ^0. The final f_θ^5 connectome serves as a highly compressed wiring diagram (Appendix A.3).

For each training generation, we use a batch size of 512 and a learning rate schedule consisting of 250 linear warm-up steps, 5,000 hold steps at 0.0018 followed by 1,750 steps of linear decay to zero as in (36). We use the AdamW optimizer and weight decay of 0.1.

2.1.4 DEVELOPMENTAL INNER LOOP: LEARNING LANGUAGE REPRESENTATIONS

We take the final connectome, f_θ^5, and use it to initialize a new sparse model, which we train on the smaller dataset $\boldsymbol{S}$ (just 100M tokens, different from the $\boldsymbol{L}$ dataset). This model is trained for 2,000 iterations (250 warm-up steps, 1,750 decay steps) using the same batch size, maximum learning rate, weight decay, and optimizer class as in the generational loop. We denote this model as **Connectome**. We train two control models, also on $\boldsymbol{S}$, to have a set of minimally differing control models: i) **RandomConnectome**, a model with a similar initialization mask to **Connectome**, except that the sparsity and the positive/negative weights are randomly sampled, and ii) **NoConnectome**, a normally initialized unpruned dense model with ($\mu = 0, \sigma = 0.02$).

2.2 MODEL EVALUATION

We evaluate models on NLP tasks (Section 3.1), behavioral alignment (Section 3.2), and neural alignment (Section 3.2), and the respective sections contain brief methods. For detailed methods, see Appendix A.4.

3 RESULTS

3.1 THE CONNECTOME MODEL OUTPERFORMS CONTROL MODELS ON NLP BENCHMARKS

We evaluate our models on the FineWeb validation loss—a popular LLM next-word prediction benchmark (35). As shown in Figure 1B (panel i), the **Connectome** model strongly outperforms the standard **NoConnectome** baseline when both are trained on the small dataset $\boldsymbol{S}$, despite **Connectome** using only 25% of the weights (purple vs. green line). To contextualize the **Connectome** model's performance, we compare it to an upper bound: a dense model (similarly 124M weights) trained on the full large dataset ($\boldsymbol{L}$) (dotted horizontal line in Figure 1B). Among the three models trained on the small dataset, **Connectome** comes closest to this upper bound. Finally, the **RandomConnectome** model performs substantially worse than both **Connectome** and **NoConnectome**, indicating that the iterative pruning procedure is carving out an efficient subspace for learning language. Next, we evaluate our models on two standard NLP benchmarks; HellaSwag (37), a task

[1] Note that we performed pruning in a layer-wise fashion, as pilot experiment ablations showed that global pruning led to significantly worse performance.

