Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data

Armin W. Thomas
Department of Psychology
Stanford University
athms@stanford.edu

Christopher Ré
Department of Computer Science
Stanford University
chrismre@stanford.edu

Russell A. Poldrack
Department of Psychology
Stanford University
poldrack@stanford.edu

Abstract

Self-supervised learning techniques are celebrating immense success in natural language processing (NLP) by enabling models to learn from broad language data at unprecedented scales. Here, we aim to leverage the success of these techniques for mental state decoding, where researchers aim to identify specific mental states (e.g., the experience of anger or joy) from brain activity. To this end, we devise a set of novel self-supervised learning frameworks for neuroimaging data inspired by prominent learning frameworks in NLP. At their core, these frameworks learn the dynamics of brain activity by modeling sequences of activity akin to how NLP models sequences of text. We evaluate the frameworks by pre-training models on a broad neuroimaging dataset spanning functional Magnetic Resonance Imaging data from $11,980$ experimental runs of $1,726$ individuals across 34 datasets and subsequently adapting the pre-trained models to benchmark mental state decoding datasets. The pre-trained models transfer well, generally outperforming baseline models trained from scratch, while models trained in a learning framework based on causal language modeling clearly outperform the others.

1 Introduction

In mental state decoding, researchers aim to identify certain mental states (e.g., deciding to accept or reject a gamble or the experience of happiness or fear) from brain activity [1]. To this end, researchers train predictive models to correctly identify (i.e., decode) these mental states from measured brain activity. At first sight, this approach seems straightforward, yet, researchers interested in training mental state decoding models are faced with the challenge that individual neuroimaging datasets, particularly functional Magnetic Resonance Imaging (fMRI) data, are often of high dimension and low sample size, such that individual samples can comprise many hundred thousand dimensions while a dataset includes only a few hundred samples for each of tens to hundreds of individuals. In this setting, predictive models are prone to overfitting, thereby not generalizing well to new datasets, which severely limits any generalizable insights that can be gained from training these models about the studied mental states and brain activity.

In spite of the low sample size of individual datasets, neuroimaging research has recently entered a big data era, as individual researchers more frequently share their collected datasets publicly [2, 3]. In addition, several efforts have been made to standardize the data structure [4] and preprocessing [5] of neuroimaging data. Together, these developments open up new possibilities for the application of pre-training in neuroimaging at scale by training models on broad, public neuroimaging data so that the resulting models can be adapted well to the datasets collected by individual researchers, without the need for much new data.

A wealth of empirical evidence has demonstrated that models pre-trained on large neuroimaging datasets generally achieve better mental state decoding performance on new data, while requiring less

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

training time and data, when compared to models trained from scratch (e.g., [6, 7, 8, 9]). Yet, much of this work has either pre-trained models on large but homogenous datasets, such as data from many individuals who all perform the same few tasks at the same few acquisition sites [10, 7], which can limit the generalizability of the resulting models due to potential systematic biases in homogeneous datasets (e.g., specific to the acquisition site or experimental paradigm [11, 12, 13, 14, 15]), or required highly-preprocessed input data (e.g., statistical maps summarizing the measured sequences of brain activity [16]). In addition, researchers often use standard supervised pre-training tasks, training models to decode specific mental states assigned to each sample of the data [6, 7, 10, 17]. While this kind of supervised training can be fruitful within individual neuroimaging datasets, it is difficult to extend to many datasets, as neuroimaging researchers usually do not adhere to standardized labeling schemes for mental states [18] when assigning labels to the mental states of their experiments. Due to this lack of standardization, it is often unclear whether two datasets contain the same or distinct mental states (for a detailed discussion, see [19]).

Here, we aim to overcome these limitations by leveraging recent advances in self-supervised learning to pre-train deep learning (DL) models on broad neuroimaging data. In particular, we take inspiration from natural language processing (NLP), where the application of self-supervised learning has led to unprecedented breakthroughs in the adaptive capabilities of models trained on public text data (e.g., [20, 21, 22, 23]), and test whether modeling sequences of brain activity akin to sequences of text enables the learning of generalizable representations of brain activity and its dynamics. To enable this, we represent sequences of brain activity akin to the embedded representation of words in sequences of text in NLP [24, 25] by representing each time point as a vector that describes whole-brain activity at this point in time as the activity of a set of functionally-independent brain networks [26].

Contributions. By representing sequences of brain activity similar to the embedded representation of text in NLP, we are able to devise novel self-supervised learning frameworks for neuroimaging data that are inspired by prominent learning frameworks in NLP, namely, sequence-to-sequence autoencoding [27], causal language modeling [20], and masked language modeling [21]. We pre-train DL models with these frameworks on a large-scale neuroimaging dataset comprising fMRI data from $11,980$ experimental runs of 1726 individuals in 34 datasets. To our knowledge, this represents one of the broadest neuroimaging datasets used for pre-training to date, spanning a wealth of experimental conditions and a diverse set of acquisition sites. We evaluate the downstream adaptation performance of the pre-trained models in two benchmark mental state decoding datasets and show that models pre-trained in a learning framework based on causal language modelling clearly outperform the others, while all pre-trained models generally outperform baseline models trained from scratch. To enable others to build on this work, we make our code, training data, and pre-trained models publicly available [1].

2 Methods overview

2.1 Input data: from brain volumes to brain networks

FMRI data are conventionally represented in four dimensions, such that the measured blood-oxygen-level-dependent (BOLD) signal is indicated in temporal sequences $S = \{V_1, ..., V_t\}$ of 3-dimensional volumes $V \in \mathbb{R}^{x \times y \times z}$ that show the BOLD signal intensity for each spatial location of the brain (as indicated by the three spatial dimensions x, y, and z). However, due to the strong spatial correlations of brain activity, the measured BOLD signal is often summarized differently by representing it as a set $\Theta = \{\theta_1, ..., \theta_n\}$ of n brain networks (i.e., parcels) θ, each describing the BOLD signal for some subset of voxels $v_{x,y,z} \subset V$ with correlated activity patterns.

Here, we utilize such a functional parcellation of brain activity to represent the measured BOLD signal, namely, Dictionaries of Functional Modes (DiFuMo), which was recently proposed by [28] and learned across millions of fMRI volumes. DiFuMo defines Θ through a sparse dictionary matrix $D \in \mathbb{R}^{p \times n}$ containing n networks, each defined by p values (one for each voxel $v \in V$, flattened over the three spatial dimensions). The BOLD signal of volume V_t is then described as a linear combination of these networks by the use of weights $\alpha_t \in \mathbb{R}^n$, such that $V_t = D\alpha_t$ and $\alpha_t = D^{\dagger} V_t$, where $D^{\dagger}$ indicates the pseudo-inverse of D: $D^{\dagger} = (D^T D)^{-1} D^T \in \mathbb{R}^{n \times p}$. We utilize a parcelation with $n = 1024$ networks, such that the resulting BOLD parcelation $X \in \mathbb{R}^{t \times n}$ describes

