

A GENERALIST INTRACORTICAL MOTOR DECODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Mapping the relationship between neural activity and motor behavior is a central aim of sensorimotor neuroscience and neurotechnology. Most progress to this end has relied on restricting complexity: studying specific simple behaviors, in limited subjects, with interpretable computational models. However, current trends in deep learning suggest that modeling a breadth of neural and behavioral data may be both possible and beneficial. We accordingly developed Neural Data Transformer 3 (NDT3) as a foundation model for motor decoding of neural data from intracortical microelectrodes. We pretrained NDT3 with 2000 hours of neural population spiking activity paired with diverse motor covariates from over 30 monkeys and humans from 10 labs. Pretrained NDT3 is broadly useful, benefiting decoding on 8 downstream decoding tasks and generalizing to a variety of neural distribution shifts. However, we find signs that scaling over diverse neural datasets may be challenging, as scaling from 200 to 2000 hours already requires increasing model size to 350M parameters to avoid model degradation, and several downstream datasets scarcely benefit from either data or model scale. We provide two demonstrations that this scaling is at least partially limited by variability in input and output spaces across neural datasets, which pretraining alone may not resolve.

1 INTRODUCTION

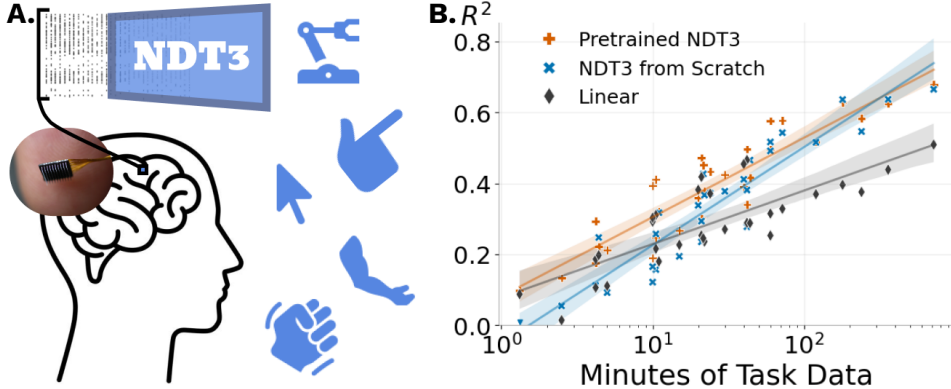


Figure 1. A. NDT3 is a deep network for decoding intracortical spiking activity into low-dimensional time series for various motor effectors¹. B. We measure decoding performance on downstream tasks with variable amounts of task-specific data. While from-scratch models only reliably outperform a linear baseline after 15 minutes of data, tuning a pretrained NDT3 provides consistently superior performance.

Intracortical neural data collection is growing rapidly. This growth comprises not only larger individual datasets with more neurons and higher behavioral complexity (Urai et al., 2022; Stevenson, 2023), but also an increase in the collective number of datasets. This growth marks a new regime of neural data science that may be best supported and exploited by new computational tooling capable of modeling a broad diversity of datasets. Large deep networks appear very suitable for this task, so much so as to justify terming large pretrained networks on broad domain data as foundation

¹Photo courtesy of REDACT and The Chicago Tribune.

models (Bommasani et al., 2022). Efforts to create foundation models are now proliferating beyond their origins in natural language processing (NLP) and computer vision (CV) into many domains of engineering and science (Wang et al., 2023a). Here, we propose a foundation model for motor decoding from intracortical spiking activity, which we call Neural Data Transformer 3 (NDT3).

Motor decoding is a valuable initial domain for characterizing neural data foundation models. Academic, clinical, and industrial efforts to create iBCIs for movement neuroprosthetics provide a path for scaling data collection from hundreds to thousands of subject-hours, and also fuel a need for pretrained models that generalize quickly and perform robustly for new users and settings. To this end, behavior prediction also provides more intuitive metrics for benchmarking progress than neural data prediction or the most abstract endpoint of providing scientific insight (e.g. with latent variable models or *in silico* models) (Pei et al., 2021; Wang et al., 2023b). Finally, recent work has shown that deep networks are able to transfer learn across motor cortical datasets collected at different timepoints, subjects, or tasks (Azabou et al., 2024; Ye et al., 2023; Schneider et al., 2023). These ingredients provide the means and motivation for scaling neural data modeling.

However, scaling may be complicated by the design and heterogeneity of contemporary neural datasets. Many motor cortical datasets are designed to probe specific hypotheses by limiting behavioral complexity and diversity. These behavioral limits can constrain the complexity of the observed neural data (Gao and Ganguli, 2015). Both constraints would seem to limit the benefits of scaling complex decoders. Beyond the limitations of individual datasets, each neural dataset inherently contains unique variability distinguishing them from others. This is most salient when comparing across the datasets we aggregate to enable pretraining, where different neurons are recorded in each subject and distinct output dimensions are required for each effector. To illustrate this, consider a 2-neuron toy setting, where one neuron fires on leftward motion and the other fires on rightward motion. No amount of scaling could reduce the data needed to determine which neuron corresponds to which direction.

To probe the value of scaled pretraining on heterogeneous spiking activity, we developed Neural Data Transformer 3 (NDT3). NDT3 uses simple neural and behavioral tokenization strategies to enable pretraining over diverse decoding datasets and fine-tuning to new tasks without introducing any new parameters (Fig. 1A). We pretrained NDT3 using up to 2000 hours of neural and behavioral data from motor neuroscience experiments with monkeys and clinical iBCI trials with humans. We then evaluated NDT3’s decoding performance on eight diverse motor tasks (Section 3.1) and find that tuning NDT3 yields models that either improve or match task-specific models trained from scratch, with prominent gains when task data is under 1.5 hours (Fig. 1B). Further, these gains persist under a number of distribution shifts Section 3.3. These performance gains in low-data regimes may enable both more complex experimental design and potentially decrease the burden of decoder training for people using iBCIs. However, we find that scaling pretraining data from 200 to 2K hours required raising model capacity to 350M parameters to mitigate a performance drop. We provide initial analyses of NDT3’s sensitivity to the specific inputs and outputs seen during fine-tuning to illustrate how we might understand this scaling behavior.

2 APPROACH

2.1 DATA

NDT3 models datasets of paired neural spiking activity and behavior (Fig. 2). Given our focus on motor decoding, most of the data comes from devices implanted in motor cortex of various monkeys and humans (Fig. 2A). These devices are intracortical multielectrode arrays or probes that record 30 kHz extracellular potentials. Spikes are extracted from these potentials, typically by bandpass-filtering the data between 300 and 3000 Hz, and marking a spike when the voltage signal crosses a preset threshold value. The neural data in our pretraining are diverse (Fig. 2B top). Data can have markedly different profiles across electrodes due to being from different electrode arrays in the same subject (left), have many silent channels (middle), or be densely active due to noise (right).

The typical behaviors in the pretraining data are different types of upper-limb reaching and grasping, nearly all from experimental paradigms that consist of short, repeated trials. While neural data were always recorded from microelectrodes, motor covariate signals came from various sensors. In monkey datasets, these sensors measure actual limb activity (e.g., Fig. 2A, left: limb kinematics from optical

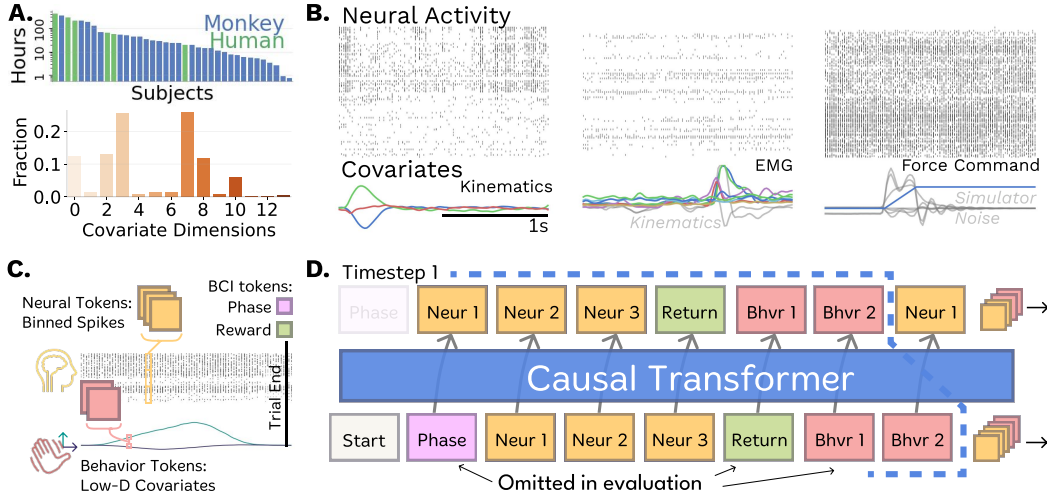


Figure 2. NDT3 Data and Model Design: **A.** NDT3 models paired neural spiking activity and behavioral covariate timeseries. We plot the distribution of 2000 hours of pretraining data volume by subjects (top) and covariate dimensionality (bottom). **B.** Examples of the neural and behavioral data for each of the three types of behavioral covariates in pretraining: kinematics, EMG (electromyography), or forces. Not all modeled dimensions in data are meaningfully task-related (right, grey behavior). **C.** Neural spiking activity is tokenized in time by binning the number of spikes every 20 ms, and in “space” using patches of channels (usually 32), as in NDT2 (Ye et al., 2023). Behavior is low-dimensional in our data, so we use 1 token per behavior dimension, also per 20 ms timestep. NDT3 also pretrains on data from BCI control, which we annotate with two additional tokens. The phase token indicates whether the user is controlling or observing the behavior and the reward token indicates if the BCI task was completed. **D.** NDT3 models tokens in a single flat stream with linear readins and readouts. Every real-world timestep (shown by the blue cutout) yields several tokens, which we order to allow causal decoding in evaluation. During evaluation, we omit non-neural tokens.

tracking, middle: electromyography (EMG)). In human datasets, physical movements are typically not possible, so the data’s behavior signals are programmatically generated. These signals are “paired” with the neural data in that they are cued or otherwise instructed to the person, who will attempt or imagine the corresponding behavior, such as grasping at a specified force level (Fig. 2B right). This force panel also shows that in pretraining, we cannot always automatically discern the primary task covariates (e.g., blue line, force, in the panel) from other recorded behavioral variables (grey). Thus, behavior data may include unpredictable variables. Finally, we also include closed loop iBCI data, where some behavior is generated by an iBCI decoder (not NDT3, see modeling strategy in Section 2.2).

The pretraining datasets are composed of archives from several experimental labs and some public datasets, and contain data from non-human primate neuroscience experiments and human clinical trials for neuroprosthetics. The grassroots nature of this aggregate dataset presents a heterogeneity in neural data processing, motor effectors, and experimental setup, most comparable to aggregate robotics datasets (OpenX et al., 2024). We detail the pretraining data composition in and provide references for the 100-so hours of public data in Section C.4.

We minimize preprocessing of these data to maximize the applicability of our generalist model. Kinematics signals are all converted to velocities, and all behavior (kinematics, EMG, force) is normalized per dataset such that the maximum absolute value of each variable is one. Data are cut into segments of two seconds - or concatenated to this length if the individual pieces of data (e.g. trials) are shorter - without additional annotation of data discontinuity. Two seconds was selected as it is roughly the timescale of behavior in our data (Fig. 2A). This segmenting is useful as it provides a more uniform format for forming training batches, which improves GPU memory utilization. Segments with no spikes or covariate variability are discarded. In total, this yielded about 3 million sequences, or 1750 hours, which we sample uniformly for pretraining. We round this to 2 khrs in subsequent text for simplicity.