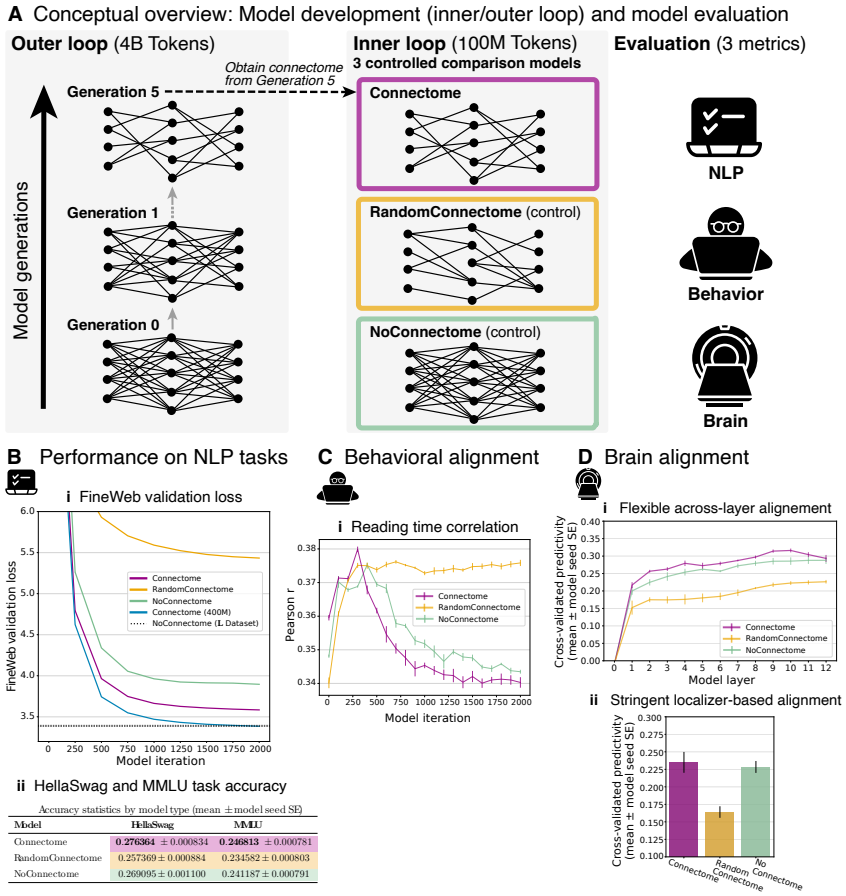

Figure 1: **A.** Conceptual overview, see description in Sections 1 and 2. **B.** Performance evaluation on standard NLP benchmarks: FineWeb validation loss (panel i), HellaSwag and MMLU (panel ii). **C.** Alignment with human reading times on naturalistic stories. **D.** Model-brain alignment, flexibly mapping all units within each model layer to brain responses (panel i) or through a more stringent procedure which localizes language-selective model units (panel ii). To ensure robustness of our results, all analyses are conducted with four seeds per model instantiation, and plots report the standard error of the mean (SE) across seeds.

that requires selecting the most plausible continuation for sentences concerning everyday scenarios, and MMLU (38), a multiple-choice benchmark assessing LLMs' knowledge and reasoning on various topics. In line with the superior performance of **Connectome** on the FineWeb validation loss (Figure 1B, panel i), **Connectome** also outperforms **RandomConnectome** and **NoConnectome** on the HellaSwag and MMLU (Figure 1B, panel ii), highlighting its strong performance as a language model beyond simply next-word prediction.

3.1.1 Scaling up: Exploratory analysis

So far (Figure 1B) we have compared three models (**Connectome**, **RandomConnectome**, and **NoConnectome**)–each with 124M weights (with **Connectome** and **RandomConnectome** having only 25% active weights, i.e., 31M), and found that the **Connectome** model consistently outperforms the two control models. But what happens if we take this a step further and scale up? In an exploratory analysis, we initialized a model with 417M weights and pruned it down to 109M active weights over a six-generation procedure similar to that of **Connectome**. As expected, the 417M pruned model (Figure 1B, blue line) outperforms the 124M **Connectome** model (purple line). Interestingly, the 417M pruned model obtains a loss comparable to the 124M dense model trained on the *large* dataset (***L***) (Figure 1B, dotted line), despite the 417M pruned model being trained only on the *small* dataset (***S***). These findings suggest that a large pruned model–only trained on the small dataset–matches the performance of a small dense model trained on the large dataset, underscoring the data efficiency of the connectome approach. More broadly, these results demonstrate the power of efficient model ini-

tialization through iterative pruning, and open the door for future work scaling to even larger models and evaluating on a larger array of tasks.

3.2 The Connectome model aligns better or on par with human reading times and brain responses

Behavioral alignment: Building on prior work demonstrating that per-word surprisal estimates from LLMs correlate with self-paced reading times (39; 40; 41) (a measure of language processing difficulty, (42; 43)), we tested our models against a large dataset of reading times from 179 participants reading 5-10 naturalistic stories (44). We tested the behavioral alignment between the models' estimated per-word surprisal and the reading time obtained from human participants, as done in prior work (45; 46). Based on Figure 1C, we note two main findings. First, we replicate the finding that for larger LLMs trained on large amounts of data, alignment with reading times degrades (47; 48; 49; 50; 41)–here, indicated by a peak behavioral alignment around iteration 250. In line with this finding, the **RandomConnectome** model, remaining in a high-loss regime, maintains a surprisingly high prediction of reading times. Second, we note that the **Connectome** model has the highest peak behavioral alignment with human reading times compared to the control models–although this result should be interpreted in light of the inverse correlation of training data amount and fit to human reading times.