[1] github.com/athms/learning-from-brains

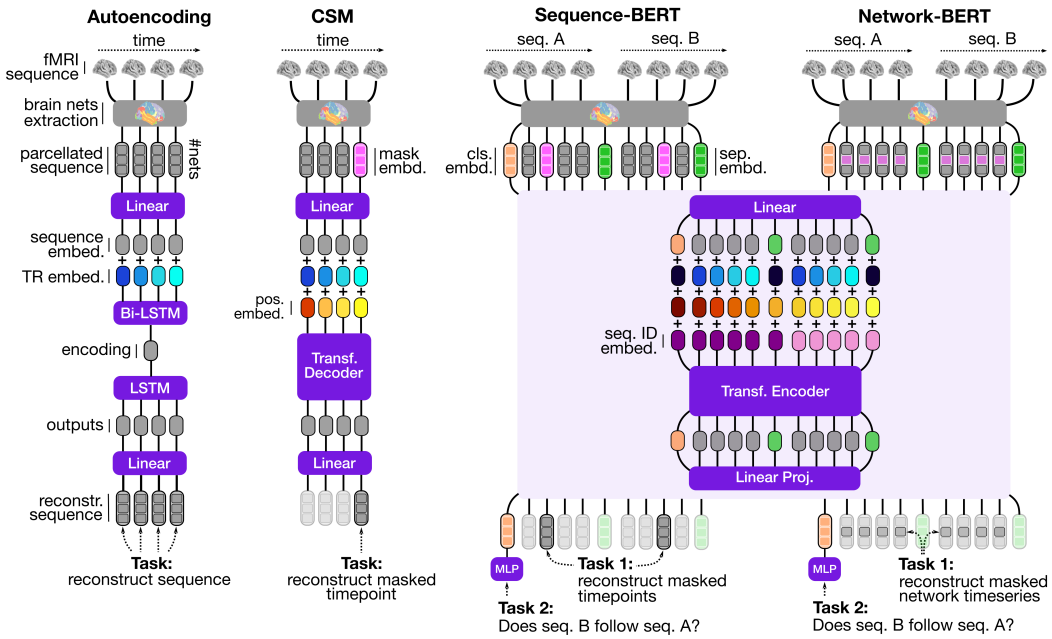

Figure 1: Proposed self-supervised learning frameworks for neuroimaging data, inspired by NLP.

the measured BOLD signal of each of these networks at each time point t of the input sequence S. This representation of the input data is thereby analogous to the embedded representation of sentences in NLP, where each word is represented by some embedding vector [24, 25].

2.2 Self-supervised learning frameworks: modeling sequences of brain activity akin to sequences of text

We first establish several commonalities among the frameworks before describing them individually in more detail (for an overview, see Fig. 1).

2.2.1 Commonalities among frameworks

First, each framework takes as input a parcellated BOLD sequence $X \in \mathbb{R}^{t \times n}$ (with t time points for n networks) as well as the corresponding sampling frequency (i.e., repetition time or TR) at which the fMRI sequence was collected (typically in the range of 0.7 s to 4 s).

Second, all frameworks linearly project the parcellated BOLD sequence X into an embedding representation $E^X \in \mathbb{R}^{t \times e}$: $E^X_{t,e} = b_e + \sum_n X_{t,n} w_{n,e}$, where $w \in \mathbb{R}^{n \times e}$ and $b \in \mathbb{R}^e$ indicate weights and biases, thereby reducing the dimensionality of the input: $\mathbb{R}^{t \times n} \rightarrow \mathbb{R}^{t \times e}$.

Third, to account for any differences in the sampling frequency of the BOLD sequences, the frameworks add one of r learnable TR embeddings $E^{TR} \in \mathbb{R}^{r \times e}$ to each time point $t \in E^X$. Each TR embedding corresponds to one possible time point k of the input. We set $r = 1500$ with a time increment of 0.2 s between subsequent TR embeddings, such that $k \in 0, 0.2, ..., 300$ s. For example, for an input sequence of 21 s that was collected at a TR of 0.7 s, we first round each time point $t \in 0, 0.7, ..., 21$ to the closest k and then add the corresponding E^{TR}_k to E^X_k. Note that we always label time points relative to the beginning of an input sequence, such that time point $k = 0$ is assigned to the first sequence element. Importantly, TR embeddings carry different information than the position embeddings commonly used in transformer models [29] because the same position in an input can correspond to different time points, depending on the sampling frequency at which the data were collected.

Fourth, all frameworks utilize standard sequence-to-sequence DL model architectures $f(\cdot)$ that take E^{in} as input and predict a corresponding output sequence $f(E^{in}) \in \mathbb{R}^{t \times e}$.

Fifth, during upstream learning, the output sequence $f(E^{in})$ is linearly projected back to the dimensionality of the parcellated BOLD sequence X, such that $\hat{X}_{t,n} = b_n + \sum_e f(E^{in})_{t,e} w_{e,n}$, where w and b again indicate weights and biases. This step is crucial, as all proposed learning frameworks involve reconstructing X (or parts of X) after some of its elements have been masked or it has been encoded in a lower-dimensional representation. To measure reconstruction performance, we utilize the mean absolute difference between the parcellated BOLD sequence X and its reconstruction $\hat{X}$: $L_{rec} = \frac{1}{n(J)} \sum_{j \in J} |X_j - \hat{X}_j|$, where J indicates the set of elements that is to be reconstructed and $n(J)$ the number of elements in J.

Sixth, for downstream adaptation to a new mental state decoding task, the frameworks utilize a simple decoding head $p(\cdot)$, composed of a dense hidden layer with e model units (one for each embedding dimension, with $tanh$ activation) as well as a $softmax$ output layer with one model unit i for each considered mental state in the data. $p(x)_i$ therefore indicates the probability that input x belongs to mental state i. Consequently, all frameworks treat the decoding task as a multinomial classification problem and use a standard cross entropy loss objective: $L_{cls} = -\sum_i y_i \log p(x)_i$, where y_i indicates a binary variable that equals 1 if i is the correct mental state and 0 otherwise.

2.2.2 Autoencoding

The autoencoding framework is inspired by recurrent neural machine translation models (e.g., [27]). These models take as input some text sequence, encode this sequence in a lower-dimensional representation, and predict an output sequence from the lower-dimensional encoding (e.g., a translation to another language).

Accordingly, the autoencoding framework uses a standard recurrent autoencoder architecture $f_a(\cdot)$, with an encoder and decoder part, each comprised of a stack of long short-term memory (LSTM [30]) units. The encoder $e(\cdot)$ takes E^{in} as input and encodes it in a lower-dimensional representation $h \in \mathbb{R}^e$. The decoder $d(\cdot)$ then takes h as input and predicts a corresponding output sequence $d(h) \in \mathbb{R}^{t \times e}$, such that $f_a(E^{in}) = d(h)$. Our encoder is composed of a stack of bidirectional LSTM units, whereas the decoder is build from a stack of unidirectional LSTM units. Note that we apply residual connections between the LSTM units in a stack: Let $LSTM^i$ and $LSTM^{(i+1)}$ be the i-th and $(i+1)$-th LSTM units in a stack (with weights w^i and w^{i+1}). At the t-th step of an input sequence, the inputs and outputs of units i and $(i+1)$ are defined as: $h_t^i, c_t^i = LSTM^i(h_{t-1}^i, c_{t-1}^i, x_{t-1}^{i-1}; w^i)$; $x_t^i = h_t^i + x_{t-1}^{i-1}$; $h_{t+1}^{i+1}, c_{t+1}^{i+1} = LSTM^{i+1}(h_t^{i+1}, c_t^{i+1}, x_t^i; w^{i+1})$, where $x_t^i \in \mathbb{R}^e$ represents the input to $LSTM^i$ at step t, whereas $h_t^i \in \mathbb{R}^e$ and $c_t^i \in \mathbb{R}^e$ represent hidden and cell state respectively. We define the sequence encoding h that is forwarded to the decoder as the mean of the final hidden states of the two LSTM units contained in the encoder's final bidirectional LSTM unit.