2.2 MODEL

NDT3 is a causal Transformer with linear readin and readout layers for its various modalities, similar to GATO or TDMs (Reed et al., 2022; Schubert et al., 2023; Chameleon, 2024). For use with a Transformer, the data must be tokenized (Fig. 2C). We tokenize neural data by patching spike counts (Ye et al., 2023); each token is a flattened vector of the binned spikes in a chosen temporal resolution (20 ms) and spatial dimension (32 channels). For example, neural activity sampled from an electrode array with 100 channels would patch into $4 = \lceil 100/32 \rceil$ 32D neural tokens per 20 ms timestep. As the behavioral variables are already low-dimensional, we simply assign 1 token per dimension at the same temporal resolution. Finally, we add tokens marking whether the behavior are generated by a BCI system or by physical limb movement. While measured kinematics, EMG, or force will reflect a natural relationship with neural activity, behavioral data from BCI tasks are controlled by a program or learned decoder. We frame BCI-driven behavior as a suboptimal demonstration (Merel et al., 2016), and adopt a scheme inspired by Decision Transformers (Chen et al., 2021; Lee et al., 2022). In this scheme, we use a Phase token to track the timesteps where behavior is at least driven by neural activity and under decoder control, or only under programmatic, open loop control. We also use a Return token reflecting controller quality based on task completion. Note that these signals are only considered for pretraining, and are ablated entirely from the model at evaluation. Similarly, input behavior tokens are masked out in inference, so that the model input only indicates how many tokens must be predicted. NDT3 is trained with mean-squared error for prediction of behavioral variables, and categorical cross-entropy losses for prediction of neural spike count and reward.

All modalities are flattened into a single token stream, with the order of tokens in each real-world timestep respecting a canonical order required for control (Fig. 2D). As in GATO, individual tokens are annotated with learned position embeddings identifying token modality and sub-modality “position.” We additionally add a rotary embedding (Su et al., 2023) to track real-world timesteps.

Pretraining We pretrain NDT3 models over variable pretraining data and in sizes of 45M and 350M parameters to assess the impact of data and model scaling. Pretraining is early stopped according to validation loss or terminated at a maximum of 400 epochs. The 200 hour, 45M model trains for 480 A100-hours while the 2000 hour (2kh) 350M model takes 20K A100-hours. Fine-tuning maintains the pretraining objectives and updates all parameters. Data is segmented into 2 second intervals in pretraining, and mainly 1 second intervals in fine-tuning.

2.3 EVALUATION STRATEGY

Evaluation datasets and tuning Our main evaluation (Section 3.1) uses four human and four monkey datasets sampling varied upper limb movements, which we detail in Section C.4. Each dataset contains multiple sessions of data, typically from a single monkey or human. We will refer to each such setting as a “task,” distinguished from the behavioral procedure performed in each dataset. Each session has unique variability, so fine-tuning procedure may greatly impact decoding results. Prior work (Azabou et al., 2024; Ye et al., 2023; Zhang et al., 2024) ran focused evaluations by tuning and evaluating separately for each evaluation session. To manage compute and storage demands and to reflect that real world datasets are rarely collected or analyzed in isolation, we fine-tune NDT3 jointly over data combined from multiple evaluation sessions.

Downstream Hyperparameters All tuned and from-scratch NDT3 models use a search over 3 learning rates. This sweep is limited to make computational demands tractable, but also demonstrate the simple versatility of the base model. Importantly, the same search space is used for all tasks; we list the space and show its sufficiency relative to wider sweeps in (Section C.2). The best learning rate is chosen according to average validation performance across three random seeds, and we report their mean on the evaluation data.

Baselines We compare against Wiener filters (WF) and multi-session NDT2. WFs are a conventional linear method for both motor decoding and control in iBCI devices (Pandarinath and Bensmaia, 2022), and we implement them as ridge regression with multi-timestep history. NDT2 is a Transformer that uses MAE-style (He et al., 2021) self-supervision and has been demonstrated to improve with heterogeneous neural datasets, including the multi-session data we use it for here. Other Transformer variants have been proposed for pretraining on spiking data (Azabou et al., 2024; Zhang et al., 2024), but the field yet lacks consensus benchmarks to distinguish the most promising proposal to scale.

3 RESULTS

NDT3’s pretraining effort advances prior intracortical models from data volumes of 200 to 2000 hours and model sizes an order of tens to hundreds of millions of parameters. In Section 3.1, we show the increased data scale actually degrades aggregate downstream performance unless simultaneously increasing model scale. We propose that the performance drop from scaling data alone is due to high variability across intracortical motor decoding datasets. In Section 3.2, we illustrate this challenge by showing NDT3’s sensitivity to shifts in data input or output. Section 3.3 concludes by showing that despite this challenge for further data scaling, NDT3’s pretraining already provides gains that generalize to various novel settings, establishing NDT3 as a useful foundation for motor decoding.

3.1 MULTI-SCALE EVALUATION ACROSS MOTOR DECODING TASKS

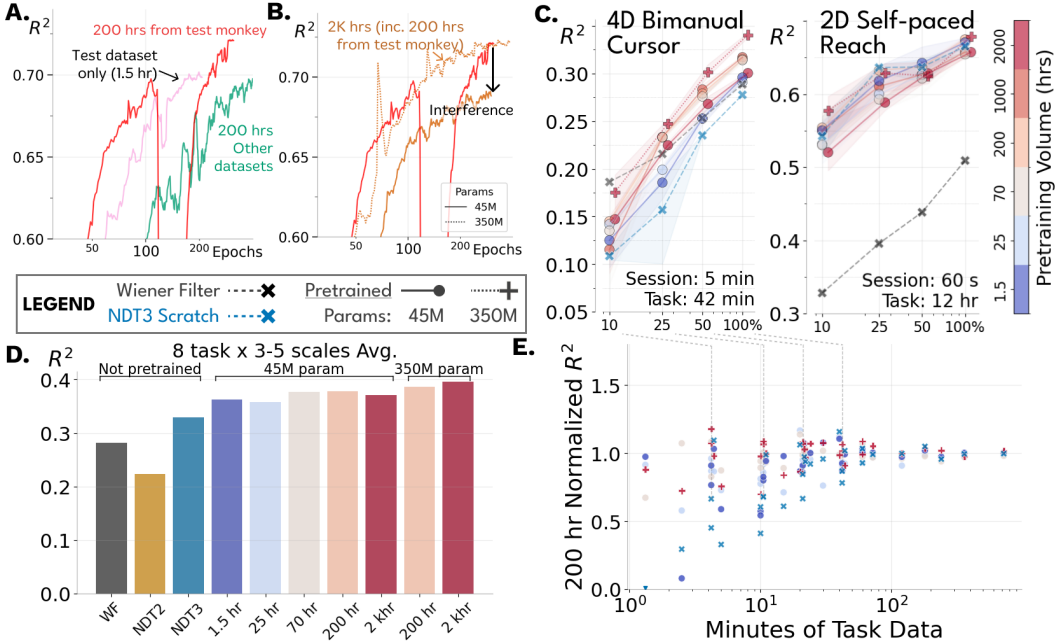


Figure 3. Evaluation on diverse motor tasks: **A.** Test-split pretraining R^2 compared for 3 models. All model pretraining data includes 1.5 hours of calibration data for the test dataset. We compare a model with just this data (Test dataset only) vs using 200 hours of additional data either from the test monkey or from over 10 other monkeys. Only the additional test monkey data improves over the calibration model. Models terminate at different points due to early stopping. **B.** Pretraining R^2 for models with up to 2000 hours (2K hrs) of pretraining data. The 2K hr model degrades in performance vs the 200 hr model at 45M parameters, and merely maintains performance at 350M parameters. **C.** Examples of good and bad data-scaling in downstream multiscale evaluation on two datasets. The x-axis scales the task and per-session data available by random subsampling of the full dataset. Shading shows standard deviation on 3 model seeds. The bottom right text shows total per-session and total time in each dataset, and the session time denotes the data in each evaluation session. 2 khr models are offset for visual clarity, but use the same amount of tuning data as other models. Increasing pretraining data yields performance gains at all downstream scales in the 4D task, but unclear effects in the self-paced reach task. **D.** Downstream performance averaged for all scales and datasets (31 settings in total) for different NDT3 and baselines. 45M models improve with data until 200 hrs and degrade at 2 khrs. Increasing model size to 350M parameter resolves interference. **E.** Per-task performance, normalized by the 350M 200 hr NDT3 performance, is shown against task time for different NDT3 models. Each vertical band shows models trained on some scaling of an evaluation dataset, e.g. dashed lines show the evaluations from the bimanual dataset. Model variability vanishes by 100 minutes of data. ▼ indicate outliers clipped for clarity.

To set expectations for how data scale and model size will impact model performance, we first examine pretraining curves computed on a test split. This test split contains multiple sessions of 2D reaching behavior mainly from one monkey. From the dataset this test split was drawn from, we sample an additional 1.5 hours of data and include this in the pretraining of all models to allow learning this specific test task. Fig. 3A shows the test performance over the course of pretraining in a model using just this 1.5 hours of “calibration” data, and two models using 200 hours of data. One

of these 200 hour models used data from over 10 other monkeys performing a variety of reaching tasks, but did not improve over the minimal 1.5 hour model on the test data. In contrast, using 200 hours more from the test monkey (from a separate set of experiments with similar behavior) achieved a small improvement in performance. Thus, only closely related data appears to improve a model that already sees sufficient task-specific data, in this case 1.5 hours. To further emphasize this sense of dataset distance, Fig. 3B shows that a 2 khr model that sees the same extra test data degrades in performance, indicating interference in learning of the test task. Pretraining literature has shown that increasing model size and dataset size in tandem is important for performance Dosovitskiy et al. (2021); Kolesnikov et al. (2020); Aghajanyan et al. (2023), and indeed increasing model size to 350M parameters remedied the drop. However, we still do not see gains on the test task.

This upstream saturation motivated a downstream evaluation conducted at multiple data scales. We illustrate this evaluation for two tasks in Fig. 3C. These two datasets are from a human performing open loop iBCI calibration for bimanual cursor use (Deo et al., 2024), and a monkey performing self-paced reach to random targets (O’Doherty et al., 2017). In both cases, the individuals are held-out from pretraining entirely, so the task-specific data is only seen in tuning. In the bimanual task, NDT3 performance does scale smoothly with pretraining performance at all downstream scales, up to the full 42 minutes of task data. The degradation in the 2000 hour model and subsequent rescue by increased model size is also replicated. The self-paced reach shows a much less clear result. For example, the from-scratch NDT3 achieves nearly the best performance at most data scales. The mixed effects here may result from the high data volume in this dataset, as the 10% scale still uses 1.2 hours of task data.

These two tasks reflect just two of several different downstream trends on the eight different evaluation datasets we study. We defer individual discussion to Section B.4, and next consider summary performance in Fig. 3D. To create this summary, we tuned over 2000 models in 31 evaluation settings, to identify overall benefit of pretraining scale in spite of per-evaluation variability Brown et al. (2020). As in (Entezari et al., 2023), we find that minimal pretraining (1.5 hrs) already achieves a large performance gain, but scaling data to 200 hours yields further improvements. However, as in upstream evaluation, effective use of 2000 hours of pretraining data requires raising model size to 350M parameters. We note this aggregate view obscures high variability in individual task trends, which we cover in Section B.4. In particular contrast with pretraining, we see the least improvement and even degradation in three of our evaluation tasks which use data from humans in the 2 khr pretraining data.

