The Neural Pile: 476 billion tokens of broad-coverage spiking neural activity data

A. Emin Orhan[†] **Feiyi Wang**
National Center for Computational Sciences
Oak Ridge National Laboratory
[†]orhanae@ornl.gov

Abstract

Foundation models pretrained with large-scale and rich domain-specific datasets facilitate scientific discovery and technological advances. Systems neuroscience currently lacks such foundation models, due mainly to two obstacles: (i) a lack of large-scale datasets and (ii) scarcity of large-scale compute to train high-capacity models. Here, we aim to address both challenges. We first introduce *the Neural Pile*, a large-scale curated dataset of spiking neural activity recorded from both primates and rodents. The dataset contains 34B uncompressed tokens of neural data from primates and 441B uncompressed tokens of neural data from rodents, involving multiple species and covering a wide range of brain regions, behaviors, and tasks. We provide a separate test split that is intended as a challenging neural prediction benchmark for evaluating neural foundation models. Secondly, as a strong baseline on this benchmark, we also release large-scale models (8B parameter models with a context length of 131k tokens) pretrained on *the Neural Pile*.

1 Introduction

High-capacity machine learning models trained with large-scale domain-specific datasets, commonly known as foundation models (Bommasani et al., 2021), are helping domain scientists and driving progress in a diverse range of scientific fields: *e.g.* protein folding (Jumper et al., 2021; Baek et al., 2021; Abramson et al., 2024), materials science (Zeni et al., 2025), genomics (Brixi et al., 2025; Hayes et al., 2025), weather and climate modeling (Bodnar et al., 2024; Schmude et al., 2024), astronomy (Nguyen et al., 2023b; Parker et al., 2024), among many others.

By and large, systems neuroscience currently lacks such foundation models. We attribute this predicament to two main factors: (i) a lack of large-scale, machine learning ready, broad-coverage, public datasets in neuroscience, and (ii) the scarcity of large-scale compute in academic labs to train high-capacity foundation models. To provide quantitative evidence for these claims, we estimate that currently the largest public spiking neural activity datasets that have been used for training machine learning models in neuroscience are only on the order of a few billion tokens in size and the largest models trained thus far are at most on the order of a few hundred million parameters in size and usually much smaller than this (Azabou et al., 2023; Ye et al., 2023; Zhang et al., 2024; Ye et al., 2025; Zhang et al., 2025) (Table 1). These numbers are exceedingly small compared to the data and model sizes commonly used in training large language models and foundation models in many other scientific domains. These twin challenges of scale hinder progress in training highly capable and generally useful foundation models in neuroscience.

In this paper, we aim to address both of these scaling challenges. We first introduce *the Neural Pile*, the largest machine learning ready public dataset of spiking neural activity data released to date, which is two orders of magnitude larger than previously available public datasets. The dataset contains a total of 476 billion tokens of curated neural data recorded from primates and rodents,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

Table 1: Recent studies training models on spiking neural activity data compared with the current work. Data size is measured in terms of uncompressed token counts. "Params" indicates the size of the largest model trained on the data and "Data acc." indicates the data accessibility status, *i.e.* whether the data are publicly accessible or private. The middle line separates the primate and rodent data.

Name	Data size	Params	Data acc.	Notes
POYO (Azabou et al., 2023)	∼2B tok.	13M	Public	Primate motor/premotor ctx data only
NDT-2 (Ye et al., 2023)	∼3B tok.	3M	Mixed	Primate motor/premotor ctx data only
NDT-3 (Ye et al., 2025)	∼30B tok.	350M	Private	Primate motor/premotor ctx data only
The Neural Pile (primate)	34.3B tok.	8B	Public	Wide variety of areas/tasks/sources
IBL-MtM (Zhang et al., 2024)	∼1B tok.	3M	Public	Rodent data (subset of IBL only)
NEDS (Zhang et al., 2025)	∼2B tok.	12M	Public	Rodent data (subset of IBL only)
The Neural Pile (rodent)	441.5B tok.	8B	Public	Wide variety of areas/tasks/sources

covering multiple species and a wide range of brain regions, tasks, and behaviors. Secondly, we also release 8 billion parameter generative models pretrained on *the Neural Pile*, again roughly two orders of magnitude larger than previous models trained on similar data. We hope that *the Neural Pile* and the large-scale pretrained generative models we are releasing with the current paper will catalyze the development of more capable and useful foundation models for neuroscience.

2 *The Neural Pile* dataset

To facilitate progress on training large-scale foundation models for neuroscience, we release *the Neural Pile* dataset. *The Neural Pile* consists of 476 billion tokens of spiking neural activity recorded from rodents and primates. We make the dataset available in two separate but identically structured Hugging Face dataset repositories, one for the rodent data and one for the primate data. Both repositories are accessible from the following Hugging Face collection: eminorhan/neural-pile.

Data curation: Data were curated from major data hosting websites popular among experimental neuroscience labs (specifically, the DANDI Archive, Dryad, Zenodo, Figshare, OSF, and G-Node) through simple keyword searches. Recordings from non-primate or non-rodent species, recordings performed under anesthesia or under unnatural stimulation conditions, and data types other than spiking neural activity were excluded from consideration. For the DANDI Archive, we made sure to search through the top 50 spiking activity datasets by token count (*i.e.* number of recorded neurons $\times$ time bins) using the DANDI API. For the other data hosting platforms, we did not aim to be comprehensive and we believe there is considerable room for extending *the Neural Pile* with additional public datasets through more comprehensive searches in the future.

The rodent subset of the dataset consists of 441.5 billion tokens of neural data curated from a total of 18 different sources. The primate subset, on the other hand, consists of 34.3 billion tokens of neural data curated from 23 different sources. The majority of our data come from repositories hosted on the DANDI Archive (34 of 41 data sources). The full list of data sources for both rodent and primate data, as well as additional information about the data from each source, can be found in Appendix A. The data contain recordings from multiple species (human, macaque, marmoset, rat, mouse) and multiple subjects, and cover a wide range of brain regions, stimuli, tasks, and behaviors.

In terms of token counts, the bulk of the rodent data comes from several large-scale recordings in mice performed by the Allen Institute and the International Brain Laboratory (IBL). The rodent data also include recordings from hippocampus and entorhinal cortex during various spatial navigation tasks; brain-wide recordings during spontaneous behaviors; recordings from piriform cortex in response to odors, *etc.* The primate data contain recordings from ventral visual areas in response to a variety of visual stimuli (including natural images); recordings from motor and premotor cortices during a variety of performed or attempted movements; recordings from parietal and frontal cortices during the performance of more cognitive tasks, among many others.