Brain alignment: Building on prior work showing that internal LLM representations can predict human brain responses (17; 18; 19; 2), we here evaluated our models' brain alignment using a published benchmark consisting of responses from the human language network (51) collected from five participants reading 1,000 linguistically diverse sentences (52) (see additional details in Appendix A.4). First, we followed the alignment procedure of Tuckute et al. (52) (Figure 1D, panel i) fitting a linear encoding model that flexibly maps all units within each LLM layer to brain responses. The **Connectome** model consistently outperforms the **RandomConnectome** model while the **NoConnectome** model is more comparable. The **Connectome** model achieves $r = 0.32$ for layer 10 (the noise ceiling for this dataset is $r = 0.56$, i.e., 57%). Finally, we turn to a more stringent model-brain alignment procedure by focusing on a subset of LLM units that are intended to functionally correspond to the language network in the human brain. To identify these units, we adopt a well-established approach in neuroscience: selecting units that respond more strongly to well-formed sentences than to lists of non-words (51). Following AlKhamissi et al. (53) (see details and validation of the approach in this paper), we identify the top 1% most language-selective units across all layers of each model (yielding 92 units) and apply the same voxelwise encoding procedure used in our previous analyses. This localization step enables a more principled comparison to brain data, with the aim of targeting functionally similar units in models and brains. Although overall performance drops and there is variability across model seeds, the **Connectome** model is still more or at least as brain-aligned as the control models (Figure 1D, panel ii). Similar patterns hold when expanding to the top 10% language-selective units (Appendix A.5), with overall higher alignment. In conclusion, the **Connectome** model shows better or on par alignment with brain responses during language processing, highlighting the biological plausibility of our generational modeling framework.

4 Limitations & Discussion

In this paper, we show that a generational learning framework–transmitting sparse connections across model generations–provides an effective model initialization for learning in low-data regimes. Our **Connectome** model outperforms control models on NLP tasks and achieves better or on par alignment with human behavior and brain responses during language processing.

One way to interpret the success of model connectomes is through the lens of model distillation research (54; 55), which suggests that large models effectively perform a parallel search over many useful sub-networks during training. The lottery ticket hypothesis (23; 29) builds on this idea, demonstrating that these performant sub-networks can be efficiently extracted from the final model state. We further constrain the problem to a setting where only a binary connectome—capturing the presence and sign of connections—is inherited from an ancestor model trained on a large dataset, and show that this highly compressed structure serves as powerful initializations for efficient learning. This aligns with prior work that has demonstrated the relative importance of sign over magnitude of weight initializations for downstream task performance (29). Alternatively, the model connec-

tome can be thought of as a graph, where the outer loop training procedure identifies several good trajectories through which information flows across the model (56; 57). Because these pathways contribute strongly to the model's output, focusing optimization solely on them allows for efficient learning from limited data.

Limitations of this work exist. We by no means claim that our framework mimics the exact process of evolution: Biological organisms do not begin with a dense initialization that "prunes away" less important information over generations. Nevertheless, our framework demonstrates that a highly compressed connectome can transmit sufficient information across generations to evolve a strong initialization, without explicitly optimizing for compression. Another limitation is a limited set of alignment metrics investigated here; future work will explore additional behavioral and brain benchmarks for a more comprehensive view.

Future work includes investigating model behavior across checkpoints in both the outer and inner training loops, scaling up the models (for which we have demonstrated promising results, see Section 3.1.1), and establishing the relationship between the amount of data transmitted through the connectome and performance on various tasks. Additionally, machine learning interpretability techniques (e.g., (58; 59)) could be applied on the connectome to explain specific pruned circuits, and potentially relating them to hypothesized neural circuits.

ACKNOWLEDGMENTS

We thank Badr AlKhamissi for sharing the brain-score version of the Futrell2018 behavioral benchmark. We thank Dan Yamins for inspiring conversations. We also thank the Stanford HAI, Stanford Data Sciences and the Marlowe team for computing support. Greta Tuckute acknowledges funding support from MIT's McGovern Institute for Brain Research.

REFERENCES

[1] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, 2024.

[2] Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47, 2024.

[3] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

[4] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[5] Geoffrey E Hinton and Steven J Nowlan. How learning can guide evolution. *Complex Systems*, 1(3):495–502, 1987.

[6] Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):3770, 2019.

[7] Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, December 2017.

[9] Alex Warstadt and Samuel R Bowman. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press, 2022.