We finally return to the issue of downstream data scale. Recall that pretraining suggested that task performance saturates given 1.5 hours in the test task. Fig. 3E aims to generalize this observation by plotting all NDT3 evaluations against the downstream data size each evaluation tuned on. Model variability steadily decreases and is negligible by 100 minutes of data. This mark provides a heuristic for users to estimate whether their own decoding task will benefit from NDT3. We believe this threshold will be effectively irreducible by further scaling over heterogeneous pretraining data. This limit may be driven by each task’s unique variability rather than neural signal quality, as in most tasks we consider, such as in Fig. 3C, doubling task data reliably improves performance, more so than scaling pretraining data two orders of magnitude.

3.2 PRETRAINING DOES NOT OVERCOME INPUT-OUTPUT VARIABILITY

Section 3.1 showed pretraining may overall be limited by a fundamental variability in each neural dataset. This variability is evident as 350M parameters were necessary to scale to only 2 khrs of pretraining data, whereas in contrast, a 0.8B audio model consistently improves as pretraining data scales from 3K to 700K hours (Radford et al., 2022), and 3M parameter language models saturate but do not degrade when scaling up to 20B tokens (~ 20 M trials) (Kaplan et al., 2020). On the other hand, modeling decisions around architecture, hyperparameters, and post-training, may all significantly influence scaling. Decoupling the effects of these potential data-driven and model-driven factors will be vital for the future scaling of neural data foundation models. We next illustrate how these factors interact in NDT3’s sensitivity to the specific neural inputs and covariate outputs seen in tuning.

To further probe what distinguishes cross-session from cross-subject transfer, we apply structured ablations of input order, as in Neyshabur et al. (2020). We specifically create a shuffle that only alters the test session’s neural token order with respect to cross-session data, hypothesizing Transformers would be more robust to this shift than full channel shuffling. Indeed, all models benefit with a small amount of token-shuffled cross-session data (Fig. 4B center). If we instead apply a half-token shift to test-session inputs, performance is greatly harmed (Fig. 4B right). This shows cross-session transfer depends greatly on the specific token dimension semantics. This is so much the case that performance degrades to the point of from-scratch cross-session transfer (blue) matching the trend of cross-subject transfer (red) to this altered test-session. This suggests that NDT3 is no more able to transfer under shifts in intra-token semantics than changes to a wholly different subject.

Outputs seen in tuning restrict NDT3 predictions. The increased sample efficiency of pretrained models suggests that NDT3 could map a new individual’s neural activity to behavior without sampling the full range of neural-behavior data. For example, having seen many instances of radial reaches, NDT3 may generalize to unseen reach directions in a new subject better than non-pretrained decoders, which fail completely (Rizzoglio et al., 2022). Structured radial reaches are a particularly simple litmus test as their underlying neural activity is easily visualized in terms of a planar subspace (Churchland et al., 2012). We assess angular generalization in Fig. 4C, by evaluating reach decoding in an isometric, force-based (Monkey J) setting and a manipulandum-based (Monkey J) setting. Beyond the change in effectors, the neural data in these datasets also vary in their separability as visualized by linear discriminant analysis (LDA). We next train decoders on every pair of angles separated by 90 degrees (one shown) and plot predictions on held-out trials from all angles. The specific failure of both WFs and NDT3 is that their predictions do not extrapolate to held-out angles, consistent with Rizzoglio et al. (2022). While NDT3 produces more organized trajectories than the WF, held-out angles are constrained to the held-in conditions. Intriguingly, even the intermediate, interpolated angle appears to disappear for the less separable monkey V dataset. We quantify prediction performance in Fig. 4D. Notably, the two datasets achieve similar R^2 patterns despite significantly different LDA projections, with the only subtle difference being that from-scratch models are slightly worse for monkey V.

These results show that NDT3 fails to generalize to held-out directions, but also hint at the reason why. One appealing interpretation is that NDT3 is incentivized to sacrifice any angular generalization to maximize held-in performance, consistent with reports that RNNs exploit attractor structures to improve performance on noisy neural data (Costello et al., 2023). The fact that in the more challenging monkey V dataset, the interpolated angle between held-in angles appears degraded, and pretrained models slightly outperform from-scratch models, suggest that pretraining may even incentivize this tradeoff. Importantly, we show in Section B.1 that a constrained linear decoder class (more so than WFs) can yield held-out generalization. Thus while NDT3 fails to generalize to held-out directions, it remains unclear whether this implies that pretraining failed to learn a viable prior for radial reach (such as a low-dimensional linear constraint), or whether we have merely failed to elicit this structure (i.e. through post-training). Even in the easiest setting of planar reaches, it can be challenging to directly probe whether the model has learned what we expect it to.

Our manipulations on model input and output highlight the difficulty of pinpointing whether scaling is limited by data-based or model-based factors. They highlight, however, basic forms of variability across neural datasets that NDT3’s generic pretraining fails to bridge.

3.3 WHERE DOES PRETRAINING HELP?

Despite challenges for scaling, NDT3’s pretraining endows a prior extracted from hundreds of hours of neural data. We conclude by showing how this prior generalizes beyond shifts in behavioral task.

Neural distribution shifts Neural data is nonstationary due to factors ranging from speculatively characterized to fully experimentally controlled. As examples, we examine shifts in time, arm posture (Marino et al., 2024), and spring load (Mender et al., 2023) (Fig. 5A top), and evaluate pretraining robustness under these shifts (bottom). Specifically, we measure the performance of models prepared in the same setting they were trained on (in-distribution, ID) vs. the shifted data (out-of-distribution, OOD). In all cases, the gains conferred by pretraining ID persist OOD. Note that the pose shift has the most significant change in firing rates but virtually no drop in performance. Decoder robustness in this case likely benefits from the fact that the neural variance associated with

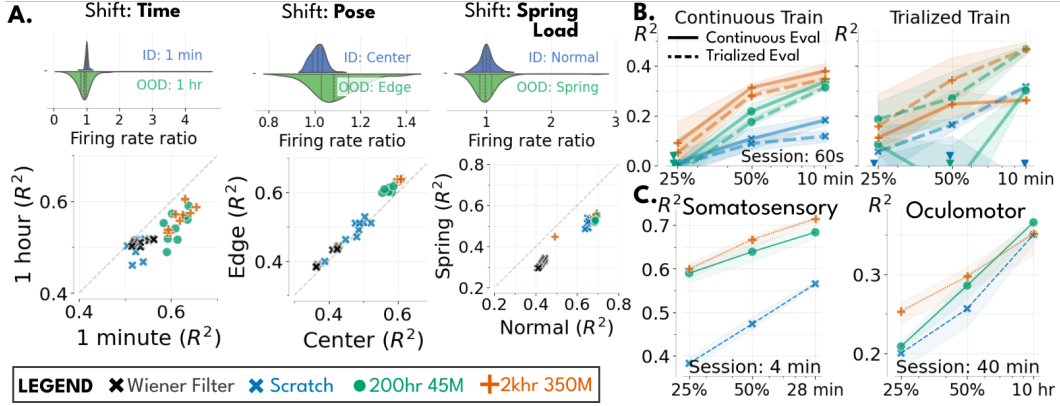


Figure 5. Enumerating where pretraining does generalize. **A.** Models fine-tuned in one distribution of data are evaluated in-distribution (ID) and out-of-distribution (OOD). Top plots show the distribution across channels of neural firing rates from OOD and ID trials, normalized by average ID firing rates. Lower plots scatter OOD vs ID performance, with each point being a single model with different hyperparameters. The **time** shift uses two human cursor datasets collected one hour apart. Models were tuned in each block and were evaluated in the second block. **Pose** shift uses a monkey center-out reach task which was performed with the hand starting in different locations in the workspace. **Spring Load** uses a dataset of monkey 1D finger motion with or without spring force feedback. **B.** Models are evaluated on a human open-loop cursor dataset prepared in two ways. Trialized training receives inputs according to trial boundaries, varying from 2-4 seconds in length. Continuous training receives random 1 second snippets (that can cross trial boundaries). Trialized evaluation matches trialized training, and continuous evaluation is done by streaming up to 1 second of history. ▼ indicates points below 0.0. Continuously trained models perform well in both evaluation settings, while models trained on trialized data fail in continuous evaluation. **C.** Multiscale fine-tuning performance of NDT3 on datasets recorded outside motor cortex, namely S1 (Somatosensory) and FEF/MT (Oculomotor).

pose is effectively orthogonal to neural variance for reach direction. Overall, while neural data present many nonstationarities, these examples provide some assurance that pretraining benefits are not dependent on brittle features specific to the precise training dataset.

Trial structure Non-pretrained DNNs tend to overfit trial structure in datasets, for example by learning what timesteps a trial begins and ends. This disrupts continuous control in iBCI use (Deo et al., 2024; Costello et al., 2023). In Fig. 5C, we assess how pretraining impacts such overfit to trial structure in an open loop human cursor control dataset. Though best assessed through control, we can also emulate probe trial structure overfitting by streaming neural inputs across trials (Continuous eval). Alternatively, we can provide full trials of data to the model (Trialized eval). We prepare models trained with trialized inputs (Trialized train), or with random 1 second intervals of data (Continuous Train), since the streaming is done with up to one second of history. That trialized evaluation of trialized trained models outperforms the best continuous analog reflects that all models can learn to exploit trial structure. However, while trialized from-scratch models become subtrivial under continuous evaluation (solid blue line is off-panel), degradation is more graceful for the pretrained models. Comparing across panels, continuous performance of pretrained models under trialized training is comparable to continuous training, though not fully. Note that the small drop in performance from continuous to trialized evaluation under continuous training comes from the extra history continuous evaluation provides at the onset of each trial. Reduced dependence on trial structure should benefit both analysis and control.

New brain areas In Fig. 5D, we return to multiscale fine-tuning to test how NDT3, pretrained on motor cortex, performs in somatosensory cortex (S1) and oculomotor areas (FEF and MT). The gain from pretraining over from-scratch models is high in S1, but also nontrivial in the Oculomotor dataset. While the former can be attributed to the close interaction of sensorimotor areas, the latter implies NDT3 has learned a broader prior. This prior could be neurophysiological (e.g. declining subject focus over time (Steinmetz et al., 2019)), or it could be a common experimental artifact, like trial structure. For example, this Oculomotor dataset contains 4 behavioral conditions, which may benefit from the tendency to learn classifiers shown in Fig. 4C rather than a prior on neural dynamics.

4 DISCUSSION

Many fields are now pursuing large scale deep learning as "a tide that lifts all boats" (Abnar et al., 2021), with the hope that improvements on the single abstract goal of effective pretraining will yield field-wide, downstream improvements. Such a unifying abstraction may be timely for neuroscience, given the increasing volume, diversity, and complexity of modern neural data. Joining other pretraining efforts on varied modalities of neural data (Section A), we trained NDT3 on 2000 hours of paired neural population activity from motor cortex and behavior, and then conducted a broad downstream decoding evaluation. Consistent with the broad foundation modeling narrative, we found the best aggregate performance from increasing data scale and model size jointly. However, these benefits from pretraining vary with the downstream dataset, with several datasets having minimal improvements from scale (Section B.4). This result may stem in various ways from our approach: insufficient hyperparameter sweeps, our focus on decoding, and generic architecture design. However, the fact NDT3 needs increased capacity to pretrain on 2000 hours and that NDT3 generalizes poorly to input and output shifts highlights potential challenges rooted in neural dataset variability.