Most of our component datasets are not "trial-based" (*i.e.* we use whole recording sessions as our data). This constitutes an important difference between our dataset and the trial-based data used in many earlier works (Zhang et al., 2024, 2025). In the few cases where a component dataset was trial-based (see Appendix A), we concatenated the trials in their proper temporal order to preserve as much temporal information as possible. We did not restrict ourselves to trial-based data, because many of our curated datasets do not have a clear trial structure (or such information is not available in

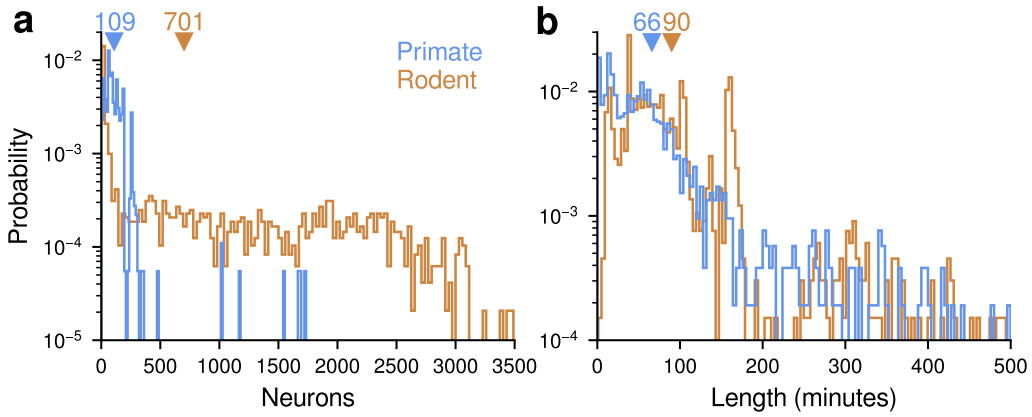

Figure 1: Distribution of simultaneously recorded neurons per session (**a**) and session lengths in minutes (**b**) for the primate and rodent subsets of *the Neural Pile* ($N = 1311$ sessions for the primate data and $N = 1676$ sessions for the rodent data). Inverted triangles (▼) indicate the means of the distributions. The rodent data have both a larger number of simultaneously recorded neurons per session and longer sessions on average.

the released data), but more importantly, considering only trial-based data would throw away a huge amount of rich information about the structure of neural activity: *e.g.* ongoing activity or activity recorded during spontaneous behaviors (not task-related) displays a rich high-dimensional structure that would be thrown away if we had only considered trial-based data (Stringer et al., 2019; Musall et al., 2019). In fact, considering only trial-based data has arguably been the main limitation of the previous studies training foundation models on neural data, as such trial-based data are inevitably severely limited both in amount and in scope.

Our entire dataset generation pipeline is fully reproducible. We make the code and detailed instructions for downloading and preprocessing the individual components of the dataset and merging them into a single dataset available in two public GitHub repositories: eminorhan/neural-pile-rodent and eminorhan/neural-pile-primate. These repositories are intended to make it as easy as possible to extend *the Neural Pile* with new datasets and we invite the users to add more datasets to it to create ever expanding versions of it.

Data preprocessing: All data were preprocessed into the common format of discrete spike counts from each recorded unit aggregated in 20 ms time bins. The spike count data from each experimental recording session were stored as a `uint8` array of size $n \times t$, where n represents the number of recorded units and t is the number of time bins in that session. We refer to each element in these arrays as an *uncompressed token* and the token counts reported in this paper are always in terms of these uncompressed tokens. Although most of our data come from sorted spikes, we make no distinction between sorted and unsorted spikes. Following common convention, we refer to the recorded units as *neurons* regardless of their sorting status.

Figure 1 shows the distribution of simultaneously recorded neurons per session (n) and session lengths (t converted to minutes) separately for the primate and the rodent subsets of *the Neural Pile*. The rodent data have both a larger number of simultaneously recorded neurons per session (701 *vs.* 109) and longer sessions on average (90 *vs.* 66 minutes), however the difference in token counts between the two subsets is primarily due to the first factor.

The dataset rows are pre-shuffled, so users do not have to shuffle them. In addition, we set aside 1% of both primate and rodent subsets of the dataset as a test split and we intend this test split to be used as a challenging neural prediction benchmark for neural foundation models.

In addition to the primary spike count data, each data row in the dataset also includes source dataset, subject, and session information. This auxiliary information can help users split the dataset in different ways to address different types of generalization questions in a rigorous way: for example, investigating *cross-session*, *cross-subject*, or *cross-dataset transfer*. Figure 2 shows some example

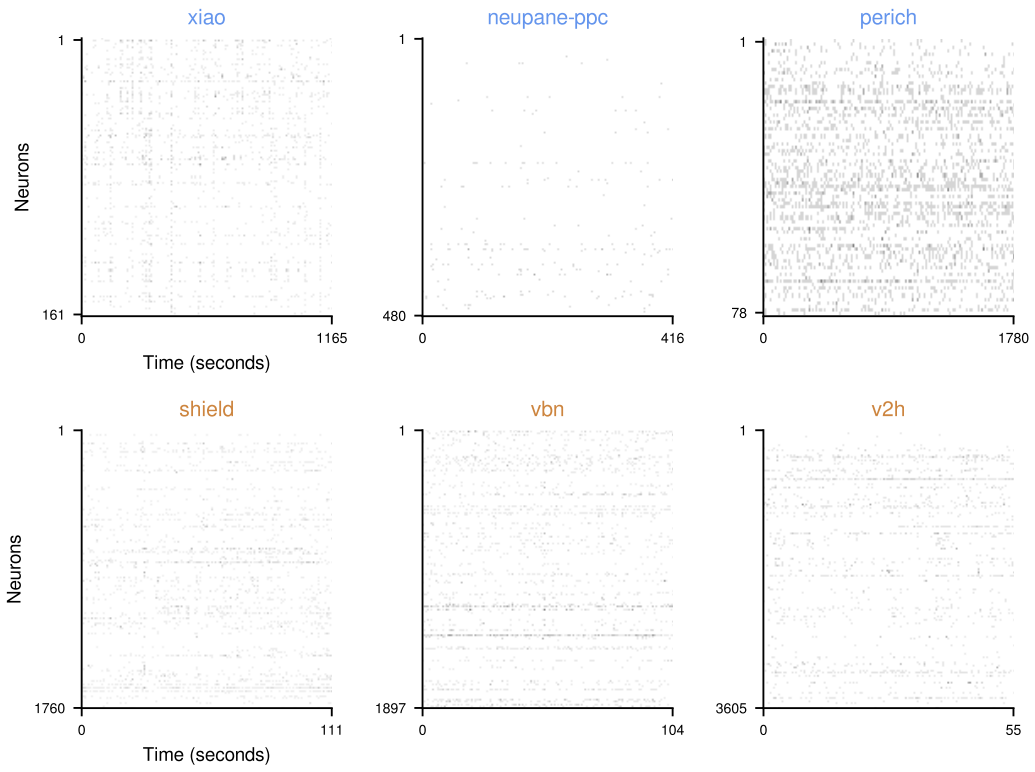

Figure 2: Example spike count arrays from the primate (upper row) and rodent (lower row) subsets of *the Neural Pile*. The source dataset for each spike count array is indicated in the title of the corresponding subplot. Details about the source datasets can be found in Appendix A.

spike count arrays from the primate and rodent subsets of *the Neural Pile*. Appendix B contains additional information about the dataset.