Upstream learning. During pre-training, we jointly train the encoder and decoder to reconstruct the parcellated BOLD sequence X by minimizing its mean absolute difference to the reconstructed sequence $\hat{X}$ (see section 2.2.1): $\min \frac{1}{t} \sum_{i=1}^{t} \frac{1}{n} \sum_{j=1}^{n} |\hat{X}_{i,j} - X_{i,j}|$. We also apply teacher-forcing [31] by using E^{in} in 50% of the cases as each next input for the decoder instead of its own preceding output.

Adaptation. To adapt the autoencoding framework to a new decoding task, we forward h as input to a corresponding decoding head $p(\cdot)$ (see section 2.2.1), thereby omitting the decoder $d(\cdot)$.

2.2.3 Causal sequence modeling (CSM)

The causal sequence modeling (CSM) framework is inspired by recent advances in causal language modeling [20], where models are trained to predict the next word in a text sequence. Specifically, these models receive some text sequence as input, represented as a matrix $X \in \mathbb{R}^{n \times e}$ with n words each encoded by an e-dimensional vector, where the last sequence element (representing the last word) is masked, for example, by replacing it with a learned mask embedding $E^{msk} \in \mathbb{R}^e$ (e.g., "When it rains, it [msk]"). The model's task is then to predict the masked word. In causal language modeling, models are thereby solely concerned with the part of the input sequence that precedes the masked word.

To adapt causal language modeling to neuroimaging data, we first mask the last time point T of a parcellated BOLD sequence X by replacing it with a learnable mask embedding E^{msk} before transforming X to E^{in} by means of linear projection and the addition of TR embeddings E^{TR}. In line with recent advances in causal language modeling, we utilize a standard transformer decoder model $f_d(\cdot)$ (based on the GPT architecture [20]) in this learning framework [2]. As transformer models are not aware of the ordering of their input, we further add learnable position embeddings $E^{pos} \in \mathbb{R}^{p \times e}$ to E^{in}, each representing one of p possible positions in the input sequence (we set $p = 512$ for all transformer models). Given E^{in}, the transformer decoder model predicts a corresponding output sequence $f_d(E^{in}) \in \mathbb{R}^{t \times e}$ in the form of the hidden states at the output of its last layer.

Upstream learning. In line with causal language modeling, the upstream objective of the CSM framework is to reconstruct the masked last time point T of the parcellated BOLD sequence X by minimizing the mean absolute difference to its reconstruction $\hat{X}_T$: $\min \frac{1}{n} \sum_{i=1}^{n} |\hat{X}_{T,i} - X_{T,i}|$.

Adaptation. During downstream adaptation, we attach a learnable classification embedding $E^{cls} \in \mathbb{R}^n$ to the end of BOLD parcellation X and forward the resulting prediction $f(E^{in})_{T+1}$ to a corresponding decoding head $p(\cdot)$ (see section 2.2.1).

2.2.4 Sequence-BERT

The Sequence-BERT framework is based on recent advances in bidirectional masked language modeling, particularly BERT [21]. BERT is trained to jointly solve a masked-language-modeling and next-sentence-prediction task. To this end, BERT receives two masked sentences S^a (e.g., "The sky is [msk].") and S^b (e.g., "The [msk] is shining.") as input, in which some fraction of words have been randomly masked. BERT is then asked to I) predict the masked words (masked-language-modeling) and II) determine whether S^b follows S^a (next-sentence-prediction). In contrast to causal language modeling, masked language modeling is considered as a bidirectional learning task, as the model is not just concerned with the parts of the input that precede a masked word but also the parts that follow it.

To adapt BERT to neuroimaging data, we first sample two input sequences $X^a \in \mathbb{R}^{t^a \times n}$ and $X^b \in \mathbb{R}^{t^b \times n}$ from the data (with lengths t^a and t^b). In 50% of the cases, the two sequences represent sequential parts of the same underlying BOLD sequence, while in the other 50% of cases, the two sequences are randomly sampled from two distinct BOLD sequences. Due to the strong temporal autocorrelation of BOLD data, we randomly leave a gap of 1-5 TRs (generally corresponding to a gap of 1 to 20 s) between the two sequences when sampling them from the same underlying BOLD sequence (to encourage the models to consider the entire input sequences and not just their connecting time points). In line with BERT, we next randomly mask time points of each sequence (with a probability of 20%) by replacing them with a learnable mask embedding E^{msk}. We also attach a separation embedding $E^{sep} \in \mathbb{R}^n$ to the end of each sequence (to indicate their endpoints) and a classification embedding E^{cls} to the beginning of S^a (for the next-sentence-prediction task; see upstream learning below) [3]. The resulting versions of X^a and X^b are then linearly projected to obtain respective BOLD embeddings E^{X^a} and E^{X^b} to which we further add TR and position embeddings E^{TR} and E^{pos} as well as sequence ID embeddings $E^{id} \in \mathbb{R}^{2 \times e}$ to indicate each of the two input sequences. The resulting embedding representation $E^{in} = [\{E_t^{X^a} + E_t^{TR} + E_t^{pos} + E_1^{id}\}^{t \in t^a}, \{E_t^{X^b} + E_t^{TR} + E_t^{pos} + E_2^{id}\}^{t \in t^b}]$ is then forwarded to a standard transformer encoder model $f_e(\cdot)$ (based on the BERT architecture [21]), which outputs a corresponding output sequence $f_e(E^{in}) \in \mathbb{R}^{t \times e}$ in the form of the hidden states at the output of its final layer.

Upstream learning. Similar to BERT, we define the upstream objective of the Sequence-BERT framework as the sum of two learning objectives: In line with masked-language-modeling, we define the first objective as the reconstruction loss L_{rec} for all masked time points J: $L_{rec} = \frac{1}{n(J)} \sum_{j \in J} \frac{1}{n} \sum_{i=1}^{n} |X_{j,i} - \hat{X}_{j,i}|$. In line with next-sentence-prediction, we define the other objective

[2] While the described input masking procedure is not needed for transformer decoder models, due to their causal attention mechanism, we chose this formulation of the task to ensure that the CSM framework also generalizes to other model architectures.

[3] We use the same learnable TR embedding (see section 2.2.1 and Fig. 1) for any other embeddings (e.g., classification embeddings) that are inserted into an input sequence during training.

as the binary cross entropy loss for a decoding head that receives $f(E^{in})_1$ (corresponding to the transformer output for classification embedding E^{cls}) as input and predicts probability p that X^b follows X^a: $L_{cls} = -y \log p - (1-y)\log(1-p)$, where y indicates a binary variable that equals 1 if the two sequences come from the same fMRI run and 0 otherwise.