More broadly, we advocate for further consideration of how neural data can contribute to and gain from the ongoing cross-disciplinary conversation on foundation modeling. For example, our input and output sensitivity analyses were inspired by ML (Neyshabur et al., 2020; Pham et al., 2021) and neuroscientific literature (Gallego et al., 2020; Sadtler et al., 2014), respectively. Challenges to scale in neural data could deeply resemble interference in multimodal models (Aghajanyan et al., 2023; Liu et al., 2024). Inversely, neural distribution shifts have the advantage of being carefully characterized, and so the appearance of correlated ID-OOD performance, as also appears in CV, NLP, and other AI domains (Taori et al., 2020), may refine our understanding of when such correlation will occur, and thus when foundation models will be effective. Our hypothesized challenge of sensor variability should be particularly interesting to compare across the biosignals community, which must overcome analogous variability to achieve our shared goal of achieving user-general models.

4.1 ETHICS STATEMENT

The animal datasets used in this work were collected for other studies that were approved by Institutional Animal Care and Use Committees. Human datasets were also collected for other studies, with Institutional Review Board approval and as part of clinical trials conducted under FDA Investigational Device Exemptions. Informed consent was obtained prior to any experimental procedures. We discuss the potential for NDT3 to reduce user burden for iBCI-based neuroprosthetics, though the dissemination of pretrained models on these data raise the risk that the original human data may be recoverable from model weights. Since this seems technically challenging at this point, and since the source data are restricted to binned spiking activity to begin with, we deem the risk low enough to justify the potential scientific benefit of sharing our pretrained models.

4.2 REPRODUCIBILITY STATEMENT

Advancing neural data foundation modeling will require a flourishing open-source ecosystem, including data, models, and evaluations. While we will release our models and codebase, our work currently has limited reproducibility given our inability to release pretraining data. Similarly, we have tried to use open evaluations where possible, but several evaluation datasets remain private. We expect that field-wide trends toward open data releases, and larger scale academic (Koch et al., 2022) or academic-industrial collaborations, can alleviate this limitation in the near future.

REFERENCES

- Anne E. Urai, Brent Doiron, Andrew M. Leifer, and Anne K. Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 25:11–19, 2022. doi: 10.1038/s41593-021-00980-9.
- Ian H. Stevenson. Tracking advances in neural recording. Statistical Neuroscience Lab, University of Connecticut, 2023. URL <https://stevenson.lab.uconn.edu/scaling/>. Accessed September 6, 2024.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas

- Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kudithipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avani Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL <https://arxiv.org/abs/2108.07258>.
- Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023a.
- Felix C Pei, Joel Ye, David M. Zoltowski, Anqi Wu, Rameed Hasan Chowdhury, Hansem Sohn, Joseph E O’Doherty, Krishna V. Shenoy, Matthew Kaufman, Mark M Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan W. Pillow, Il Memming Park, Eva L Dyer, and Chethan Pandarinath. Neural latents benchmark ‘21: Evaluating latent variable models of neural population activity. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=KVMS3f14Rsv>.
- Eric Y Wang, Paul G Fahey, Kayla Ponder, Zhuokun Ding, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, et al. Towards a foundation model of the mouse visual cortex. *bioRxiv*, 2023b.
- Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36, 2024.
- Joel Ye, Jennifer L Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=CBbtMn1TGq>.
- Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL <https://doi.org/10.1038/s41586-023-06031-6>.
- Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.
- Collaboration OpenX, Abby O’Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Boother, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishk Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis,

- Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Llerel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minh Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Bajjal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Halder, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2024. URL <https://arxiv.org/abs/2310.08864>.
- Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yuri Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent, 2022. URL <https://arxiv.org/abs/2205.06175>.
- Ingmar Schubert, Jingwei Zhang, Jake Bruce, Sarah Bechtle, Emilio Parisotto, Martin Riedmiller, Jost Tobias Springenberg, Arunkumar Byravan, Leonard Hasenclever, and Nicolas Heess. A generalist dynamics model for control, 2023. URL <https://arxiv.org/abs/2305.10912>.
- Team Chameleon. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- Josh Merel, David Carlson, Liam Paninski, and John P Cunningham. Neuroprosthetic decoder training as imitation learning. *PLoS computational biology*, 12(5):e1004948, 2016.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021. URL <https://arxiv.org/abs/2106.01345>.
- Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Winnie Xu, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, and Igor Mordatch. Multi-game decision transformers, 2022. URL <https://arxiv.org/abs/2205.15241>.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL <https://arxiv.org/abs/2104.09864>.
- Yizi Zhang, Yanchen Wang, Donato Jimenez-Beneto, Zixuan Wang, Mehdi Azabou, Blake Richards, Olivier Winter, The International Brain Laboratory, Eva Dyer, Liam Paninski, et al. Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution. *arXiv preprint arXiv:2407.14668*, 2024.
- Chethan Pandarinath and Sliman J Bensmaia. The science and engineering behind sensitized brain-controlled bionic hands. *Physiological Reviews*, 102(2):551–604, 2022.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL <https://arxiv.org/abs/2111.06377>.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2020. URL <https://arxiv.org/abs/1912.11370>.

- Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023.
- Darrel R Deo, Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. Brain control of bimanual movement enabled by recurrent neural networks. *Scientific Reports*, 14(1): 1598, 2024.
- Joseph E. O’Doherty, Mariana M. B. Cardoso, Joseph G. Makin, and Philip N. Sabes. Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology, May 2017. URL <https://doi.org/10.5281/zenodo.788569>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Rahim Entezari, Mitchell Wortsman, Olga Saukh, M Moein Shariatnia, Hanie Sedghi, and Ludwig Schmidt. The role of pre-training data in transfer learning. *arXiv preprint arXiv:2302.13602*, 2023.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL <https://arxiv.org/abs/2212.04356>.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.
- Fabio Rizzoglio, Ege Altan, Xuan Ma, Kevin L. Bodkin, Brian M. Dekleva, Sara A. Solla, Ann Kennedy, and Lee E. Miller. Monkey-to-human transfer of brain-computer interface decoders. *bioRxiv*, 2022. doi: 10.1101/2022.11.12.515040. URL <https://www.biorxiv.org/content/early/2022/11/13/2022.11.12.515040>.
- Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.
- Joseph T Costello, Hisham Temmar, Luis H Cubillos, Matthew J Mender, Dylan M Wallace, Matthew S Willsey, Parag G Patil, and Cynthia A Chestek. Balancing memorization and generalization in rnns for high performance brain-machine interfaces. *bioRxiv*, pages 2023–05, 2023.
- Patrick J Marino, Lindsay Bahureksa, Carmen Fernández Fisac, Emily R Oby, Adam L Smoulder, Asma Motiwala, Alan D Degenhart, Erinn M Grigsby, Wilsaan M Joiner, Steven M Chase, et al. A posture subspace in primary motor cortex. *bioRxiv*, pages 2024–08, 2024.
- Matthew J Mender, Samuel R Nason-Tomaszewski, Hisham Temmar, Joseph T Costello, Dylan M Wallace, Matthew S Willsey, Nishant Ganesh Kumar, Theodore A Kung, Parag Patil, and Cynthia A Chestek. The impact of task context on predicting finger movements in a brain-machine interface. *eLife*, 12:e82598, jun 2023. ISSN 2050-084X. doi: 10.7554/eLife.82598. URL <https://doi.org/10.7554/eLife.82598>.
- Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.
- Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training, 2021. URL <https://arxiv.org/abs/2110.02095>.
- Thang M. Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks?, 2021. URL <https://arxiv.org/abs/2012.15180>.
- Juan A. Gallego, Matthew G. Perich, Rameed H. Chowdhury, Sara A. Solla, and Lee E. Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23(2):260–270, Feb 2020. ISSN 1546-1726. doi: 10.1038/s41593-019-0555-4. URL <https://www.nature.com/articles/s41593-019-0555-4>.
- Patrick T Sadtler, Kristin M Quick, Matthew D Golub, Steven M Chase, Stephen I Ryu, Elizabeth C Tyler-Kabara, Byron M Yu, and Aaron P Batista. Neural constraints on learning. *Nature*, 512(7515):423–426, 2014.

- 702 Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need of a modeling language for
703 distribution shifts: Illustrations on tabular datasets, 2024. URL <https://arxiv.org/abs/2307.05284>.
704
- 705 Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring
706 robustness to natural distribution shifts in image classification, 2020. URL <https://arxiv.org/abs/2007.00644>.
707
- 708 Christof Koch, Karel Svoboda, Amy Bernard, Michele A Basso, Anne K Churchland, Adrienne L Fairhall,
709 Peter A Groblewski, Jérôme A Lecoq, Zachary F Mainen, Mackenzie W Mathis, et al. Next-generation brain
710 observatories. *Neuron*, 110(22):3661–3666, 2022.
- 711 Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with
712 tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.
713 URL <https://openreview.net/forum?id=QzTpTRVtrP>.
- 714 Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal
715 foundation model for clinical applications, 2024. URL <https://arxiv.org/abs/2402.10251>.
716
- 717 Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. Biot: Cross-data biosignal learning in the wild, 2023.
718 URL <https://arxiv.org/abs/2305.10351>.
- 719 Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou.
720 Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals,
721 2024. URL <https://arxiv.org/abs/2405.17766>.
- 722 Armin W. Thomas, Christopher Ré, and Russell A. Poldrack. Self-supervised learning of brain dynamics from
723 broad neuroimaging data, 2023.
724
- 725 Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed A Rizvi, Matteo Rosati, Christopher Averill,
726 James L Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, and David van
727 Dijk. BrainLM: A foundation model for brain activity recordings. In *The Twelfth International Conference*
728 *on Learning Representations*, 2024. URL <https://openreview.net/forum?id=RwI7ZEFr27>.
- 729 Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases,
730 and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The*
731 *Eleventh International Conference on Learning Representations*, 2023c. URL [https://openreview.net/](https://openreview.net/forum?id=xmcYx_reUn6)
732 [forum?id=xmcYx_reUn6](https://openreview.net/forum?id=xmcYx_reUn6).
- 733 Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong
734 Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of
735 intracranial activity, 2024. URL <https://arxiv.org/abs/2406.03044>.
- 736 Sabera J Talukder, Jennifer J. Sun, Matthew K Leonard, Bingni W Brunton, and Yisong Yue. Deep neural
737 imputation: A framework for recovering incomplete brain recordings. In *NeurIPS 2022 Workshop on Learning*
738 *from Time Series for Health*, 2022. URL <https://openreview.net/forum?id=c9qFg8UrIcn>.
- 739 Samuel M Peterson, Shiva H Singh, Benjamin Dichter, Kelvin Tan, Craig DiBartolomeo, Devapratim Theogara-
740 jan, Peter Fisher, and Josef Parvizi. Ajile12: Long-term naturalistic human intracranial neural recordings
741 and pose. *Scientific Data*, 9(1):184, 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01280-y. URL
742 <https://doi.org/10.1038/s41597-022-01280-y>.
- 743 Adrien Doerig, Rowan Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace Lindsay, Konrad
744 Kording, Talia Konkle, Marcel A. J. Van Gerven, Nikolaus Kriegeskorte, and Tim C. Kietzmann. The
745 neuroconnectionist research programme, 2022. URL <https://arxiv.org/abs/2209.03718>.
- 746 Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fmri,
747 2024. URL <https://arxiv.org/abs/2305.11863>.
- 748 Tyler Benster, Guy Wilson, Reshef Elisha, Francis R Willett, and Shaul Druckmann. A cross-modal approach to
749 silent speech with llm-enhanced recognition. *arXiv preprint arXiv:2403.05583*, 2024.
- 750 Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding.
751 In *European Conference on Computer Vision (ECCV)*, 2024.
- 752 Quilee Simeon, Leandro Venâncio, Michael A Skuhersky, Aran Nayeibi, Edward S Boyden, and Guangyu Robert
753 Yang. Scaling properties for artificial neural network models of a small nervous system. In *SoutheastCon*
754 *2024*, pages 516–524. IEEE, 2024.

- Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data, 2024. URL <https://arxiv.org/abs/2407.07595>.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM, January 2019. doi: 10.1145/3287560.3287596. URL <http://dx.doi.org/10.1145/3287560.3287596>.
- Edward H Adelson, James R Bergen, et al. *The plenoptic function and the elements of early vision*, volume 2. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of ..., 1991.
- Andrea Colins Rodriguez, Matt G Perich, Lee E Miller, and Mark D Humphries. Motor cortex latent dynamics encode spatial and temporal arm movement parameters independently. *Journal of Neuroscience*, 44(35), 2024.
- Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction, 2024. URL <https://arxiv.org/abs/2403.06963>.
- Colin Lea, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation, 2016. URL <https://arxiv.org/abs/1608.08242>.
- Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL <https://arxiv.org/abs/2310.10688>.
- Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15, 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.653659. URL <https://www.frontiersin.org/articles/10.3389/fnhum.2021.653659>.
- Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective work best for zero-shot generalization?, 2022. URL <https://arxiv.org/abs/2204.05832>.
- Nur Muhammad Mahi Shafiullah, Zichen Jeff Cui, Ariuntuya Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone, 2022. URL <https://arxiv.org/abs/2206.11251>.
- Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, and Rishabh Agarwal. Stop regressing: Training value functions via classification for scalable deep rl, 2024. URL <https://arxiv.org/abs/2403.03950>.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL <https://arxiv.org/abs/2104.14294>.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.
- Brianna M Karpowicz, Joel Ye, Chaofei Fan, Pablo Tostado-Marcos, Fabio Rizzoglio, Clay Washington, Thiago Scodeler, Diogo de Lucena, Samuel R Nason-Tomaszewski, Matthew J Mender, Xuan Ma, Ezequiel Matias Arneodo, Leigh R Hochberg, Cynthia A Chestek, Jaimie M Henderson, Timothy Q Gentner, Vikash Gilja, Lee E Miller, Adam G Rouse, Robert A Gaunt, Jennifer L Collinger, and Chethan Pandarinath. Few-shot algorithms for consistent neural decoding (falcon) benchmark. *bioRxiv*, 2024. doi: 10.1101/2024.09.15.613126. URL <https://www.biorxiv.org/content/early/2024/09/16/2024.09.15.613126>.
- Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03506-2. URL <https://www.nature.com/articles/s41586-021-03506-2>.
- Jonas Geiping and Tom Goldstein. Cramming: Training a language model on a single gpu in one day, 2022. URL <https://arxiv.org/abs/2212.14034>.
- Robert D Flint, Eric W Lindberg, Luke R Jordan, Lee E Miller, and Marc W Slutzky. Accurate decoding of reaching movements from field potentials in the absence of spikes. *Journal of neural engineering*, 9(4): 046006, 2012.

- Chethan Pandarinath, Paul Nuyujukian, Christine H Blabe, Brittany L Sorice, Jad Saab, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson. High performance communication by people with paralysis using an intracortical brain-computer interface. *elife*, 6:e18554, 2017.
- Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain-computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.
- Beata Jarosiewicz, Anish A Sarma, Daniel Bacher, Nicolas Y Masse, John D Simeral, Brittany Sorice, Erin M Oakley, Christine Blabe, Chethan Pandarinath, Vikash Gilja, et al. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. *Science translational medicine*, 7(313): 313ra179–313ra179, 2015.
- Mariana P. Branco, Lisanne M. De Boer, Nick F. Ramsey, and Mariska J. Vansteensel. Encoding of kinetic and kinematic movement parameters in the sensorimotor cortex: A brain-computer interface perspective. *European Journal of Neuroscience*, 50(5):2755–2772, September 2019. ISSN 0953-816X, 1460-9568. doi: 10.1111/ejn.14342. URL <https://onlinelibrary.wiley.com/doi/10.1111/ejn.14342>.
- Kristin M Quick, Jessica L Mischel, Patrick J Loughlin, and Aaron P Batista. The critical stability task: quantifying sensory-motor control during ongoing movement in nonhuman primates. *Journal of Neurophysiology*, 120(5):2164–2181, 2018.
- Xuan Ma, Fabio Rizzoglio, Eric J. Perreault, Lee E. Miller, and Ann Kennedy. Using adversarial networks to extend brain computer interface decoding accuracy over time. Aug 2022. doi: 10.1101/2022.08.26.504777. URL <https://www.biorxiv.org/content/10.1101/2022.08.26.504777v1>.
- Kendra K Noneman and J. Patrick Mayo. Gaze decoding with sensory and motor cortical activity. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, page 1–3, Glasgow United Kingdom, June 2024. ACM. ISBN 9798400706073. doi: 10.1145/3649902.3655655. URL <https://dl.acm.org/doi/10.1145/3649902.3655655>.
- Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy H Wilson, Leigh Hochberg, Krishna V. Shenoy, Jaimie M. Henderson, and Francis R Willett. Plug-and-play stability for intracortical brain-computer interfaces: A one-year demonstration of seamless brain-to-text communication. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=STqMqhtDi>.
- Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL <https://arxiv.org/abs/2307.08691>.
- Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. URL <https://arxiv.org/abs/2302.05442>.
- Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie E Everett, Alexander A Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=d8w0pmvXbZ>.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories, 2021. URL <https://arxiv.org/abs/2012.14913>.

A RELATED WORKS AND A PROPOSED TAXONOMY

Neural data is sufficiently diverse so as to support many distinct efforts to train large neural data models. The scale of pretraining is somewhat larger in the non-implanted modalities, where data is more abundant. The largest EEG models have reached a scale of 2.5K (Jiang et al., 2024) to 40K hours of data (Yuan et al., 2024), or higher volumes if also considering non-brain biosignals (EKG) (Yang et al., 2023; Thapa et al., 2024). Current fMRI models operate in the 1K (Thomas et al., 2023) to 7K (Caro et al., 2024) hour range. The largest models in these studies are in the 0.1B-1B parameter range. Intracranial modalities, including sEEG (Wang et al., 2023c; Chau et al., 2024), ECoG (Talukder et al., 2022; Peterson et al., 2022), and spiking activity (Wang et al., 2023b), have thus far been studied at an order of magnitude smaller scales of data and model size (20-1000 hours, <0.1B parameters).

Direct scaling on neural data modeling should be distinguished from NeuroAI efforts (Doerig et al., 2022) to measure how models of the human sensorimotor experience (e.g. language, vision, audio models) predict neural data (Antonello et al., 2024). However, as multimodal efforts begin to blur this distinction (Benster et al., 2024; Xia et al., 2024), care will be required to distinguish advances in modeling neural data, embodied data, or their interaction.

Comparing neural data models Current efforts to understand scaling in neural data Simeon et al. (2024); Sato et al. (2024) will have their reach limited by the specificity of every neural dataset. A meta-challenge for the field is understanding how different parameters (species, brain area, modality, task) impact scaling properties. This would be greatly aided by development of reporting practices for different neural data models. To facilitate comparison, we create a model card (Mitchell et al., 2019) for NDT3 in Section D. In addition to the standard model card, we propose reporting an additional taxonomy to aid comparisons across neural data models, using two concepts.

First: neural data models can be conceptualized as modeling slices of the *plenneural function*, inspired by the plenoptic function in vision (Adelson et al., 1991). The plenoptic function is a model of an idealized eye which parameterizes all possible images with 7 dimensions: 4D to describe the global spacetime of the view, 2D to describe viewing angle (spherical) or coordinate (Cartesian) of the image, and 1D for wavelength. Since neural data models are primarily interested in circumscribed systems rather than the physical world, a similar global coordinate system (e.g. 4D for all possible electric potentials) would be uninformative. We thus propose reporting more qualitative coordinates:

1. Identity: The network or individual being recorded.
2. Task: The behavior, stimuli, or other activity the network is reflecting.
3. Spacetime: Coordinates specified in a network-local coordinate frame (e.g. brain area).

Second: The modeled extent of this plenneural function is conveniently discretized in three resolutions in a Transformer-like sequence modeling framework: the token, the sequence, and the full training data. The token is the most granular unit of data being modeled; NDT3 models neural populations 32 neurons at a time, in 20ms bins. At the sequence input level, NDT3 models inputs from single humans or monkeys, across 128-256 neurons in 2 second snippets, while performing effectively one “movement.” Finally, NDT3’s pretraining spans dozens of individuals, records motor and premotor areas over 2.5K hours, over a variety of arm and hand movements.

B SUPPLEMENTARY RESULTS

B.1 EXTRAPOLATION IN CENTER-OUT DECODING IS POSSIBLE WITH RESTRICTED LINEAR DECODERS.

In Section 3.2, we proposed that pretrained models like NDT3 should be able to generalize to held-out, and in particular, extrapolated reach angles. We hypothesized this due to the frequent appearance of 2D linear projections of high dimensional neural activity showing clear separation by reach angle (e.g. (Rodriguez et al., 2024)), which would imply NDT3 meta-learning of an explicit planar prior would enable our desired generalization. Here we make this intuition explicit, by constructing a linear decoder that generalizes to held-out reach directions, and thus illustrate that NDT3’s failure to extrapolate is not due to an inherently unconstrained generalization task but is rather due to a

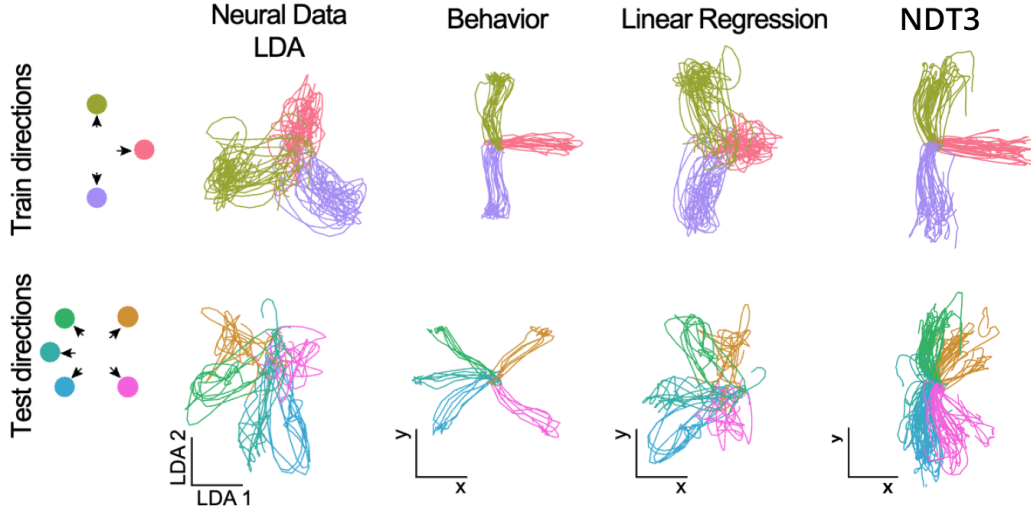


Figure 6. We illustrate how explicitly restricting to a linear decoder plane allow reconstruction of held-out behavior (Linear Regression), in contrast with NDT3 predictions which are restricted to held-in conditions even when shown three cardinal directions.

lack of proper objective (i.e. requires post-training to make generalization desirable) or a failure of pretraining.