3 Models

We trained 8 billion parameter autoregressive generative models with a context length of 131072 tokens on the rodent and the primate subsets of *the Neural Pile* (models are available from the following Hugging Face collection: eminorhan/neural-pile-models). The models have the same architecture as the `Llama-3.1-8B` language model released by Meta (Grattafiori et al., 2024). The only difference is in the vocabulary size: the original `Llama-3` class models use a tokenizer with a large vocabulary size of 128k, whereas we use a much smaller vocabulary size of 256, aimed at directly modeling the discrete spike counts without any additional tokenization stage. We thus replace the embedding and the output layers in the original architecture with their appropriately resized versions. We choose to operate directly on discrete spike counts without tokenization (*i.e.* on uncompressed tokens), because this allows us to sidestep issues related to ensuring the generation of correctly shaped outputs when sampling autoregressively from the model. This choice is analogous to *byte-level language modeling* (Xue et al., 2022; Wu et al., 2024; Zheng et al., 2025), which is similarly motivated by a desire to avoid tokenization artifacts and to achieve more universal applicability. However, we encourage the users of the dataset to experiment with different tokenization schemes. Training large-scale models with very long context windows on the complete dataset will likely remain infeasible for academic compute budgets without some sort of tokenization that achieves significant compression rates (lossless or lossy compression).

To feed the data to the model, we flatten the spike count arrays in column-major order (*i.e.* in temporal order): first the spike counts of all neurons at $t = 0$, followed by the spike counts of all neurons at $t = 1$, *etc.* We separate the spike counts corresponding to different time bins in this flattened array with a special token that is distinct from all other tokens representing the spike counts (see Figure 4

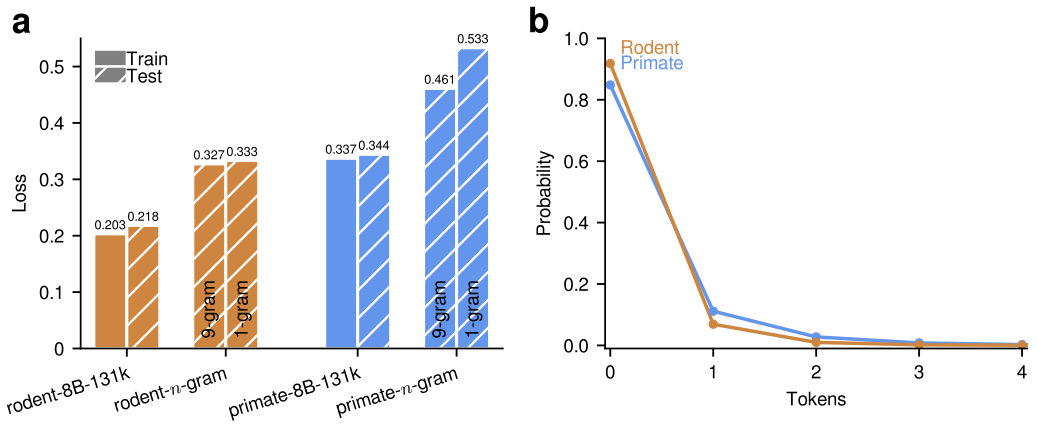

Figure 3: (**a**) Cross-entropy loss on the training and test splits of *the Neural Pile*. The losses are calculated on the rodent or the primate subsets of *the Neural Pile*. Training losses are estimated from the last 400 training steps (∼7.5B tokens). (**b**) Marginal token probabilities for the first five tokens estimated over the training splits.

in Appendix C). This special token effectively functions like a "beginning-of-sentence" (bos) token in language models and helps ensure the generation of correctly sized outputs at inference time while generating samples from the model. To keep the model as general as possible (and strongly motivated by the analogy with large language models), we do not explicitly feed any other auxiliary information (such as subject or session identifiers, or the recorded areas) to the model, relying instead on the ruthless efficiency of the language modeling objective to discover any relevant latent structure in the data. Further training details can be found in Appendix C.

4 Results

Figure 3a shows the final training and test losses of the 8B parameter models trained on the rodent and the primate subsets of *the Neural Pile* (denoted as `rodent-8B-131k` and `primate-8B-131k`, respectively). We first observe that the rodent data seem to be more predictable than the primate data (*e.g.* `rodent-8B-131k` *vs.* `primate-8B-131k` in Figure 3a). This is likely due to the fact that the marginal token count distribution is more skewed for the rodent data, making the rodent data less entropic than the primate data (Figure 3b): in particular, there is a significantly larger fraction of zero-spike tokens, *i.e.* time bins with a spike count of zero, in the rodent *vs.* primate data (92% *vs.* 85%). Secondly, we also trained n-gram models on both subsets of the data as baselines (more specifically, we trained 1-gram and 9-gram models using Laplace smoothing) and Figure 3a compares the performance of these n-gram models with the performance of the 8B parameter models. Even the more expressive 9-gram model performs substantially worse than the corresponding 8B parameter model, demonstrating that achieving high predictive performance on *the Neural Pile* requires more sophisticated models that can model the long-term dependencies between tokens.

5 Discussion

We have introduced *the Neural Pile*, the largest public dataset of broad-coverage spiking neural activity data released to date and 8B parameter foundation models trained on this dataset, again the largest such models released to date. We intend *the Neural Pile* to be used primarily as a large-scale pretraining dataset (similar to the large-scale text datasets like *the Pile* (Gao et al., 2020) that inspired it in the first place) to train high-capacity models that can learn the rich structure of high-dimensional neural activity across a variety of contexts. Such models would be expected to benefit neural decoding applications in general (see Appendix D for a more detailed discussion); however, current neural decoding tasks are likely not challenging enough to effectively leverage large-scale pretraining (Ye et al., 2025). Developing more challenging neural decoding tasks should be a priority for future work. In the meantime, we advocate using neural prediction, as opposed to neural decoding, as the primary method to quantitatively evaluate large-scale neural foundation models. With its large size, long-term temporal structure, and the wide variety of data sources, brain regions, and behaviors it represents, *the Neural Pile* would serve as an ideal neural prediction benchmark for neural foundation models.