Adaptation. During downstream adaptation, we forward the transformer's output for classification embedding $f_e(E^{in})_1$ as input to a corresponding decoding head $p(\cdot)$ (see section 2.2.1).

2.2.5 Network-BERT

All so far presented frameworks model their input on the level of the individual time points of a sequence by trying to reconstruct the parcellated whole-brain BOLD signal for specific time points. Yet, the dynamics of brain activity data can also be viewed differently by focusing on the interaction of network activities over time instead of on the distribution of whole-brain activity at individual time points. Consequently, we created a variant of the Sequence-BERT framework that randomly (at a rate of 10%) masks the activity of individual networks n in the parcellated BOLD sequence X by replacing them with a learnable network mask embedding $E^{msk} \in \mathbb{R}^t$ instead of masking whole-brain activity at individual time points. With the exception of the masking procedure, the Network-BERT and Sequence-BERT frameworks are identical.

Upstream learning. During upstream learning, Network-BERT uses the same cross-entropy loss L_{cls} as Sequence-BERT and adds it to a variant of the reconstruction loss that measures reconstruction performance for the masked network time courses J: $L_{rec} = \frac{1}{n(J)} \sum_{j \in J} \frac{1}{t} \sum_{i=1}^{t} |X_{i,j} - \hat{X}_{i,j}|$.

Adaptation. During downstream adaptation, Network-BERT forwards the transformer's output for classification embedding $f_e(E^{in})_1$ to a respective decoding head $p(\cdot)$ (see section 2.2.1).

2.3 Training details

We train all models with stochastic gradient descent and the ADAM optimizer (with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$) [32], if not reported otherwise. We also apply a linear learning rate decay schedule (with a warmup phase of 1% of the total number of training steps), gradient norm clipping at 1.0, and $L2$-regularisation (weighted by 0.1).

During upstream learning, we set the maximum learning rates to $2e^{-4}$, $5e^{-4}$, $1e^{-4}$, and $1e^{-4}$ for the autoencoding, CSM, Sequence-BERT, and Network-BERT frameworks, respectively (based on the learning rates used in corresponding NLP studies [20, 21, 27]) and randomly sample sequences X of 10 to 55 TRs from our upstream fMRI runs. For the Sequence- and Network-BERT frameworks, each sequence is split into two sequences X_i^a and X_i^b of equal length (with a randomly sampled gap of 1 to 5 TRs between the two sequences) and X_i^b randomly exchanged (with 50% probability) with another X_j^b (where $i \neq j$).

During downstream adaptation, we begin training with the pre-trained model parameters and then allow all parameters to be changed freely during training. Input sequences are drawn from the downstream fMRI runs according to the on- and offset of each experimental trial (with durations generally ranging between 2 and 30 s), when accounting for the temporal delay of the hemodynamic response function.

Compute resources. Upstream learning was performed on Google Compute Engine n1-highmem-64 nodes with four Nvidia Tesla P100 GPUs, 64 CPU threads, and 416 GB RAM memory, while downstream adaptations were performed on compute nodes of the Texas Advanced Computing Center with one Nvidia 1080-TI GPU, 32 CPU threads, and 128 GB RAM memory. Training times varied depending on model size but were generally in the range of 1 to 3 days for upstream learning and 0.5 to 3 hours for downstream adaptations.

3 Experiments

3.1 Datasets

All unprocessed fMRI data used in this study are publicly available through OpenNeuro.org [3] and the Human Connectome Project (HCP [33]). For an overview of our preprocessing, see Appendix A.3.

3.1.1 Upstream: $11,980$ fMRI runs from 1726 individuals across 34 datasets

Our upstream dataset comprises $11,980$ fMRI runs from $1,726$ individuals and 34 datasets (see Fig. 2 and Appendix Table A1). To our knowledge, this represents one of the broadest fMRI datasets used for pre-training in neuroimaging to date, spanning many acquisition sites and a diverse set of experimental conditions and domains. A few prominent examples of included datasets are The Midnight Scan Club [34], Individual Brain Charting [35], Amsterdam Open MRI collection [36], BOLD5000 [37], and the Narratives collection [38]. We randomly split the upstream data into distinct training and evaluation datasets by randomly designating 5% of the fMRI runs of each included fMRI dataset as evaluation data (at a minimum of 2 runs per dataset) and using the rest of the runs for training. At each evaluation step, we randomly sample $640,000$ sequences from the evaluation dataset.

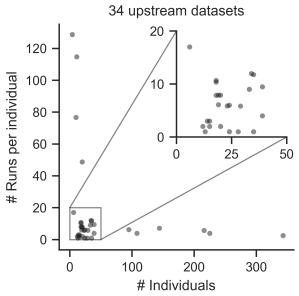

Figure 2: Number of individuals and average number of runs per individual of each upstream dataset.

3.1.2 Downstream: benchmark datasets with many mental states

To evaluate the adaptation performance of the pre-trained models, we utilize two benchmark mental state decoding datasets that span a wide range of mental states, namely, task-fMRI data from the HCP [33] and the Multi-Domain Task Battery (MDTB [39]). Our HCP dataset spans 100 participants and 19 distinct mental states across seven experimental tasks (see Appendix A.2). For each task, two fMRI runs were collected. We also include two resting state fMRI runs per HCP participant, in which individuals simply rest in the fMRI without any particular task, increasing the total number of mental states to 20. The MDTB dataset spans 24 individuals and 26 experimental tasks, which participants performed across two scanning sessions with eight fMRI runs per session. Due to the strong similarity of the experimental conditions within individual experimental tasks of the MDTB, we define each of its tasks as one target mental state (see Appendix A.2).

3.2 Hyper-parameter evaluation: larger embeddings but not deeper models

In our first experiment, we compared the performance of different hyper-parameter settings for the proposed frameworks. Specifically, we evaluated three different embedding dimensions e (192, 384, and 768; see section 2.2.1) and three different numbers of hidden layers for the used DL models (2, 4, and 6 hidden layers for the encoder and decoder parts of the autoencoding framework and 4, 8, and 12 hidden layers for the transformer models). In line with other work [40], we scaled the number of attention heads n_α per layer of the transformer models with the embedding dimension, such that $n_\alpha = \frac{e}{64}$. We trained one model for each point of the resulting 9-point grid of each framework for $200,000$ training steps, using a mini-batch size of 256 random sequences. Overall, model performances improved (in terms of validation loss) with larger embedding dimensions but not with the number of hidden layers (Fig. 3). We therefore decided to continue all further analyses with models comprising 768 embedding dimensions and 2 hidden layers for the encoder and decoder parts of the autoencoding framework and 4 hidden layers for the transformer models.