[10] Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. Call for papers–the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*, 2023.

[11] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[12] Meta AI. Llama v3: Next-generation foundation language model, 2023. Unpublished release. Accessed on 2025-02-05.

[13] DeepSeek Team. Deepseek llm: Advancing deep information retrieval with large language models, 2023. Online release/preprint. Accessed on 2025-02-05.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models, 2020.

[15] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

[16] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

[17] Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, November 2020.

[18] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

[19] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

[20] Lukas Nagel, Diego Acuna, Elmar Schmerling, Marion Schenk, and Erich Elsen. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2022. arXiv preprint.

[21] Victor Sanh, Thomas Wolf, and Sebastian Ruder. Movement pruning: Adaptive sparsity by fine-tuning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 195–202. Association for Computational Linguistics, 2020.

[22] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[23] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[24] Mansheej Paul et al. Unmasking the lottery ticket hypothesis: What's encoded in a winning ticket's mask? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[25] Ashwinee Panda et al. Lottery ticket adaptation: Mitigating destructive interference in llms, 2024. arXiv preprint, `https://arxiv.org/abs/2406.16797v1`.

[26] FirstName Zheng and Others. Lottery tickets in llms: Robustness to adversarial examples via binary masking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022. Association for Computational Linguistics. Long paper.

[27] Mukund Varma T, Xuxi Chen, Zhenyu Zhang, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Sparse winning tickets are data-efficient image recognizers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[28] Hidenori Tanaka, Daniel Kunin, Daniel L Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in neural information processing systems*, 33:6377–6389, 2020.

[29] Hattie Zhou et al. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 3592–3602, 2019.

[30] Dániel L Barabási, Taliesin Beynon, Ádám Katona, and Nicolas Perez-Nieves. Complex computation from developmental priors. *Nature Communications*, 14(1):2226, 2023.

[31] Sergey Shuvaev, Divyansha Lachi, Alexei Koulakov, and Anthony Zador. Encoding innate ability through a genomic bottleneck. *Proceedings of the National Academy of Sciences*, 121(38):e2409160121, 2024.

[32] Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot, and Uri Hasson. Enhancing interpretability using human similarity judgements to prune word embeddings. *arXiv preprint arXiv:2310.10262*, 2023.

[33] Nhut Truong, Uri Hasson, et al. Pruning sparse features for cognitive modeling. In *The 7th annual conference on Cognitive Computational Neuroscience*, 2024.

[34] Betty Hart and Todd R Risley. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental psychology*, 28(6):1096, 1992.

[35] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[36] Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024.

[37] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11867–11878. AAAI Press, 2019.

[38] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Zou, David Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. arXiv preprint.

[39] Ethan G Wilcox. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*, 2020.

[40] Byung-Doh Oh and William 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, 2023.

[41] Cory Shain. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201, 2024.

[42] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

[43] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

[44] Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77, 2021.

[45] Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*, 2023.

[46] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. Brain-like language processing via a shallow untrained multihead attention network. *arXiv preprint arXiv:2406.15109*, 2024.

[47] Byung-Doh Oh and William Schuler. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *arXiv preprint arXiv:2304.11389*, 2023.

[48] Julius Steuer, Marius Mosbach, and Dietrich Klakow. Large gpt-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. *arXiv preprint arXiv:2311.04547*, 2023.

[49] Byung-Doh Oh, Shisen Yue, and William Schuler. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. *arXiv preprint arXiv:2402.02255*, 2024.

[50] Andrea De Varda and Marco Marelli. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, 2023.

[51] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

[52] Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, pages 1–18, 2024.

[53] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The llm language network: A neuroscientific approach for identifying causally task-relevant units. *arXiv preprint arXiv:2411.02280*, 2024.

[54] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[55] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[56] Chrisantha Fernando, Daniil Banarse, Charles Blundell, Yori Zwols, Oriol Vinyals, Daniel Apthorp, Daan Wierstra, and Koray Kavukcuoglu. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

[57] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[58] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[59] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

[60] Xiang Zhang et al. Root mean square layer normalization, 2019. arXiv preprint.