Acknowledgments and Disclosure of Funding

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500.

Allen Institute for Brain Science (2019). Allen Brain Observatory: Neuropixels Visual Coding. Technical report. Accessed: 2025-04-28.

Allen Institute for Brain Science (2022). Allen Brain Observatory: Visual Behavior Neuropixels. Technical report. Accessed: 2025-04-28.

Athalye, V. R., Khanna, P., Gowda, S., Orsborn, A. L., Costa, R. M., and Carmena, J. M. (2023). Invariant neural dynamics drive commands to control different movements. *Current Biology*, 33(14):2962–2976.

Azabou, M., Arora, V., Ganesh, V., Mao, X., Nachimuthu, S., Mendelson, M., Richards, B., Perich, M., Lajoie, G., and Dyer, E. (2023). A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36:44937–44956.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876.

Bennett, C., Ouellette, B., Ramirez, T. K., Cahoon, A., Cabasco, H., Browning, Y., Lakunina, A., Lynch, G. F., McBride, E. G., Belski, H., et al. (2024). Shield: Skull-shaped hemispheric implants enabling large-scale electrophysiology datasets in the mouse brain. *Neuron*, 112(17):2869–2885.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Vaughan, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., et al. (2024). A foundation model for the earth system. *arXiv preprint arXiv:2405.13063*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., et al. (2025). Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, pages 2025–02.

Chowdhury, R. H., Glaser, J. I., and Miller, L. E. (2020). Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. *eLife*, 9:e48198.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405):51–56.

Dehaqani, A. A., Michelon, F., Patella, P., Petrucco, L., Piasini, E., and Iurilli, G. (2024). A mechanosensory feedback that uncouples external and self-generated sensory responses in the olfactory cortex. *Cell Reports*, 43(4).

Even-Chen, N., Sheffer, B., Vyas, S., Ryu, S. I., and Shenoy, K. V. (2019). Structure and variability of delay activity in premotor cortex. *PLoS Computational Biology*, 15(2):e1006808.

Fan, C., Hahn, N., Kamdar, F., Avansino, D., Wilson, G., Hochberg, L., Shenoy, K. V., Henderson, J., and Willett, F. (2023). Plug-and-play stability for intracortical brain-computer interfaces: A one-year demonstration of seamless brain-to-text communication. *Advances in Neural Information Processing Systems*, 36:42258–42270.

Finkelstein, A., Fontolan, L., Economo, M. N., Li, N., Romani, S., and Svoboda, K. (2021). Attractor dynamics gate cortical information flow during decision-making. *Nature Neuroscience*, 24(6):843–850.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gonzalez, A. and Giocomo, L. M. (2024). Parahippocampal neurons encode task-relevant information for goal-directed navigation. *eLife*, 12:RP85646.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. (2025). Simulating 500 million years of evolution with a language model. *Science*, page eads0018.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Hobbhahn, M. and Besiroglu, T. (2022). Trends in GPU Price-Performance. https://epoch.ai/blog/trends-in-gpu-price-performance.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Huszár, R., Zhang, Y., Blockus, H., and Buzsáki, G. (2022). Preconfigured dynamics in the hippocampus are guided by embryonic birthdate and rate of neurogenesis. *Nature Neuroscience*, 25(9):1201–1212.

International Brain Laboratory, Benson, B., Benson, J., Birman, D., Bonacchi, N., Bougrova, K., Bruijns, S. A., Carandini, M., Catarino, J. A., Chapuis, G. A., et al. (2023). A brain-wide map of neural activity during complex behaviour. *bioRxiv*, pages 2023–07.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Karpowicz, B., Ye, J., Fan, C., Tostado-Marcos, P., Rizzoglio, F., Washington, C., Scodeler, T., de Lucena, D., Nason-Tomaszewski, S., Mender, M., et al. (2024). Few-shot algorithms for consistent neural decoding (falcon) benchmark. *Advances in Neural Information Processing Systems*, 37:76578–76615.

Lanzarini, F., Maranesi, M., Rondoni, E. H., Albertini, D., Ferretti, E., Lanzilotto, M., Micera, S., Mazzoni, A., and Bonini, L. (2025). Neuroethology of natural actions in freely moving monkeys. *Science*, 387(6730):214–220.

Li, N., Chen, T.-W., Guo, Z. V., Gerfen, C. R., and Svoboda, K. (2015). A motor cortex circuit for motor planning and movement. *Nature*, 519(7541):51–56.

Liang, W., Liu, T., Wright, L., Constable, W., Gu, A., Huang, C.-C., Zhang, I., Feng, W., Huang, H., Wang, J., et al. (2024). TorchTitan: One-stop PyTorch native solution for production ready LLM pre-training. *arXiv preprint arXiv:2410.06511*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Makin, J. G., O'Doherty, J. E., Cardoso, M. M., and Sabes, P. N. (2018). Superior arm-movement decoding from cortex with a new, unsupervised-learning algorithm. *Journal of Neural Engineering*, 15(2):026010.

Mallory, C. S., Hardcastle, K., Campbell, M. G., Attinger, A., Low, I. I., Raymond, J. L., and Giocomo, L. M. (2021). Mouse entorhinal cortex encodes a diverse repertoire of self-motion signals. *Nature Communications*, 12(1):671.

Mehrotra, D., Levenstein, D., Duszkiewicz, A. J., Carrasco, S. S., Booker, S. A., Kwiatkowska, A., and Peyrache, A. (2024). Hyperpolarization-activated currents drive neuronal activation sequences in sleep. *Current Biology*, 34(14):3043–3054.

Mehta, M. R., Purandare, C., Jha, S., Lecoq, J., Durand, S., Gillis, R., Belski, H., Bawany, A., Carlson, M., Peene, C., Wilkes, J., Johnson, T., Naidoo, R., Suarez, L., Han, W., Amaya, A., Nguyen, K., Ouellette, B., Swapp, J., and Williford, A. (2025). Allen Institute Openscope - Vision2Hippocampus project.

Mineault, P., Zanichelli, N., Peng, J. Z., Arkhipov, A., Bingham, E., Jara-Ettinger, J., Mackevicius, E., Marblestone, A., Mattar, M., Payne, A., et al. (2024). NeuroAI for AI safety. *arXiv preprint arXiv:2411.18526*.

Moore, D. D., MacLean, J. N., Walker, J. D., and Hatsopoulos, N. G. (2024). A dynamic subset of network interactions underlies tuning to natural movements in marmoset sensorimotor cortex. *Nature Communications*, 15(1):1–16.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., and Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22(10):1677–1686.