3.3 Upstream performance: models learn well in all frameworks

In our second experiment, we fully pre-trained one model in each framework according to our previously selected set of hyper-parameters (see section 3.2) by training models for $300,000$ training steps at a mini-batch size of 768 sequences. All models learned well over the course of their training,

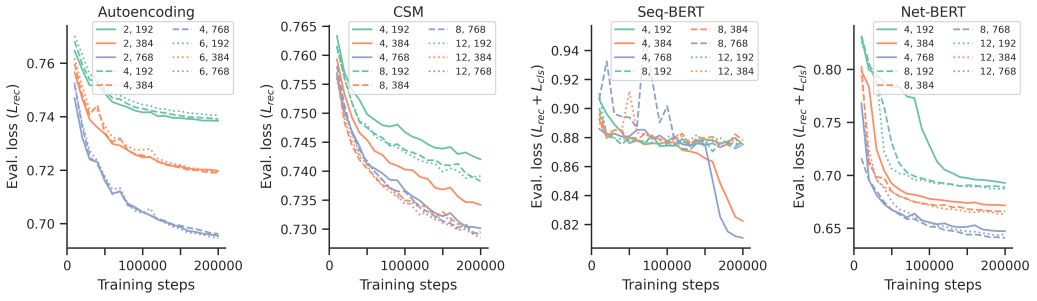

Figure 3: Hyper-parameter evaluation. Model performances improve with larger embedding dimensions but not with model depth. Note that we do not plot the evaluation loss for the Sequence-BERT variant with 12 hidden layers and 768 embedding dimensions because the loss of this model variant did not meaningfully change over the course of its training (see Appendix B.1).

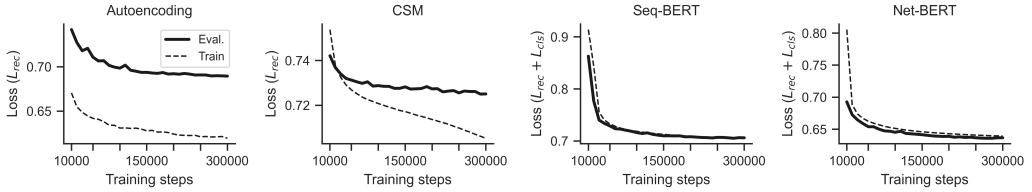

Figure 4: Upstream learning. Models learn well in each framework, with gradually decreasing training and evaluation losses.

with gradually decreasing training and validation losses (Fig. 4). Note that the transformer model trained in the CSM framework slightly overfitted the upstream data, whereas the gap in training and evaluation performance of the autoencoding framework results from the application of teacher-forcing during training (see section 2.2.2). We also evaluated the reconstruction performance of the final models for different parts of the brain and found that they exhibited similar overall distributions of reconstruction error, with relatively higher errors in the posterior parietal, occipital, and cingulate cortices as well parts of the limbic system (see Appendix B.2).

3.4 Adapting pre-trained language models does not yield upstream performance gains

Given that the formatting of our input data is highly similar to the embedded representation of sentences in language modeling, where each word of a sentence is represented by an embedding vector [41, 21], we next tested whether adapting pre-trained language models to our upstream data would yield meaningful performance gains, when compared to training models from scratch. To this end, we adapted pre-trained language model variants of GPT-2 [41] and BERT [21] (with 12 hidden layers and an embedding dimension of 768; as provided by Hugging Face's Transformer library [42]) to our upstream dataset (see Appendix B.3). The upstream validation performance of these two language models did not meaningfully improve upon the performance of our models trained from scratch (see section 3.3), at final upstream validation losses of 0.73 (CSM with GPT-2) and 0.80 (Sequence-BERT with BERT) (Appendix Fig. B3).

3.5 Downstream adaptation: CSM outperforms other frameworks

Lastly, in our fourth experiment, we evaluated the downstream adaptation performance of our pre-trained models in two benchmark mental state decoding datasets (HCP and MDTB; see section 3.1.2). To gauge the effectiveness of pre-training, we adapted the pre-trained models to varying sizes of the downstream datasets by randomly assigning individuals to distinct training, validation, and test datasets. Specifically, for each evaluated dataset size, we first randomly selected 10 (HCP) and 3 (MDTB) individuals whose data we use for validation and 20 (HCP) and 9 (MDTB) other individuals whose data we use for testing before randomly sampling the given number of individuals whose

data we use for training (1 to 48 (HCP) and 11 (MDTB)) from the remaining pool of individuals. We then adapted each pre-trained model with varying learning rates and numbers of training steps to the training data (see Appendix B.4) and used the model variant that achieved the highest final mental state decoding accuracy in the validation data for any further analyses (for an overview of validation decoding accuracies, see Appendix Fig. B6 and B7). For comparison, we also trained a linear baseline model (inline with the state-of-the-art in mental state decoding [43]) and a variant of our decoding head $p(\cdot)$ that we applied directly to the parcellated BOLD sequences X (see Appendix B.5).

Table 1: HCP test decoding performances. Models are trained on varying training dataset sizes ($N = 1, 3, 6, 12, 24, 48$ individuals) and tested on a distinct dataset ($N_{test} = 20$). F1-scores are macro-averaged and reported with standard errors. Bold font indicates best metric value. Asterisks indicate meaningfully better predictive accuracy than the respective second-best model in a McNemar test at $\alpha = 0.00083$ (multiple-comparison corrected: $0.005/6$).

Framework	Metrics	$N = 1$	$N = 3$	$N = 6$	$N = 12$	$N = 24$	$N = 48$
Linear	Acc, F1	24.2(±.69), 19.1	29.0(±.72), 27.1	34.2(±.75), 31.8	46.3(±.79), 47.0	51.8(±.79), 54.8	51.9(±.79), 55.6
$p(X)$	Acc, F1	53.5(±.79), **59.0**	54.1(±.79), 59.5	53.9(±.79), 59.4	56.0(±.78), 60.8	34.9(±.75), 39.9	50.2(±.79), 56.5
Autoencoding	Acc, F1	60.2(±.77), 47.7	69.3(±.73), 60.6	75.6(±.68), 67.4	82.9(±.60), 77.0	88.1(±.51), 84.1	91.5(±.44), 88.1
CSM	Acc, F1	**64.1(±.76)**∗, 51.4	**81.2(±.62)**∗, **72.9**	**86.8(±.54)**∗, **81.0**	**90.1(±.47)**∗, **85.8**	**92.7(±.41)**∗, **89.2**	**94.8(±.35)**∗, **92.0**
Seq-BERT	Acc, F1	39.5(±.77), 16.0	37.2(±.76), 21.6	52.1(±.79), 38.7	73.6(±.70), 65.6	73.4(±.70), 65.9	89.2(±.49), 84.5
Net-BERT	Acc, F1	36.9(±.76), 11.9	45.8(±.79), 31.4	59.0(±.78), 43.5	73.4(±.70), 64.4	82.8(±.60), 76.5	89.8(±.48), 85.2

Table 2: MDTB test decoding performances. Conventions as in Table 1 ($N_{test} = 9$).