To do this, we follow a classic neural data analysis procedure on a single session of the monkey J dataset (as in Fig. 4). As a reminder, the behavior here is an isometric task where the monkey generates isometric forces against a small box equipped with a six-degree-of-freedom load cell (JR3 Inc., CA). The forces were linearly mapped to control cursor movement: wrist flexion/extension moved the cursor left/right, and radial/ulnar deviation moved it up/down. The monkey had to move the cursor from a central position toward one of eight peripheral targets, in a classic center-out task. We extracted successful trials for each target from 0.5 seconds before to 1 second after movement onset.

With this preparation, we can now project the high-dimensional (96) neural activity at each timestep onto a candidate plane that reflects reach related variance, after which a simple rotation would allow generalized decoding. If we were fitting neural activity from all conditions, the top PCs from Principal Components Analysis (PCA) would typically be sufficient to identify this plane. Since we would like to find our general decoding plane without fitting all conditions, PCA alone is not reliable enough to extract the plane we desire. We thus first fit PCA to the three held-in conditions, and then used Linear Discriminant Analysis (LDA) on the top 10 PCs to find the 2D plane that best separated the three directions. This yields a plane where neural activity is well separated by their reach direction in a consistent manner for train and test directions (Fig. 6 Neural Data and Behavior). This allows a ridge regression to generalize to from held-in trajectories (variance accounted for (VAF): 0.70 ± 0.0) to both the interpolated (VAF: 0.48 ± 0.02) and extrapolated held-out (VAF: 0.44 ± 0.05) directions (Fig. 6 Linear Regression). In contrast, NDT3 trajectories are, as before, clearly constrained to held-in directions.

B.2 ABLATIONS

We ablate the major design decisions made to enable NDT3’s large scale pretraining. These ablations give us confidence that NDT3 overcomes the basic challenges we encountered in development, but compute restrictions prevent more exhaustive comparisons or exploration of model design space. We encourage further work exploring the influence of different hyperparameters. In these plots, we distinguish validation split performance and evaluation split performance, which is computed by batch-mode prediction (not the costly streaming evaluation used throughout main experiments).

Covariate dropout We find the default next-step prediction objective fails for learning decoding of highly autocorrelated covariate timeseries, perhaps because simply relying on teacher-forced

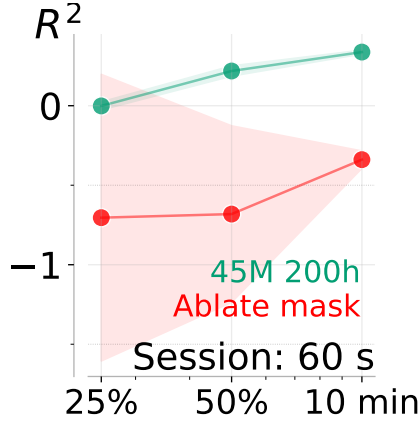


Figure 7. Ablation of covariate masking on an open 2D Cursor + Click dataset. Covariate inputs are completely masked in inference for the default NDT3, and autoregressively generated in the ablation.

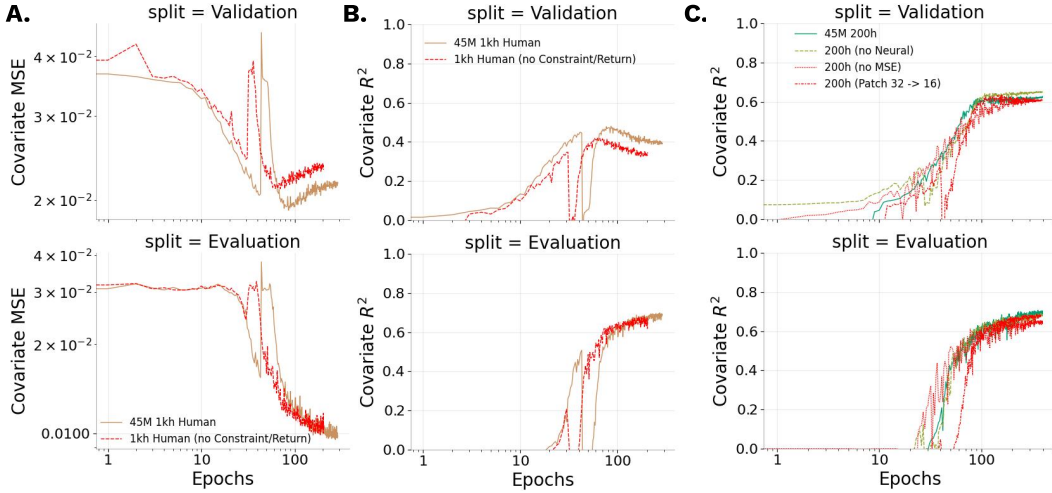


Figure 8. Ablations evaluated based on upstream evaluation split. **A., B.** Ablation of BCI control tokens **C.** Ablation of neural objective and covariate MSE objective in favor of classification over quantized covariates.

behavioral inputs provides a severe shortcut that prevents learning of a proper neural to behavior decoding map (Bachmann and Nagarajan, 2024). Different time-series models have addressed this by adopting convolutional input-output layers (Lea et al., 2016), tokenizing along temporal dimensions (Das et al., 2024), or learning with contrastive objectives (Chau et al., 2024; Kostas et al., 2021). We avoid introducing architectural modifications and instead adopt a simple dropout procedure that masks a portion of covariate inputs some fraction of the time. Specifically, on every training batch, two random numbers are drawn. The first, $M \sim U[0, 1]$, determines what fraction of covariate inputs should be masked. On 90% of batches, we also sample $T \sim U[0, 2]$ seconds, such that the mask is only applied after timestep T . That is, on 90% of batches, the model is provided a prefix-prompt. We do not block losses on this prefix as in prefix-LMs (Wang et al., 2022). Pretraining metrics for validation and evaluation are always computed with a prefix and full masking of non-prefix timesteps. In Fig. 7, we ablate covariate masking (which also removes the prefix logic), and tune on a 2D Cursor + Click task. The ablated model performs subtrivially with student-forced predictions provided as input at test time. Note that the ablated model performs trivially with masked inputs (not shown).

BCI-phase and return conditioning NDT3’s pretraining includes several hundred hours of BCI control data, where the covariates were set by another decoder. We introduced phase and return conditioning tokens to differentiate the several types of BCI control data from recorded behavior. Specifically, in BCI data, NDT3 receives input tokens specifying what fraction of the behavior reflects

neural input (BCI control is on) vs programmatic input (BCI control is off, as in open loop BCI calibration). Further, we provide inputs encoding reward (trial success) when trials change, and return (future reward over a 10 second horizon, which crosses data boundaries). This design is intended to evaluate the potential for a Decision-Transformer like offline learning strategy for improved online control, but we do not discuss this in this work. In Fig. 8A, B, we focus on whether these inputs improves pretraining loss and R^2 in validation splits, which has contains BCI data, and the held-out evaluation split containing only monkey behavior. The figures show that the ablation significantly decreases validation split performance, and causes a slightly earlier stopping point leading to worse evaluation performance. Note both models early before the full training budget of 400 epochs.

Neural reconstruction objective All main NDT3 models used a neural reconstruction objective inherited from the self-supervised learning pretraining from NDT2. We ablate this choice post-hoc and see it may actually minorly harm pretraining (validation split), though the neural objective doesn't harm evaluation split decoding (Fig. 8C). Note the scalar weighting of neural vs covariate objectives were set to be roughly balanced in pretraining.

MSE over classification In robotics and certain generalist models (Schubert et al., 2023), continuous action spaces are sometimes better decoded and controlled when quantized (Shafuallah et al., 2022). This is because MSE is an insufficient objective when the output distribution is multimodal (e.g. one of two possible paths in robotics). While it seems unlikely that the close relationship between movement behavior and motor cortex is multimodal, multimodal behavior may be appropriate when pretrained on heterogeneous data, i.e. when similar neural activity corresponds to different behavior in two datasets. We attempted such a quantization, including HL-Gauss smoothing (Farebrother et al., 2024) which we found to help; but this does not recover the performance of the default MSE objective (Fig. 8C) on the evaluation split. We found this performance gap persisted under fine-tuning (not shown). This suggests that NDT3 is differentiates neural data inputs from different datasets.

Patch size NDT2 and NDT3 both tokenize neural data by patching them into fixed size clusters. It is unclear whether transfer learning might occur for sub-token features, which motivates the use of smaller tokens in larger datasets that might afford it (Caron et al., 2021). We change patch size to 16 and show this performs slightly worse in the 45M 200h model (Fig. 8)C. Smaller patches (and subsequent increased neural tokens) may be more beneficial in the larger scale models, but their benefit must be weighed against their increased compute burden.

B.3 PRETRAINING DOES NOT BENEFIT FALCON H2 (HANDWRITING)

We also evaluated NDT3 for decoding of letters in a human-open loop handwriting task (FALCON H2). Although this is also a motor cortical decoding task, we excluded H2 from NDT3's aggregate evaluation since it is a sequence-to-sequence as opposed to continuous task. To apply NDT3 to this task, we pool neural tokens at each timestep and add a linear projection and optimize with a CTC loss (Graves et al., 2006). We maintain the default neural reconstruction loss and causal attention mask, and do not apply data augmentation.

Note that RNNs are the current standard architecture for communication tasks like H2 (Karpowicz et al., 2024; Willett et al., 2021). Training and tuning was less stable than for our continuous decoding tasks and required more extensive hyperparameter tuning, perhaps because the overall dataset size remains small (<1k samples), specific parameters are listed in the codebase. We observe three regimes in both training and fine-tuning. First, the model can fail to achieve an initial learning period. Second, the model can achieve reasonable nontrivial solutions, comparable to expected performance for unaugmented RNNs (though we do not quantify this). Third, some models will exhibit learning instabilities that resolve in significantly improved performance. We illustrate these regimes in example validation curves below. Overall, the third regime is rarely achieved. More relevant to the main narrative of this work, fine-tuning appears to degrade both final solution quality and reduces the range of nontrivial hyperparameters (not shown). Investigating a sequence to sequence objective over CTC loss would be valuable future work.