Namima, T., Kempkes, E., Zamarashkina, P., Owen, N., and Pasupathy, A. (2025). High-density recording reveals sparse clusters (but not columns) for shape and texture encoding in macaque v4. *Journal of Neuroscience*, 45(5).

Nason, S. R., Mender, M. J., Vaskov, A. K., Willsey, M. S., Kumar, N. G., Kung, T. A., Patil, P. G., and Chestek, C. A. (2021). Real-time linear prediction of simultaneous and independent movements of two finger groups using an intracortical brain-machine interface. *Neuron*, 109(19):3164–3177.

Neupane, S., Fiete, I., and Jazayeri, M. (2024). Mental navigation in the primate entorhinal cortex. *Nature*, 630(8017):704–711.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. (2023a). HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36:43177–43201.

Nguyen, T. D., Ting, Y.-S., Ciucă, I., O'Neill, C., Sun, Z.-C., Jabłońska, M., Kruk, S., Perkowski, E., Miller, J., Li, J., et al. (2023b). AstroLLaMA: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2309.06126*.

Papale, P., Wang, F., Self, M. W., and Roelfsema, P. R. (2025). An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*, 113(4):539–553.

Parker, L., Lanusse, F., Golkar, S., Sarra, L., Cranmer, M., Bietti, A., Eickenberg, M., Krawezik, G., McCabe, M., Morel, R., et al. (2024). AstroCLIP: A cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011.

Pei, F., Ye, J., Zoltowski, D., Wu, A., Chowdhury, R. H., Sohn, H., O'Doherty, J. E., Shenoy, K. V., Kaufman, M. T., Churchland, M., et al. (2021). Neural Latents Benchmark'21: Evaluating latent variable models of neural population activity. *arXiv preprint arXiv:2109.04463*.

Perich, M. G., Gallego, J. A., and Miller, L. E. (2018). A neural population mechanism for rapid learning. *Neuron*, 100(4):964–976.

Petersen, P. C. and Buzsáki, G. (2020). Cooling of medial septum reveals theta phase lag coordination of hippocampal cell assemblies. *Neuron*, 107(4):731–744.

Rajalingham, R., Sohn, H., and Jazayeri, M. (2025). Dynamic tracking of objects in the macaque dorsomedial frontal cortex. *Nature Communications*, 16(1):346.

Rouse, A. G. and Schieber, M. H. (2015). Spatiotemporal distribution of location and object effects in reach-to-grasp kinematics. *Journal of Neurophysiology*, 114(6):3268–3282.

Schmude, J., Roy, S., Trojak, W., Jakubik, J., Civitarese, D. S., Singh, S., Kuehnert, J., Ankur, K., Gupta, A., Phillips, C. E., et al. (2024). Prithvi WxC: Foundation model for weather and climate. *arXiv preprint arXiv:2409.13598*.

Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., and Dao, T. (2024). FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685.

Shin, H., Ogando, M. B., Abdeladim, L., Durand, S., Belski, H., Cabasco, H., Loefler, H., Bawany, A., Hardcastle, B., Wilkes, J., et al. (2023). Recurrent pattern completion drives the neocortical representation of sensory inference. *bioRxiv*.

Sohn, H., Narain, D., Meirhaeghe, N., and Jazayeri, M. (2019). Bayesian computation through cortical latent dynamics. *Neuron*, 103(5):934–947.

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., and Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273.

Stevenson, I. H. and Körding, K. P. (2011). How advances in neural recording affect data analysis. *Nature Neuroscience*, 14(2):139–142.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., and Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Westerberg, J., Durand, S., Cabasco, H., Belski, H., Loeffler, H., Bawany, A., Peene, R. C., Han, W., Nguyen, K., Ha, V., Johnson, T., Grasso, C., Hardcastle, B., Young, A., Swapp, J., Gillis, R., Ouellette, B., Caldejon, S., Williford, A., Groblewski, A. P., Olsen, S., Kiselycznyk, C., Lecoq, J., Maier, A., and Bastos, A. (2024). Allen Institute Openscope - Global/Local Oddball project.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., et al. (2023). A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

Wodlinger, B., Downey, J., Tyler-Kabara, E., Schwartz, A., Boninger, M., and Collinger, J. (2014). Ten-dimensional anthropomorphic arm control in a human brain- machine interface: difficulties, solutions, and limitations. *Journal of Neural Engineering*, 12(1):016011.

Wójcik, M. J., Stroud, J. P., Wasmuht, D., Kusunoki, M., Kadohisa, M., Buckley, M. J., Myers, N. E., Hunt, L. T., Duncan, J., and Stokes, M. G. (2023). Learning shapes neural geometry in the prefrontal cortex. *bioRxiv*, pages 2023–04.

Wu, S., Tan, X., Wang, Z., Wang, R., Li, X., and Sun, M. (2024). Beyond language models: Byte models are digital world simulators. *arXiv preprint arXiv:2402.19155*.

Xiao, W., Sharma, S., Kreiman, G., and Livingstone, M. S. (2024). Feature-selective responses in macaque visual cortex follow eye movements during natural vision. *Nature Neuroscience*, 27(6):1157–1166.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Ye, J., Collinger, J., Wehbe, L., and Gaunt, R. (2023). Neural Data Transformer 2: Multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36:80352–80374.

Ye, J., Rizzoglio, F., Smoulder, A., Mao, H., Ma, X., Marino, P., Chowdhury, R., Moore, D., Blumenthal, G., Hockeimer, W., et al. (2025). A generalist intracortical motor decoder. *bioRxiv*, pages 2025–02.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Wang, Z., Shysheya, A., Crabbé, J., Ueda, S., et al. (2025). A generative model for inorganic materials design. *Nature*, pages 1–3.

Zhang, Y., Wang, Y., Azabou, M., Andre, A., Wang, Z., Lyu, H., Laboratory, T. I. B., Dyer, E., Paninski, L., and Hurwitz, C. (2025). Neural encoding and decoding at scale. *arXiv preprint arXiv:2504.08201*.

Zhang, Y., Wang, Y., Jiménez-Benetó, D., Wang, Z., Azabou, M., Richards, B., Tung, R., Winter, O., Dyer, E., Paninski, L., et al. (2024). Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution. *Advances in Neural Information Processing Systems*, 37:80495–80521.

Zheng, L., Zhao, X., Wang, G., Wu, C., Dong, D., Wang, A., Wang, M., Du, Y., Bo, H., Sharma, A., Li, B., Zhang, K., Hu, C., Thakker, U., and Kong, L. (2025). EvaByte: Efficient byte-level language models at scale. https://hkunlp.github.io/blog/2025/evabyte.