Framework	Metrics	$N = 1$	$N = 3$	$N = 6$	$N = 11$
Linear	Acc, F1	56.1(±.47), 42.2	70.2(±.44), 65.6	75.6(±.41), 72.9	78.1(±.39), 77.7
$p(X)$	Acc, F1	71.8(±.43), 65.4	74.5(±.42), 70.0	77.6(±.40), 76.7	68.0(±.44), 59.8
Autoencoding	Acc, F1	68.8(±.44), 55.8	78.4(±.39), 70.8	83.4(±.35), 77.8	86.7(±.32), 83.4
CSM	Acc, F1	**77.1(±.40)**∗, **69.9**	**85.1(±.34)**∗, **83.1**	**88.5(±.30)**∗, **87.1**	**90.0(±.29)**∗, **89.7**
Seq-BERT	Acc, F1	63.1(±.46), 36.9	75.6(±.41), 57.9	82.5(±.36), 72.5	86.3(±.33), 79.6
Net-BERT	Acc, F1	71.6(±.43), 54.7	81.1(±.37), 72.1	84.6(±.34), 78.8	87.5(±.32), 84.3

The pre-trained models clearly outperformed the linear baseline model across all training dataset sizes, while their performances also scaled well with the number of individuals in the training data (Tables 1 and 2; at chance-levels of 5.0% (HCP) and 3.9% (MDTB)). Overall, the transformer decoder model trained in the CSM framework clearly outperformed the other models by achieving the highest test decoding accuracies throughout all evaluated sizes of the training dataset. The other pre-trained models performed on par with the decoding head $p(\cdot)$ in smaller training dataset sizes (≤ 6 (HCP) and ≤ 3 (MDTB) individuals), while clearly outperforming it in larger training datasets. For the HCP data, for which other reported decoding performances exist in the literature, the pre-trained models performed on par with models trained on much larger datasets (often comprising data of many hundred individuals [10, 7, 6, 44]), which generally report test decoding accuracies between 80% and 93%. A feature ablation analysis of the models' mental state decoding decisions in both datasets further revealed that the these are strongly dependent on the activity of the occipital and inferior temporal cortex as well as parts of the pre- and postcentral gyrus (see Appendix B.6).

We also tested whether the reported downstream model performances are stable over the non-deterministic aspects of their training [45] by replicating our adaptation analysis with different random seeds and splits of the data (see Appendix B.7). This replication confirmed our results, yielding final test decoding accuracies that were not meaningfully different from our initial analysis (Appendix Tables A3, A4, A5, A6).

4 Conclusion

In this work, we propose a set of novel self-supervised learning frameworks for neuroimaging data inspired by prominent learning frameworks in NLP. By the use of these frameworks, we are able to pre-train DL models on broad, public neuroimaging data, comprising many individuals, experimental domains, and acquisition sites. The pre-trained models generalize well to benchmark mental state decoding datasets and generally outperform baseline models trained from scratch, while models pre-trained in causal sequence modelling clearly outperform the others. We hope that this work

will inspire others to explore pre-training in functional neuroimaging at scale, using self-supervised learning.

Limitations. This work neglects a key aspect of mental state decoding, namely, the ability to draw inferences about the association between the decoded mental states and the underlying brain activity from the trained models. First empirical evidence indicates that attribution methods from explainable artificial intelligence (XAI [46]) research are well-suited to provide insights in the mental state decoding decisions of DL models [19]. Similarly, this work does not provide any insights into how the proposed self-supervised learning frameworks compare to contrastive learning techniques, which have demonstrated empirical success in computer vision [47] and medical imaging [48].

Potential negative social impact. We are currently not aware of any direct negative social impacts that could follow from pre-training DL models on public neuroimaging data that are de-identified and collected and shared with human consent in a way approved by institutional review boards.

Acknowledgments. We gratefully acknowledge the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ARL under No. W911NF-21-2-0251 (Interactive Human-AI Teaming); ONR under No. N000141712266 (Unifying Weak Supervision); ONR N00014-20-1-2480: Understanding and Applying Non-Euclidean Geometry in Machine Learning; N000142012275 (NEPTUNE); NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, Google Cloud, Salesforce, Total, the HAI-GCP Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), the Texas Advanced Computing Center (TACC) at The University of Texas at Austin, and members of the Stanford DAWN project: Facebook, Google, and VMWare. The OpenNeuro data archive is supported by NIH grant R24MH117179. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government.

References

[1] J.-D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, no. 7, pp. 523–534, 2006.

[2] C. Horien, S. Noble, A. S. Greene, K. Lee, D. S. Barron, S. Gao, D. O'Connor, M. Salehi, J. Dadashkarimi, X. Shen, *et al.*, "A hitchhiker's guide to working with large, open-source neuroimaging datasets," *Nature human behaviour*, vol. 5, no. 2, pp. 185–193, 2021.

[3] C. J. Markiewicz, K. J. Gorgolewski, F. Feingold, R. Blair, Y. O. Halchenko, E. Miller, N. Hardcastle, J. Wexler, O. Esteban, M. Goncavles, *et al.*, "The openneuro resource for sharing of neuroscience data," *Elife*, vol. 10, p. e71774, 2021.

[4] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, *et al.*, "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[5] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, *et al.*, "fmriprep: a robust preprocessing pipeline for functional mri," *Nature methods*, vol. 16, no. 1, pp. 111–116, 2019.

[6] A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux, "Extracting representations of cognition across neuroimaging studies improves brain decoding," *PLoS computational biology*, vol. 17, no. 5, p. e1008795, 2021.

[7] A. W. Thomas, U. Lindenberger, W. Samek, and K.-R. Müller, "Evaluating deep transfer learning for whole-brain cognitive decoding," *arXiv preprint arXiv:2111.01562*, 2021.

[8] U. Mahmood, M. M. Rahman, A. Fedorov, Z. Fu, and S. Plis, "Transfer learning of fmri dynamics," *arXiv preprint arXiv:1911.06813*, 2019.

[9] T. He, L. An, P. Chen, J. Chen, J. Feng, D. Bzdok, A. J. Holmes, S. B. Eickhoff, and B. Yeo, "Meta-matching as a simple framework to translate phenotypic predictive models from big to small data," *Nature Neuroscience*, vol. 25, pp. 795–804, 2022.

[10] Y. Zhang, L. Tetrel, B. Thirion, and P. Bellec, "Functional annotation of human cognitive states using deep graph convolution," *NeuroImage*, vol. 231, p. 117847, 2021.

[11] P. A. Kragel, M. Kano, L. Van Oudenhove, H. G. Ly, P. Dupont, A. Rubio, C. Delon-Martin, B. L. Bonaz, S. B. Manuck, P. J. Gianaros, *et al.*, "Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex," *Nature neuroscience*, vol. 21, no. 2, pp. 283–289, 2018.

[12] T. T. Liu, "Noise contributions to the fmri signal: An overview," *NeuroImage*, vol. 143, pp. 141–151, 2016.

[13] D. Sahoo and C. Davatzikos, "Learning robust hierarchical patterns of human brain across many fmri studies," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29034–29048, 2021.

[14] P.-H. C. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. Haxby, and P. J. Ramadge, "A reduced-dimension fmri shared response model," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[15] M. Yousefnezhad, A. Selvitella, L. Han, and D. Zhang, "Supervised hyperalignment for multisubject fmri data alignment," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 475–490, 2020.

[16] A. Mensch, J. Mairal, D. Bzdok, B. Thirion, and G. Varoquaux, "Learning neural representations of human cognition across many fmri studies," *Advances in neural information processing systems*, vol. 30, 2017.

[17] X. Wang, X. Liang, Z. Jiang, B. A. Nguchu, Y. Zhou, Y. Wang, H. Wang, Y. Li, Y. Zhu, F. Wu, *et al.*, "Decoding and mapping task states of the human brain via deep learning," *Human brain mapping*, vol. 41, no. 6, pp. 1505–1519, 2020.