[61] Ruixiang Su, Xinjian Lu, Wei Li, Rui Xiong, Xiaoyu Li, Hao Chen, et al. Roformer: Enhanced transformer with rotary position embedding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[62] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[63] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

A Appendix

A.1 Model architecture

We utilize the upgraded GPT-2 architecture which has become the standard baseline implementation for this model family (36). It differs from the original GPT-2 paper (11) in three ways: firstly it uses RMSNorm without trained parameters (60), secondly it removes all bias parameters such that the model consists solely of 2D weight matrices, and thirdly it uses RoPE positional embedding (61) instead of learned positional embeddings. These architectural design choices were fixed before evaluating our generational pruning approach.

A.2 Dataset details

We utilize the first shard of the FineWeb (35) dataset - consisting of 100M tokens which make up 74,248,643 words - for the small dataset $\boldsymbol{S}$, and the subsequent 40 shards - consisting of 4B tokens which make up 2,969,847,300 words - for the large dataset $\boldsymbol{L}$.

A.3 Connectome compression

Our generational pruning framework iteratively zeros out 20% of the weights with the lowest magnitudes at each generation. After six generations (see Section 2, the final f_θ^5 connectome is a highly compressed representation consisting of just 124M ternary values where 25% are non-zero. The Shannon entropy of this connectome can be computed as:

$$H = -\left[0.75\log_2(0.75) + 0.125\log_2(0.125) + 0.125\log_2(0.125)\right] = 1.06$$

which means that encoding 1.06 bits of information per weight would result in a final compressed size of roughly $124M * 1.6$ bits = 131.44 Mb = 16MB. This corresponds to approximately a 15x compression over naïvely storing the dense model weights (248MB at 16 bits per weight), and an over 500x compression of the entire dataset $\boldsymbol{L}$ (8GB at 16 bits per token).

A.4 Model evaluation detailed methods

A.4.1 Natural language processing tasks

We evaluate our models on three standard NLP benchmarks: FineWeb validation loss (35), a diverse corpus of web-based text spanning various online content, HellaSwag (37), a benchmark that tests for commonsense reasoning in scenario completion, and MMLU (38), a multiple choice benchmark which tests for multi-domain knowledge and reasoning.

A.4.2 BEHAVIORAL ALIGNMENT

We analyze the self-paced reading dataset from Futrell et al. (2021) (44) (through Brain-Score (62)) consisting of word-by-word reading times from 179 participants across 5-10 naturalistic stories. The same stories are processed by our language models, and behavioral alignment is measured using the Pearson correlation between LLM per-word perplexity (summing sub-token perplexities for words that are split into multiple tokens per (63)) and human reading times.

A.4.3 BRAIN ALIGNMENT

We analyze fMRI data from Tuckute et al. (2024) (52) consisting of brain responses from 5 participants during a sentence-reading experiment. Participants read 1,000 6-word-long semantically and stylistically diverse sentences, and we investigated responses in the left-hemisphere language network (51), averaged across 5 participants. Following (52), brain alignment is measured by predicting brain responses using a ridge regression encoding model and computing the Pearson correlation between predicted and actual responses via 5-fold cross-validation.

A.5 BRAIN ALIGNMENT USING TOP 10% LANGUAGE-SELECTIVE UNITS

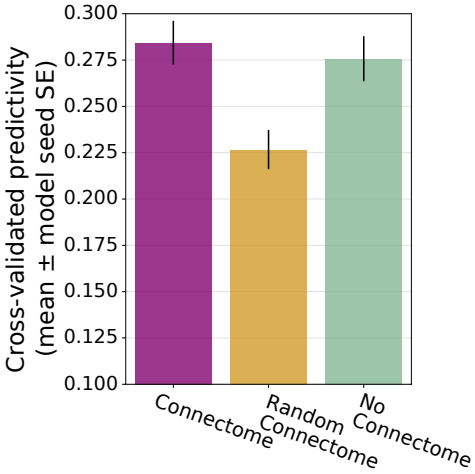

Figure 2: In the main text, we present model-brain alignment results using the top 1% language-selective units, per prior work (53) (Figure 1B, panel ii). However, in neuroscience, the top 10% units are typically used (51), and in this supplement we select the top 10% language-selective units in our models. The pattern is the same as in the top 1% case (Figure 1D, panel ii), but the overall correlations are higher. The error bars show show the standard error of the mean (SE) across model seeds.