A Individual components of *the Neural Pile* dataset

Table 2 and Table 3 show the full list of component datasets making up the primate and rodent subsets of *the Neural Pile* dataset, respectively. In cases where the dataset already had a name, we used the same name for it; otherwise, we used the name of the listed contact person or the name of the first author of the paper introducing the dataset. For each dataset, we provide below some basic information about the recorded brain region(s), as well as the task, behavior, and/or stimulus conditions. Further details about the recordings and the underlying data from each dataset can be obtained from the corresponding reference.

Table 2: Individual components of the primate subset of *the Neural Pile*.

Name	Tokens	Source	Species	Subjects	Sessions
Xiao	17,695,820,059	dandi:000628	macaque	13	679
Neupane (PPC)	7,899,849,087	dandi:001275	macaque	2	10
Willett	1,796,119,552	dryad:x69p8czpq	human	1	44
Churchland	1,278,669,504	dandi:000070	macaque	2	10
Neupane (EC)	911,393,376	dandi:000897	macaque	2	15
Kim	804,510,741	dandi:001357	macaque	2	159
Even-Chen	783,441,792	dandi:000121	macaque	2	12
Papale	775,618,560	g-node:TVSD	macaque	2	2
Perich	688,889,368	dandi:000688	macaque	4	111
Wojcik	422,724,515	dryad:c2fqz61kb	macaque	2	50
Makin	375,447,744	zenodo:3854034	macaque	2	47
H2	297,332,736	dandi:000950	human	1	47
Lanzarini	259,179,392	osf:82jfr	macaque	2	10
Athalye	101,984,317	dandi:000404	macaque	2	13
M1-A	45,410,816	dandi:000941	macaque	1	11
M1-B	43,809,344	dandi:001209	macaque	1	12
H1	33,686,576	dandi:000954	human	1	40
Moore	30,643,839	dandi:001062	marmoset	1	1
Temmar	27,388,320	dandi:001201	macaque	1	12
Rajalingham	14,923,100	zenodo:13952210	macaque	2	2
DMFC-rsg	14,003,818	dandi:000130	macaque	1	2
M2	12,708,384	dandi:000953	macaque	1	20
Area2-bump	7,394,070	dandi:000127	macaque	1	2

A.1 Primate subset of *the Neural Pile*

Xiao: Recordings from six areas in the macaque ventral visual pathway (V1, V2 ,V4, PIT, CIT, AIT) in response to thousands of natural images (Xiao et al., 2024).

Neupane (PPC): Recordings from the macaque posterior parietal cortex (PPC) during a mental navigation task (Neupane et al., 2024).

Willett: Recordings from the human premotor cortex during attempted speech (Willett et al., 2023).

Churchland: Recordings from the macaque motor cortex (M1) and dorsal premotor cortex (PMd) while performing reaching tasks with the right hand (Churchland et al., 2012).

Neupane (EC): Recordings from the macaque entorhinal cortex (EC) during a mental navigation task (Neupane et al., 2024).

Kim: Recordings from the macaque ventral visual area V4 in response to shape and texture stimuli (Namima et al., 2025).

Even-Chen: Recordings from the macaque dorsal premotor cortex (PMd) during cursor movement tasks (Even-Chen et al., 2019).

Papale: Recordings from the macaque ventral visual areas V1, V4, and IT in response to $\sim 22k$ natural images from the THINGS dataset (Papale et al., 2025). This dataset is trial-based. We concatenate consecutive trials end-to-end to preserve as much temporal information as possible.

Perich: Recordings from the macaque motor cortex (M1) and dorsal premotor cortex (PMd) during reaching movements (Perich et al., 2018).

Wojcik: Recordings from the macaque prefrontal cortex (PFC) during a rule learning task (Wójcik et al., 2023). This dataset is trial-based. We concatenate consecutive trials end-to-end to preserve as much temporal information as possible.

Makin: Recordings from the macaque motor cortex (M1) and sensorimotor cortex (S1) during reaching movements (Makin et al., 2018).

H2: Recordings from the human dorsal motor cortex ("hand knob" area) during attempted handwriting trials (Fan et al., 2023). This dataset is part of the FALCON benchmark for neural decoding (Karpowicz et al., 2024).

Lanzarini: Recordings from the macaque premotor and motor cortex under conditions of head restraint and free movement (Lanzarini et al., 2025).

Athalye: Recordings from the macaque premotor and motor cortex (PMd/M1) during center-out reaching and obstacle avoidance tasks (Athalye et al., 2023).

M1-A: Recordings from the macaque motor cortex (M1) during a center-out reaching task (Rouse and Schieber, 2015). This dataset is part of the FALCON benchmark for neural decoding (Karpowicz et al., 2024).

M1-B: Recordings from the macaque motor cortex (M1) during a center-out reaching task (Rouse and Schieber, 2015). Same as above, but from a different monkey. This dataset is part of the FALCON benchmark for neural decoding (Karpowicz et al., 2024).

H1: Recordings from the human motor cortex ("hand knob" area) during the attempted control of a prosthetic limb (Wodlinger et al., 2014). This dataset is part of the FALCON benchmark for neural decoding (Karpowicz et al., 2024).

Moore: Recordings from the marmoset primary motor and somatosensory cortex during a naturalistic prey capture task (Moore et al., 2024).

Temmar: Recordings from the macaque motor cortex (M1) during a self-paced finger movement task Nason et al. (2021).

Rajalingham: Recordings from the macaque dorsomedial frontal cortex (DMFC) during a naturalistic ball interception task (Rajalingham et al., 2025). This dataset is trial-based. We concatenate consecutive trials end-to-end to preserve as much temporal information as possible.

DMFC-rsg: Recordings from the macaque dorsomedial frontal cortex (DMFC) during a time-interval reproduction task (Sohn et al., 2019). This dataset is part of the Neural Latents benchmark for neural decoding (Pei et al., 2021).

M2: Recordings from the macaque motor cortex (M1) during a finger movement task (Nason et al., 2021). This dataset is part of the FALCON benchmark for neural decoding (Karpowicz et al., 2024).

Area2-bump: Recordings from the macaque somatosensory area 2 during a reaching task with perturbations (Chowdhury et al., 2020). This dataset is part of the Neural Latents benchmark for neural decoding (Pei et al., 2021).

A.2 Rodent subset of *the Neural Pile*

Table 3: Individual components of the rodent subset of *the Neural Pile*.