B.4 MULTISCALE DECODING ON INDIVIDUAL MOTOR TASKS

Fig. 10A plots model performance for each of the 31 evaluation settings we study in the eight primary evaluation datasets we use. Studying any individual dataset will yield variable conclusions on whether

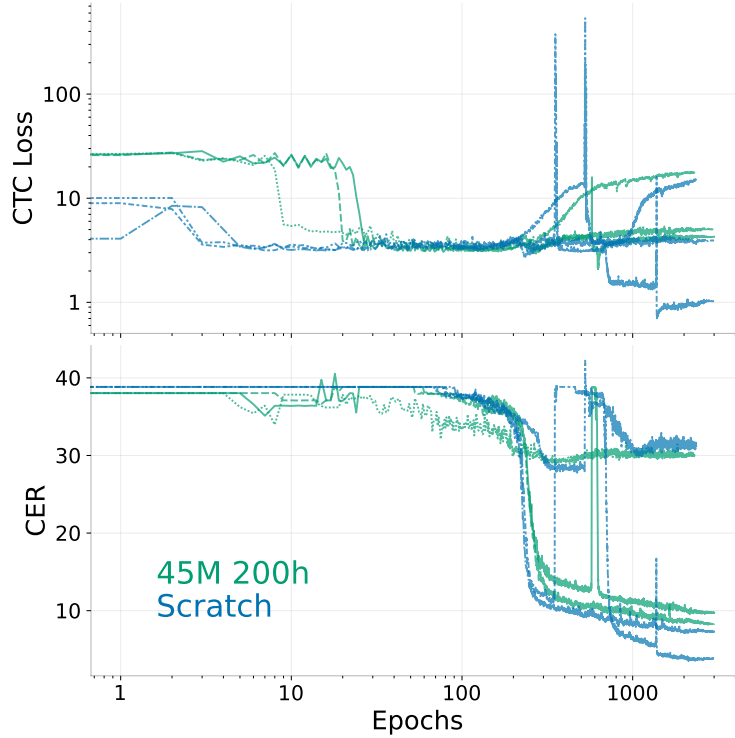


Figure 9. Three regimes of NDT3 training for handwriting decoding. We show validation loss and character error rates for example runs of from-scratch and fine-tuned NDT3s.

pretraining structure is helpful, underscoring the need for proposed foundation models to be evaluated across many different datasets. Here specifically we see the most clear scaling with pretraining data (color gradient with red on top) in the Critical Stability Task and Bimanual Task. FALCON tasks and Self-paced Reach appear minimally affected by scaling pretraining data, in that either pretrained models are generally slightly above a from scratch model at all data scales with no particular best pretrained model. The 2D + Click and Grasp datasets show uniquely show high variability in model performance and strong degradation of the 350M 2 khr model at low data scales. Grasp instability was so high that we trained 9 seeds instead of the standard 3 to better estimate model performance. We propose this degradation is due to the instability of full fine-tuning of large models at the extremely low data scales these datasets present (e.g. 2.5 minutes at the 25% scaling). Finally, we remind that the 2D + Click, FALCON H1, and 1D Grasp Force tasks are datasets from human participants that are included in the 2 khr pretraining. Surprisingly, we see no particular benefit to the 2 khr model.

These scaling plots also provide more precise context for baseline performance. NDT2 performs particularly poorly in the low data regime, while Wiener Filters perform poorly in the high data regimes.

In Fig. 10B, we illustrate qualitative predictions on private datasets. These visualizations show a diversity in covariate timescales and structure. They also illustrate that the summary R^2 obscure several features of model predictions. For example, pretrained models in Cursor Y tend have false positive deflections in movement. R^2 also is not easily comparable in tasks with continuous dynamics (CST) vs. transient dynamics (Cursor G1).

B.5 SEQUENTIAL TUNING IS SIMILAR TO JOINT TUNING

During the input shuffling analysis in Section 3.2, we showed that sequential transfer was vital for enabling cross-subject transfer in a from-scratch model Fig. 11A. In contrast, the pretrained NDT3s do not appear affected by this choice of sequential vs joint tuning. We show in Fig. 11B that in the different input shuffling conditions, pretrained model performance is similarly largely unaffected.

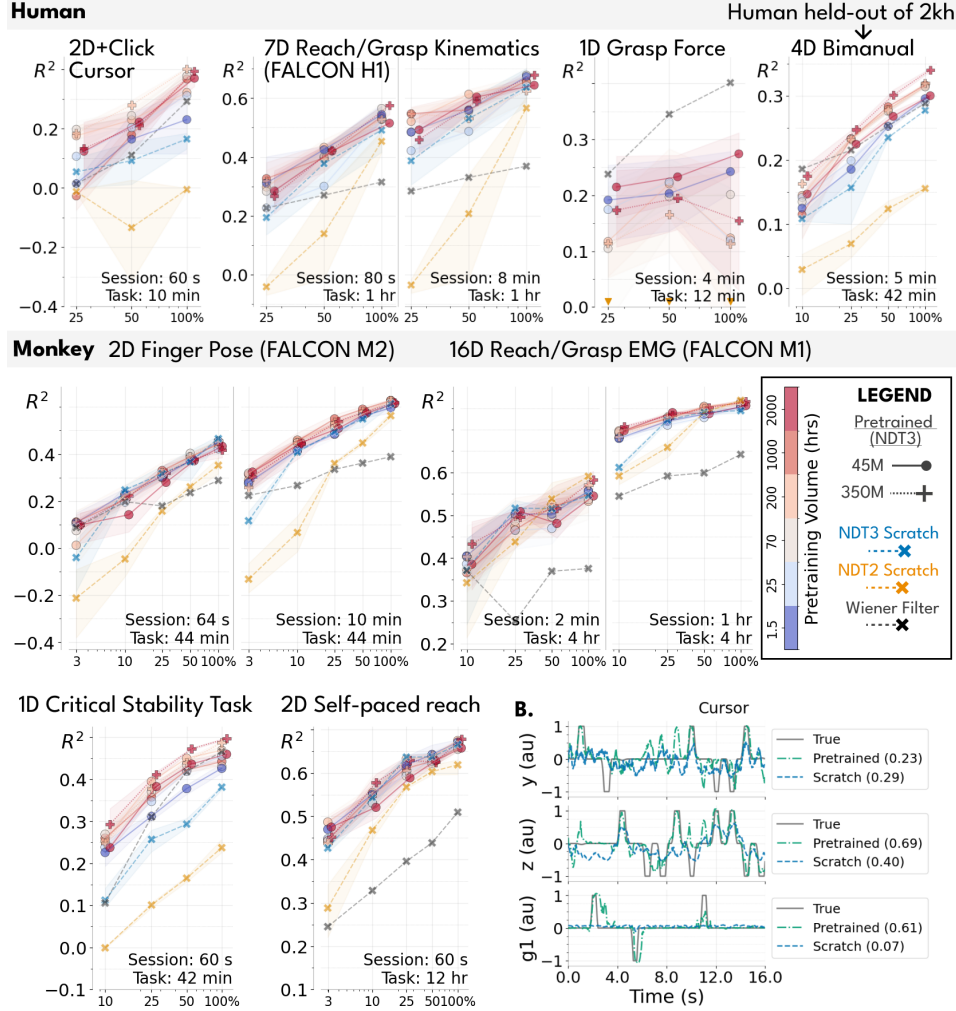


Figure 10. A. Fine-tuning evaluations for individual datasets. Performance on both held-out (left) and held-in (right) splits are shown side by side by FALCON datasets. We shade the standard deviation of 3 model seeds in fine-tuning. Different tasks show substantial variability in benefit from pretraining. B. We show example predictions of a pretrained (45M 200h) and from-scratch NDT3 for the 2D + Click Cursor task to give a sense of what different prediction performances mean in terms of open loop data prediction. Numbers in legend are the R^2 for that model’s predictions in the shown snippet.

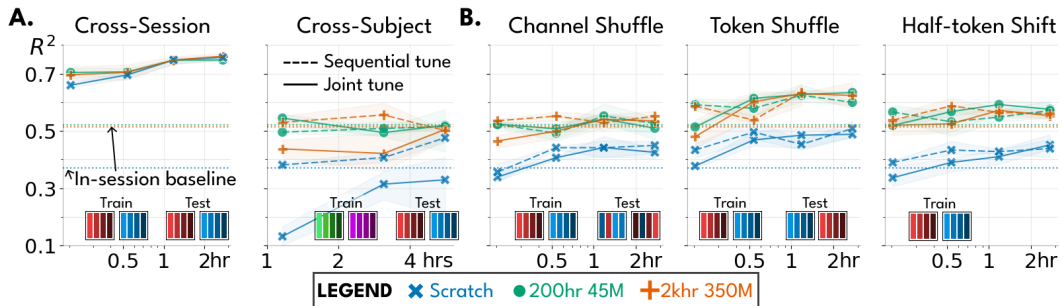


Figure 11. A replication of Fig. 4, but additionally providing sequential transfer results for input shuffling conditions in B.

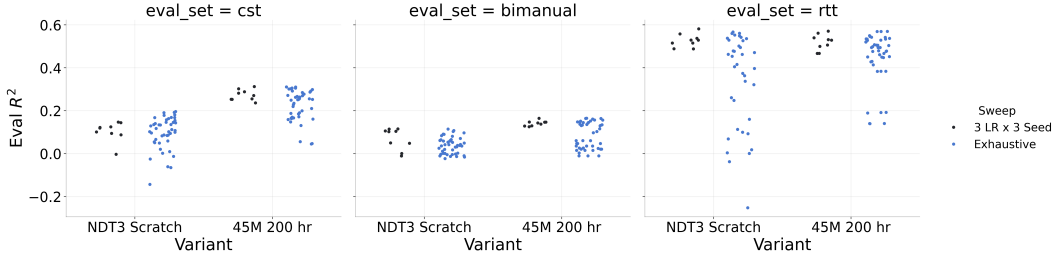


Figure 12. For 3 monkeys datasets at 10% scale, we extend a HP sweep to 5 LR and dropout in $[0.0, 0.1, 0.3]$ (vs default 0.1). For fine-tuning, we also sweep weight decay in $[0.0001, 0.01, 0.1]$ (vs default 0.1), while for from-scratch models we also sweep Transformer width ($[256, 512, 1024]$) vs default 512. This yields a 45-model sweep on 1 seed. We compare the range of scores achieved by this larger sweep against the standard 3 LR x 3 seed sweep.

C METHODS

C.1 METRICS AND EVALUATION

Throughout this work we evaluate offline decoding of continuous covariates timeseries. The metric we specifically use is the coefficient of determination, R^2 , as computed by scikit-learn’s `r2_score` function. R^2 is a useful metric over MSE as 1 represents perfect prediction and 0 is the score achieved by best-guess baseline, the mean of the data. In pretraining, R^2 is computed over the flat average of all covariate dimensions, since each datapoint has differing covariate dimensionalities. In evaluation, R^2 is computed as a variance-weighted average of R^2 s in each covariate dimension. Another difference between training and evaluation metrics is that training predictions are made over batched data, while evaluation predictions are mostly computed in a *streaming* fashion. Streaming requires continuous neural data across different behavioral epochs, and so cannot be performed for the Oculomotor and CST datasets. We also omit it for the motor cortex self-paced reach dataset, which has a very large evaluation split. Streaming allows timesteps at the beginning of each sequence to leverage neural context from the preceding sequence, which raises performance slightly, as shown in the continuous vs trialized analysis (Section 3.3). We limit history in streaming evaluations to the max history seen in tuning (1 second).

C.2 TRAINING

Pretraining hyperparameters were manually tuned in preliminary experiments at the 45M parameter models on small datasets. 350M models diverged at the chosen $4e-4$ peak LR, so we lowered peak LR to $1e-4$. For tuning, the explored LR are $1e-4, 3e-4, 5e-4$ for training from scratch and $3e-5, 1e-4, 4e-4$ for fine-tuning. While this is far from an exhaustive search, we show in Fig. 12 that other regularization hyperparameters are set to reasonable defaults such that this sweep finds near optimal results for both a from scratch model and fine-tuning the 45M model. Fine-tuning, like pretraining, is early stopped with a patience of 100 epochs. Batch size is uniformly set to 16K in pretraining, and scaled to be roughly 10-20% of dataset size in fine-tuning. NDT3 from-scratch models were trained at the 11M parameter range. Exact model configurations for different experiments are documented in the codebase.

NDT3’s simple architectural design allows us to train on batches from different tasks and dimensionalities. To avoid excess padding in training, we concatenate pretraining data that is otherwise discontinuous (trialized) into 2 second data. We do not add any separator tokens, as this does not appear to have a performance impact for language models (Geiping and Goldstein, 2022). With mixed-precision training, the 350M parameter NDT3 can fit the 4-8K tokens in each input context in the memory of 40G NVIDIA A100 GPUs. Thus we can restrict NDT3’s pretraining parallelism to data-parallelism.