[18] R. A. Poldrack, A. Kittur, D. Kalar, E. Miller, C. Seppa, Y. Gil, D. S. Parker, F. W. Sabb, and R. M. Bilder, "The cognitive atlas: toward a knowledge foundation for cognitive neuroscience," *Frontiers in neuroinformatics*, vol. 5, p. 17, 2011.

[19] A. W. Thomas, C. Ré, and R. A. Poldrack, "Challenges for cognitive decoding using deep learning methods," *arXiv preprint arXiv:2108.06896*, 2021.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

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

[24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[26] S. B. Eickhoff, B. Yeo, and S. Genon, "Imaging-based parcellations of the human brain," *Nature Reviews Neuroscience*, vol. 19, no. 11, pp. 672–686, 2018.

[27] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[28] K. Dadi, G. Varoquaux, A. Machlouzarides-Shalit, K. J. Gorgolewski, D. Wassermann, B. Thirion, and A. Mensch, "Fine-grain atlases of functional modes for fmri analysis," *NeuroImage*, vol. 221, p. 117126, 2020.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[34] E. M. Gordon, T. O. Laumann, A. W. Gilmore, D. J. Newbold, D. J. Greene, J. J. Berg, M. Ortega, C. Hoyt-Drazen, C. Gratton, H. Sun, *et al.*, "Precision functional mapping of individual human brains," *Neuron*, vol. 95, no. 4, pp. 791–807, 2017.

[35] A. L. Pinho, A. Amadon, T. Ruest, M. Fabre, E. Dohmatob, I. Denghien, C. Ginisty, S. Becuwe-Desmidt, S. Roger, L. Laurier, *et al.*, "Individual brain charting, a high-resolution fmri dataset for cognitive mapping," *Scientific data*, vol. 5, no. 1, pp. 1–15, 2018.

[36] L. Snoek, M. M. van der Miesen, T. Beemsterboer, A. van der Leij, A. Eigenhuis, and H. Steven Scholte, "The amsterdam open mri collection, a set of multimodal mri datasets for individual difference analyses," *Scientific data*, vol. 8, no. 1, pp. 1–23, 2021.

[37] N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, and E. M. Aminoff, "Bold5000, a public fmri dataset while viewing 5000 visual images," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[38] S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev, *et al.*, "The "narratives" fmri dataset for evaluating models of naturalistic language comprehension," *Scientific data*, vol. 8, no. 1, pp. 1–22, 2021.

[39] M. King, C. R. Hernandez-Castillo, R. A. Poldrack, R. B. Ivry, and J. Diedrichsen, "Functional boundaries in the human cerebellum revealed by a multi-domain task battery," *Nature neuroscience*, vol. 22, no. 8, pp. 1371–1378, 2019.

[40] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962*, 2019.

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[42] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

[43] M.-A. Schulz, B. Yeo, J. T. Vogelstein, J. Mourao-Miranada, J. N. Kather, K. Kording, B. Richards, and D. Bzdok, "Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets," *Nature communications*, vol. 11, no. 1, pp. 1–15, 2020.

[44] Y. Zhang, N. et Farrugia, and P. Bellec, "Deep learning models of cognitive processes constrained by human brain connectomes," *bioRxiv*, 2021.

[45] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, *et al.*, "Accounting for variance in machine learning benchmarks," *Proceedings of Machine Learning and Systems*, vol. 3, pp. 747–769, 2021.

[46] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019.

[47] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[48] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12546–12558, 2020.

[49] G. Xue and R. A. Poldrack, "The neural substrates of visual perceptual learning of words: implications for the visual word form area hypothesis," *Journal of cognitive neuroscience*, vol. 19, no. 10, pp. 1643–1655, 2007.

[50] R. A. Poldrack, E. Congdon, W. Triplett, K. Gorgolewski, K. Karlsgodt, J. Mumford, F. Sabb, N. Freimer, E. London, T. Cannon, *et al.*, "A phenome-wide examination of neural and cognitive function," *Scientific data*, vol. 3, no. 1, pp. 1–12, 2016.

[51] M. Hanke, F. J. Baumgartner, P. Ibe, F. R. Kaule, S. Pollmann, O. Speck, W. Zinke, and J. Stadler, "A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie," *Scientific data*, vol. 1, no. 1, pp. 1–18, 2014.

[52] C.-W. Woo, M. Roy, J. T. Buhle, and T. D. Wager, "Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain," *PLoS biology*, vol. 13, no. 1, p. e1002036, 2015.

[53] P. A. Smeets, F. M. Kroese, C. Evers, and D. T. de Ridder, "Allured or alarmed: counteractive control responses to food temptations in the brain," *Behavioural brain research*, vol. 248, pp. 41–45, 2013.

[54] J. Koster-Hale, R. Saxe, J. Dungan, and L. L. Young, "Decoding moral judgments from neural representations of intentions," *Proceedings of the National Academy of Sciences*, vol. 110, no. 14, pp. 5648–5653, 2013.

[55] J. Chen, Y. C. Leong, C. J. Honey, C. H. Yong, K. A. Norman, and U. Hasson, "Shared memories reveal shared structure in neural activity across individuals," *Nature neuroscience*, vol. 20, no. 1, pp. 115–125, 2017.

[56] J. Chen, C. Honey, E. Simony, M. J. Arcaro, K. A. Norman, and U. Hasson, "Accessing real-life episodic information from minutes versus hours earlier modulates hippocampal and high-order cortical dynamics," *Cerebral cortex*, vol. 26, no. 8, pp. 3428–3441, 2016.

[57] I. Momennejad, A. R. Otto, N. D. Daw, and K. A. Norman, "Offline replay supports planning in human reinforcement learning," *elife*, vol. 7, p. e32548, 2018.

[58] A. Shenhav, M. A. Straccia, S. Musslick, J. D. Cohen, and M. M. Botvinick, "Dissociable neural mechanisms track evidence accumulation for selection of attention versus action," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[59] R. Botvinik-Nezer, R. Iwanir, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, A. Dreber, C. F. Camerer, R. A. Poldrack, and T. Schonberg, "fmri data of mixed gambles from the neuroimaging analysis replication and prediction study," *Scientific data*, vol. 6, no. 1, pp. 1–9, 2019.

[60] M. Piva, K. Velnoskey, R. Jia, A. Nair, I. Levy, and S. W. Chang, "The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other," *Elife*, vol. 8, p. e44939, 2019.

[61] A. Shenhav, M. A. Straccia, J. D. Cohen, and M. M. Botvinick, "Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value," *Nature neuroscience*, vol. 17, no. 9, pp. 1249–1254, 2014.

[62] T. Nakai and S. Nishimoto, "Quantitative models reveal the organization of diverse cognitive functions in the brain," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[63] S. Schwettmann, J. B. Tenenbaum, and N. Kanwisher, "Invariant representations of mass in the human brain," *Elife*, vol. 8, p. e46619, 2019.

[64] J. A. Avery, A. G. Liu, J. E. Ingeholm, C. D. Riddell, S. J. Gotts, and A. Martin, "Taste quality representation in the human brain," *Journal of Neuroscience*, vol. 40, no. 5, pp. 1042–1052, 2020.