Name	Tokens	Source	Species	Subjects	Sessions
VBN	153,877,057,200	dandi:000713	mouse	81	153
IBL	69,147,814,139	dandi:000409	mouse	115	347
SHIELD	61,890,305,241	dandi:001051	mouse	27	99
VCN	36,681,686,005	dandi:000021	mouse	32	32
VCN-2	30,600,253,445	dandi:000022	mouse	26	26
V2H	24,600,171,007	dandi:000690	mouse	25	25
Petersen	15,510,368,376	dandi:000059	rat	5	24
Oddball	14,653,641,118	dandi:000253	mouse	14	14
Illusion	13,246,412,456	dandi:000248	mouse	12	12
Huszar	8,812,474,629	dandi:000552	mouse	17	65
Steinmetz	7,881,422,592	dandi:000017	mouse	10	39
Finkelstein	1,313,786,316	dandi:000060	mouse	9	98
Giocomo	1,083,328,404	dandi:000053	mouse	34	349
Steinmetz-2	684,731,334	figshare:7739750	mouse	3	3
Mehrotra	465,402,824	dandi:000987	mouse	3	14
Iurilli	388,791,426	dandi:000931	mouse	1	1
Gonzalez	366,962,209	dandi:000405	rat	5	276
Li	260,807,325	dandi:000010	mouse	23	99

Visual Behavior - Neuropixels (VBN): Large-scale recordings from the mouse visual cortical areas, including VISp, VISl, VISal, VISrl, VISam, and VISpm and several subcortical areas, including visual thalamic areas LGd and LP as well as from the hippocampus and midbrain (up to 6 probes at a time). The task is a visual change detection task (Allen Institute for Brain Science, 2022).

International Brain Laboratory (IBL) - Brain-wide map: Brain-wide recordings from mice during a decision-making task with sensory, motor, and cognitive components (International Brain Laboratory et al., 2023).

SHIELD: Simultaneous recordings from multiple cortical and subcortical areas in mice during a visual change detection task closely related to the VBN task above (Bennett et al., 2024).

Visual Coding - Neuropixels (VCN): Simultaneous recordings from multiple cortical and subcortical areas in mice in response to a variety of visual stimuli (Allen Institute for Brain Science, 2019).

Visual Coding - Neuropixels (VCN-2): Simultaneous recordings from multiple cortical and subcortical areas in mice in response to a variety of visual stimuli (Allen Institute for Brain Science, 2019). Same task and recording conditions as above.

Vision2Hippocampus (V2H): Simultaneous recordings from multiple cortical and subcortical areas in mice in response to a variety of visual stimuli (Mehta et al., 2025).

Petersen: Recordings from the rat hippocampus during a spatial navigation task (includes recordings made when the medial septum was cooled) (Petersen and Buzsáki, 2020).

Oddball: Simultaneous recordings from multiple cortical areas in mice in response to simple visual stimuli (Westerberg et al., 2024).

Illusion: Simultaneous recordings from multiple visual cortical areas (V1, LM, RL, AL, PM, AM) in mice in response to illusory contour stimuli (Shin et al., 2023).

Huszar: Recordings from the CA1 region of mouse hippocampus during a spatial alternation task in a figure-eight maze (Huszár et al., 2022).

Steinmetz: Brain-wide recordings from mice during a perceptual decision-making task (Steinmetz et al., 2019).

Finkelstein: Recordings from the anterior lateral motor cortex and the vibrissal sensory cortex of mice trained to detect optogenetic stimulation of the vibrissal sensory cortex (Finkelstein et al., 2021).

Giocomo: Tetrode recordings from the medial entorhinal cortex in mice during open field navigation and Neuropixels recordings from the medial entorhinal cortex in mice during navigation down a virtual linear track (Mallory et al., 2021).

Steinmetz-2: Brain-wide recordings from mice with eight Neuropixels probes during spontaneous behaviors (Stringer et al., 2019).

Mehrotra: Recordings from the mouse retrosplenial cortex during sleep and open field exploration (Mehrotra et al., 2024).

Iurilli: Recordings from the mouse piriform cortex in response to odors under different inhalation speeds (Dehaqani et al., 2024).

Gonzalez: Recordings from the rat parahippocampal cortex during a memory-guided spatial navigation task (Gonzalez and Giocomo, 2024).

Li: Recordings from the mouse anterior lateral motor cortex (ALM) during a whisker-based object location discrimination task (Li et al., 2015).

A.3 License information

All datasets deposited on the DANDI Archive have a CC BY 4.0 license. Among the primate datasets that are not hosted on the DANDI Archive, Willett and Wojcik are in the public domain (CC0); Papale and Makin have CC BY 4.0 licenses; Rajalingham has a CC BY-NC-ND 4.0 license. We could not obtain license information regarding the Lanzarini dataset. Among the rodent datasets that are not hosted on the DANDI Archive, Steinmetz-2 have a CC BY-NC 4.0 license. Please follow the source links provided in Table 2 and Table 3 above for more detailed license information.

B Further details about the dataset

Naming: We chose the name *The Neural Pile* as a tribute to *The Pile* (Gao et al., 2020), one of the first large-scale, diverse, high-quality, and open text datasets for language modeling. *The Pile* launched the development of the first generation of open large language models. We hope that the release of *The Neural Pile* will similarly spark community interest in developing large-scale open foundation models for neuroscience.

Storage requirements: The spike count arrays are stored in `uint8` format. The choice of `uint8` data type gives us sufficient range for the lossless representation of all our spike count data while minimizing the memory requirements for storage. The rodent subset of *the Neural Pile* takes up about 47 GB when stored as `parquet` files and 443 GB when stored as memory-mapped `arrow` files; the primate subset, on the other hand, takes up about 6 GB when stored as `parquet` files and 34 GB when stored as memory-mapped `arrow` files. The local caching system of the Hugging Face `datasets` library uses `arrow` files for storage, so users will need to have a minimum of 477 GB of free disk space in order to cache the entire *Neural Pile* dataset on disk (recommended).

Splitting of long recording sessions: When creating the dataset rows, instead of storing each session as a single row, we split long recording sessions (where the spike count array was larger than 10M tokens) into smaller equal-sized chunks of no more than 10M tokens each. For a 100-neuron recording, this roughly corresponds to 2000 second long chunks; whereas for a 1000-neuron recording, it corresponds to 200 second long chunks, so even for sessions with a large number of simultaneously recorded neurons, a substantial amount of temporal information (on the order of 100 seconds) is still preserved in each chunk. In our experience, splitting long sessions into smaller

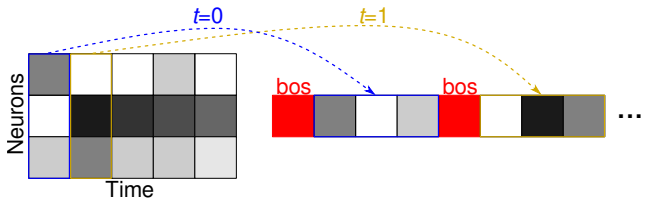

Figure 4: Spike count arrays (left) are flattened in column-major order and spike counts corresponding to different time bins are separated with a special token (represented by the red bos token).

chunks in this way makes data loading more efficient and prevents memory errors when creating Hugging Face dataset classes and pushing them to the Hugging Face Hub. The 10M-token long data rows and the relatively long-term temporal information contained in each row also make our dataset potentially suitable as a benchmark for long context training and evaluation. We expect techniques developed for efficient long context training on, for example, DNA data (Nguyen et al., 2023a) to be effective for our temporally extended spike count data as well.