Using Kaplan et al. (2020)’s equation for FLOP computation, $C_{\text{forward}} = 2N + 2n_{\text{layer}}n_{\text{ctx}}d_{\text{attn}}$, we compute the footprint of the 350M 2kh model. We use about 0.9B FLOPs per token in the forward pass, and about 0.9T neural tokens processed over training, which yields a pretraining footprint of about $2.4e21$ FLOPs.

C.3 BASELINES

Wiener Filter The Wiener Filter baseline was cross-validated over regularization strength. We also swept history of neural input up to the max length provided to NDT, and reported the R^2 of the best WF according to test data in primary evaluation (slightly advantaging the WF). Generalization plots in Section 3.3 report the performance of WF models at these different histories. For evaluating angular generalization, WFs were only swept up to 1s history due to memory limits; performance was not varying substantially with history so we do not expect this to have impacted conclusions. The WF was for simplicity directly fit on the concatenated trial data, which may have slightly negatively impacted its performance in trialized datasets (Oculomotor, CST, Generalization analyses).

In the primary evaluations in Section 3.1, we considered WFs fit either independently per session in a dataset or jointly on all sessions, which is helpful for sessions in very low data regimes. We report the better of the 2. In generalization analyses, for simplicity, we only report joint fits, which may cause a slight downward bias in performance.

Dataset	Patience	Held-In R^2	Held-Out R^2
H1	100	0.567 ± 0.034	0.453 ± 0.030
H1 (reproduction)	250	0.628 ± 0.011	0.517 ± 0.016
H1 ((Karpowicz et al., 2024))	250	0.62	0.52
M2	100	0.563 ± 0.015	0.352 ± 0.028
M2 (reproduction)	250	0.582 ± 0.002	0.391 ± 0.009
M2 ((Karpowicz et al., 2024))	250	0.63	0.43

Table 1. NDT2 H1 and M2 results when trained with 100 epochs of patience (this work) in fine-tuning vs 250 as in Karpowicz et al. (2024). We report mean and standard deviation of 3 model seeds on the FALCON evaluation (which is in turn a cross-session mean).

NDT2 NDT2 baselines were prepared with its public codebase. Max context length and patience were held constant across the models. This restriction to a patience of 100 accounts for some difference with the reported FALCON benchmark results in Karpowicz et al. (2024), as we note in Table 1. Other choices were left to NDT2 defaults. For example, NDT2 uses z-score normalization, which we kept. A major change to the NDT2 approach, for simplicity, is that we jointly trained NDT2 with its neural reconstruction loss (masking of 25%) and supervised decoding loss. This is true for all eight evaluation tasks except CST, where we used only the supervised decoding loss as the token dropout used in reconstruction can dropout all neural input. NDT2 hyperparameters were not explored widely, which likely is a source of its mediocre performance in this work. We did however sweep NDT2 over 2 model sizes (20M and 72M parameters) in addition to the standard 3 learning rates, which provides it twice the budget as NDT3.

C.4 PRETRAINING AND EVALUATION DATASETS

Pretraining datasets were comprised of historical data from several labs, the rough composition of which is shown in Fig. 2B. The evaluation behavior used during pretraining was reaching in 2 monkeys. The first monkey dataset came from a public release (Flint et al., 2012), and the second from a private dataset (REDACT lab). The latter had center-out reach in standard conditions and under visual feedback perturbations. The monkey in the second dataset is also present in the 1khr monkey and 2kh and up model dataset sizes, though performing in a different set of experiments.

Inherent to the process of large-scale scraping is a loss of detail on what precise tasks were used, so we only have a qualitative description of tasks we believe are well represented. NDT3 trains on a wide variety of reaching behaviors from relatively constrained (2D center-out reaching to fixed number of targets) to relatively unconstrained (self-paced, more targets, potentially 3D) and under experimental manipulations (delayed onset, multiple targets, different error thresholds requiring more precision). These reaching behaviors are described in both endpoint kinematics and as EMG. A smaller fraction of pretraining data are isometric and force related (force exerted against manipulandums) for wrist and arm motion. Human datasets contain a variety of iBCI tasks, with closed loop datasets reflecting both high and low quality control. These tasks include reach and grasp behavior from 1-10 degrees of freedom, as well as some individuated finger tasks for clicking.

We detail evaluation datasets in Table 2. Three datasets come from the FALCON benchmark (Karpowicz et al., 2024), two are based on public datasets ((O’Doherty et al., 2017; Deo et al., 2024)), and three are private. Note we avoid the Neural Latents Benchmark (Pei et al., 2021) as it does not directly measure decoding performance. For each evaluation dataset, we specify a tuning split and an evaluation split. Only tuning split data is changed when varying data scale. Tuning and evaluation splits are block-contiguous, i.e. trials are not interleaved, for better downstream applicability.

C.5 GENERALIZATION ANALYSES AND FURTHER EVALUATIONS

Intra-session generalization Posture, spring, and angular generalization evaluate OOD performance in the standard setup of comparing in-distribution and out-of-distribution performance directly (with changes in the underlying evaluation dataset) The intra-session temporal shift analysis is evaluated in an inverted, slightly more rigorous setting. Specifically, we trained two sets of models on the two different temporal blocks, and evaluated on an evaluation split in the later block, rather than only training on the early block and evaluating on both blocks. This way, the OOD shift is measured with respect to the same evaluation dataset.

C.6 ARCHITECTURAL DETAILS

NDT3 adopts several architectural innovations used in recent Transformer models. These were compared against baselines in preliminary experiments, but formal ablations in the final experimental setting were not conducted. We defer full description of the Transformer dimensions to the public codebase.

- FlashAttention 2 (Dao, 2023) is used to increase training and inference speeds. On the NERSC Perlmutter cluster, with FA2, 45M NDT3 trained at about 270M neural tokens per 40G A100 hour, 350M NDT3 trained at about 70M neural tokens per A100 hour. FA2 also enables use of the 350M model for real-time (<20ms) inference latency for iBCI control results.
- Positional Embeddings (Su et al., 2023): Rotary embeddings are applied to indicate the real-world timestep of every input token. Additionally, 48 categorical learned embeddings are reserved to distinguish token modality and position within a timestep (10 for neural, 16 for covariates, 16 for covariate constraints, 1 for reward/return, 1 for dummy tokens, remainder unused).
- QK Normalization (Dehghani et al., 2023; Wortsman et al., 2024): An additional layer norm is applied to the query and key embeddings, before the rotary embeddings, which helped stabilize training of the 350M parameter models.
- No context embeddings (Ye et al., 2023): Differing from NDT2, no learned embeddings for disambiguating input datasets were prepended to each input. This was removed for simplicity. Per GATO (Reed et al., 2022) and language modeling practices, we instead leave task / dataset disambiguation to the modeling process: In pretraining, the covariate maskout strategy allows for many tasks to be specified in-context (as later behavior can be inferred on the basis of earlier neural-behavioral token relationships). In fine-tuning, the tuning dataset already uniquely specifies the function to be learned.
- Cross entropy loss for spiking data prediction: We used the standard cross entropy loss to classify spike count over the Poisson loss common in many neural data architectures. Since the overall ablation of neural objective shows no large impact in this work, it is likely that this decision should be evaluated with neural data related tasks rather than decoding.

We document the Transformer model shapes considered in our work in Table 3. This shape is not systematically explored in our work, and is by historical artifact, slightly different than the shapes used in NLP/CV. Embedding parameters are negligible. One possible area of interest is that the feedforward expansion factor is 1 in our model, i.e. the MLP dimension is low. If MLPs do serve as memory stores in Transformers (Geva et al., 2021), increasing this shape may yield more performant model size scaling, given the heterogeneity of our datasets.

1350 D NDT3 MODEL CARD

1351

1352 The card is currently only provided in the codebase.

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

Table 2. Evaluation datasets used for multiscale decoding and generalization analyses. The references provide extended description of the behavioral task. Dashed line separates datasets for Section 3.1 and for analysis. Datasets use unsorted multi-unit activity and are processed in 1s chops unless otherwise mentioned.

Dataset	Description
FALCON H1, M1, M2 (Karpowicz et al., 2024)	3 separate single-subject multi-session datasets for different iBCI tasks. Data comes in a high data split (held-in), and a low-data split (held-out), with the intention on identifying methods that can achieve parity in the two settings. H1 is an open loop human dataset for calibrating 7D reach-and-grasp in a robot arm. M1 is a monkey reach-and-grasp task to different objects with EMG recordings. M2 is a monkey 2D finger movement task with manipulandum-measured kinematics. Scaling scores are reported on the test set.
Self-paced reach (RTT) (O’Doherty et al., 2017)	Monkeys reach for random targets one at a time in a small planar workspace. We decode 2D arm velocity in monkey Indy. Has neural data from M1 and S1, we use M1 in Section 3.1 and Section 3.2 and S1 in Section 3.3.
Bimanual Cursor Control (Deo et al., 2024)	A human open loop dataset where the participant attempts movement of one or both hands to control two cursors.
2D Cursor + Click (private)	Cursor control is a classic iBCI endpoint (Pandarinath et al., 2017; Wolpaw et al., 2002; Jarosiewicz et al., 2015). Two human participants attempt movement according to visually cued cursor movement and audiovisual click cues. We also use this dataset for trial structure analysis in Section 3.3.
Grasp force (private)	A open-loop dataset with two human participants attempting isometric power grasps. Specifically, participants were asked to match force output according to visual cues in a Mujoco environment. Grasps cued were both static (instant onset, hold, and offset) or dynamic (gradually increasing force). This dataset is valuable for human iBCI study because force modulation is required in many motor behaviors, and grasp force has primarily only been characterized in monkeys until now (Branco et al., 2019). Uses 2 second intervals due to long behavior timescale. We expect this dataset can be released by end of 2024.
Critical Stability Task (Quick et al., 2018) (private, trialized, sorted)	A monkey dataset collected to study continuous control relative to ballistic movement. The monkey balances a virtual cursor on a 1D workspace for up to 6 seconds.
Posture-varied Center-Out (Marino et al., 2024) (private, trialized, sorted)	A monkey center-out task, but the monkey’s hand is adjusted to one of 6 different starting positions. We use the central position as center and the rest as edge.
Spring-load (Mender et al., 2023)	A monkey moves fingers, clamped together in a manipulandum for effective 1DoF, is neutral or under spring load.
Center-out, Monkey J (Ma et al., 2022) (trialized)	Used in Section 3.2. A monkey performs an isometric center out task. Forces are measured by the manipulandum and converted to cursor velocity signals.
Center-out, Monkey V (private, trialized)	Used in Section 3.2. A monkey reaches to one of 8 radially arranged targets by moving a manipulandum (Kinarm).
Oculomotor pursuit (Noneman and Mayo, 2024) (private, trialized, sorted)	A monkey visually tracks (via smooth pursuit) a target that moves from center of workspace to one of four directions. A few dozen neurons are recorded on probes in each of frontal eye field (FEF) and area MT. We decode pupil velocity. The small number of neurons in this dataset required resetting NDT3 neural readin/readout layers.
FALCON H2	Human open loop dataset where a participant attempts movement to write letters cued on a screen (Willett et al., 2021; Fan et al., 2023). The large number of timesteps in this dataset required resetting NDT3 neural readin/readout layers (to use fewer neural tokens).

Model	Layers	Width	MLP Size	Heads	Parameters (M)
NDT3 Base	6	1024	1024	8	45
NDT3 Big	12	2048	2048	16	350

Table 3. Transformer Model Shapes used in this work.