C Further details about the models and training

Training infrastructure: The models were trained on an HPE Cray EX254n supercomputer with 168 NVIDIA GH200 superchips (42 nodes connected with Slingshot-11 interconnect $\times$ 4 GH200s on each node with each GH200 having 96 GB HBM3 high-bandwidth GPU memory). For training each model, we used a total of 128 superchips across 32 nodes of the system.

Parallelisms: We observed that fully-sharded data parallelism (FSDP), without any additional parallelization techniques, was sufficient to train our 8B parameter models with a context size of 131072 tokens. To be able to train the models efficiently at scale, in addition to FSDP, we also utilized full activation checkpointing, `bf16` mixed precision training, and the Hopper-optimized FlashAttention-3 kernels for the self-attention computations (Shah et al., 2024).

Implementation: To train the models, we used a customized version of the `torchtitan` library (Liang et al., 2024), a lightweight, PyTorch-native distributed training framework. The training code is available at: eminorhan/gpt-neuro-arch with detailed instructions for full reproduction.

Context length & batch size: The models were trained with a context length of 131072 tokens and they consumed 17M tokens globally per training step: *i.e.* **128** data-parallel ranks (FSDP) $\times$ local batch size of **1** per data-parallel rank $\times$ **131072** token context length. Each training step took around 53 seconds to complete and we were able to achieve a very respectable model FLOPS utilization (MFU) rate of around 62% consistently throughout training. The context length of 131072 tokens is large enough to span 26.2 seconds of recording for a 100-neuron recording, and 2.6 seconds for a 1000-neuron recording.

Training objective: We used the cross-entropy loss to train the models autoregressively, minimizing the mean cross-entropy between the actual and predicted tokens for each token, given all previous tokens in the flattened input sequence. Cross-entropy loss enjoys the advantage of greater expressivity over the popular Poisson negative loglikelihood loss used in most of the prior work (Ye et al., 2023; Zhang et al., 2024, 2025). Cross-entropy loss assumes that the underlying distribution of the discrete count variable is a categorical distribution. In effect, this allows us to model the whole distribution over the discrete count variable non-parametrically rather than assuming that it is Poisson. Over a finite range, categorical distribution is thus strictly more general than Poisson: for instance, Poisson cannot model multimodal count distributions, but a categorical distribution can.

Learning rate schedule: We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a linear learning rate schedule to train the models. For the rodent model, the learning rate was linearly warmed up to a maximum of 0.0003 over the initial 1000 training steps, after which it decayed back to 0 linearly over 74000 training steps. For the primate model, the learning rate was similarly warmed up to a maximum of 0.0003 over the initial 1000 training steps, after which it decayed back to 0 linearly over 24000 training steps.

D The case for large-scale datasets and foundation models for neuroscience

The case for large-scale, high-dimensional, rich datasets and high-capacity foundation models for systems neuroscience largely rests on the analogy with large language models and foundation models in other scientific domains, where the push for larger datasets and bigger models resulted in increasingly capable models and sometimes in surprising emergent capabilities (Hestness et al., 2017; Hoffmann et al., 2022; Wei et al., 2022). It is reasonable to expect a similar development in neuroscience with increasingly capable and useful foundation models as the data size, the dimensionality or the richness of the data, and the model size are all scaled up. Unlike for large language models, however, in neuroscience the data size and dimensionality are bound by the neural recording technology and this is likely to be a more significant constraint than compute in scaling up neural foundation models in the long term. For example, Stevenson and Körding (2011) estimate that the number of simultaneously recorded neurons doubles only every ~ 7 years or so, whereas currently the doubling time for GPU FLOPS is probably closer to ~ 1 year ("Huang's law") (Hobbhahn and Besiroglu, 2022).

We envision several promising use cases for large-scale datasets and foundation models in neuroscience in the coming years and decades. Some of these use cases are likely already feasible to some extent with the current technology, while others (especially the last two below) may require further advances in recording technology to enable denser sampling of neural activity in primates under more natural conditions.

Brain-computer interfaces: High-capacity models trained with large amounts of neural data would be expected to improve the performance, robustness, and versatility of brain-computer interfaces (BCIs) and neuroprosthetics. BCIs, in general, involve mapping neural activity to external signals, *e.g.* inputs to an electronic device, such as a computer or a speech synthesizer. A high-capacity model with fine-grained knowledge of the semantics of neural activity (*i.e.* what precisely different patterns of neural activity mean and how their meanings relate to each other) could facilitate the learning of fine-grained, complex mappings between neural activity and external signals, analogous to how pretrained large language models facilitate the execution of various downstream tasks, often to such an extent that they obviate the need for learning the downstream task altogether by being able to perform it zero-shot. Similar "zero-shot BCIs" may be feasible in the future with highly capable neural foundation models.

Digital twins of the brain: A model that can mimic the responses of a neural system well enough under a sufficiently broad range of conditions can serve as a digital twin of that system. We can expect high-capacity models trained with large amounts of relevant neural data to be able to function as digital twins in this sense. Digital twins of neural systems can help reduce the amount of animal experimentation, speed up the process of hypothesis generation and testing for neuroscientists, and thus facilitate scientific discoveries and treatments for brain disorders.

Neural data for inducing capabilities in machine learning models: Large-scale recordings of neural activity can also potentially be useful as an unorthodox source of data for inducing various perceptual, cognitive, and motor capabilities in machine learning models. The basic idea here is that neural activity contains a direct and detailed imprint of how the brain solves important perceptual, cognitive, and motor problems in the real world. The computational mechanisms and algorithmic details manifest in this imprint can be distilled into a machine learning model facing similar problems by training it to emulate the neural activity. This can be particularly valuable in domains where the solutions are difficult to verbalize precisely, such as learning a perceptual or motor skill.

Neural data for inducing alignment in machine learning models: Similar to the reasoning behind the previous use case, Mineault et al. (2024) recently argued that neural data and emulation of neural data could potentially be used to induce better alignment with human values, judgments, and preferences in machine learning models.