[65] M. E. Sachs, A. Habibi, A. Damasio, and J. T. Kaplan, "Dynamic intersubject neural synchronization reflects affective responses to sad music," *NeuroImage*, vol. 218, p. 116512, 2020.

[66] N. Saadon-Grosman, S. Arzy, and Y. Loewenstein, "Hierarchical cortical gradients in somatosensory processing," *Neuroimage*, vol. 222, p. 117257, 2020.

[67] E. Soreq, I. R. Violante, R. E. Daws, and A. Hampshire, "Neuroimaging evidence for a network sampling theory of individual differences in human intelligence test performance," *Nature communications*, vol. 12, no. 1, pp. 1–13, 2021.

[68] L. Tomova, K. L. Wang, T. Thompson, G. A. Matthews, A. Takahashi, K. M. Tye, and R. Saxe, "Acute social isolation evokes midbrain craving responses similar to hunger," *Nature Neuroscience*, vol. 23, no. 12, pp. 1597–1605, 2020.

[69] J. W. Antony, T. H. Hartshorne, K. Pomeroy, T. M. Gureckis, U. Hasson, S. D. McDougle, and K. A. Norman, "Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing," *Neuron*, vol. 109, no. 2, pp. 377–390, 2021.

[70] J. A. Avery, A. G. Liu, J. E. Ingeholm, S. J. Gotts, and A. Martin, "Viewing images of foods evokes taste quality-specific activity in gustatory insular cortex," *Proceedings of the National Academy of Sciences*, vol. 118, no. 2, p. e2010932118, 2021.

[71] E. Knights, C. Mansfield, D. Tonin, J. Saada, F. W. Smith, and S. Rossit, "Hand-selective visual regions represent how to grasp 3d tools: brain decoding during real actions," *Journal of Neuroscience*, vol. 41, no. 24, pp. 5263–5273, 2021.

[72] L. J. Chang, E. Jolly, J. H. Cheong, K. M. Rapuano, N. Greenstein, P.-H. A. Chen, and J. R. Manning, "Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience," *Science Advances*, vol. 7, no. 17, p. eabf7129, 2021.

[73] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] see section 3.3 and 3.5
   (b) Did you describe the limitations of your work? [Yes] see section 4
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] see section 4
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] see section 1
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see sections 2.3, 3.1.1, 3.2, 3.5, and Appendix B.5 and B.4
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We replicated our downstream adaptation analysis with different random seeds and splits of the data and found no meaningful differences in final test decoding accuracies of the individual training runs between the initial analysis and replication (see Appendix B.7)
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] see section 2.3
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] see section 3.1 and Appendix A.1
   (b) Did you mention the license of the assets? [Yes] see Appendix A.1
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] we make our pre-trained models and training data publicly available; see section 1
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] see section 4 and Appendix A.1
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] see section 4 and Appendix A.1
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Data details

A.1 Upstream datasets

Appendix Table A1 provides an overview of the datasets that we included in our upstream training [49, 50, 51, 52, 53, 54, 34, 55, 56, 37, 57, 58, 59, 60, 61, 62, 38, 35, 36, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72]. The unprocessed fMRI data of all datasets are publicly available (under a Creative Commons CC0 license) through OpenNeuro.org [3] under the specified identifier (ID) of each dataset. All fMRI data were de-identified and collected and shared with human consent in a manner approved by institutional review boards. We did not use any personally identifiable data.

Table A1: Overview of upstream datasets. For each dataset, the OpenNeuro.org identifier and DOI are given as well as the number of individuals and fMRI runs included in our upstream dataset, a brief text descriptor, and the DOI of an associated publication.

ID	DOI	#Individuals	#Runs	Text descriptor	DOI publication
ds000003	10.18112/openneuro.ds000003.v1.0.0	13	13	Rhyme judgment	10.1162/jocn.2007.19.10.1643
ds000009	10.18112/openneuro.ds000009.v1.0.0	24	144	The generality of self-control	unpublished
ds000030	10.18112/openneuro.ds000030.v1.0.0	144	1029	UCLA Consortium for Neuropsychiatric Phenomics LA5c Study	10.1038/sdata.2016.110
ds000113	10.18112/openneuro.ds000113.v1.3.0	20	976	Study Forrest	10.1038/sdata.2014.3
ds000140	10.18112/openneuro.ds000140.v1.0.0	33	297	Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain	10.1371/journal.pbio.1002036
ds000157	10.18112/openneuro.ds000157.v1.0.0	28	28	Block design food and nonfood picture viewing task	10.1016/j.bbr.2013.03.041
ds000212	10.18112/openneuro.ds000212.v1.0.0	39	370	Moral judgments of intentional and accidental moral violations across Harm and Purity domains	10.1073/pnas.1207992110
ds000224	10.18112/openneuro.ds000224.v1.0.3	10	767	The Midnight Scan Club (MSC) dataset	10.1016/j.neuron.2017.07.011
ds001132	10.18112/openneuro.ds001132.v1.0.0	15	45	Watching BBC's Sherlock	10.1038/nn.4450
ds001145	10.18112/openneuro.ds001145.v1.0.0	24	24	Watching The Twilight Zone	10.1093/cercor/bhv155
ds001499	10.18112/openneuro.ds001499.v1.3.1	4	515	BOLD5000	10.1038/s41597-019-0052-3
ds001612	10.18112/openneuro.ds001612.v1.0.2	23	135	Offline replay supports planning in human reinforcement learning	10.7554/eLife.32548
ds001715	10.18112/openneuro.ds001715.v1.0.0	34	407	Dissociable neural mechanisms track evidence accumulation for selection of attention versus action	10.1038/s41467-018-04841-1
ds001734	10.18112/openneuro.ds001734.v1.0.5	108	431	Neuroimaging Analysis Replication and Prediction Study (NARPS)	10.1038/s41597-019-0113-7
ds001882	10.18112/openneuro.ds001882.v1.0.0	19	150	Social Decision-Making Intertemporal Choice Task Dataset	10.7554/eLife.44939
ds001883	10.18112/openneuro.ds001883.v1.0.3	20	158	Social Decision-Making Risky Choice Task Dataset	10.7554/eLife.44939
ds001921	10.18112/openneuro.ds001921.v1.0.0	15	30	Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value (1)	10.1038/nn.3771
ds001923	10.18112/openneuro.ds001923.v1.0.0	14	42	Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value (2)	10.1038/nn.3771
ds002306	10.18112/openneuro.ds002306.v1.0.3	6	102	Over 100 Task fMRI Dataset	10.1038/s41467-020-14913-w
ds002345	10.18112/openneuro.ds002345.v1.1.4	343	861	Narratives Collection	10.1038/s41597-021-01033-3
ds002685	10.18112/openneuro.ds002685.v1.3.1	11	1263	Individual Brain Charting	10.1038/sdata.2018.105
ds002785	10.18112/openneuro.ds002785.v2.0.0	216	1235	Amsterdam Open MRI Collection-PIOP1	10.1038/s41597-021-00870-6
ds002790	10.18112/openneuro.ds002790.v2.0.0	225	887	Amsterdam Open MRI Collection-PIOP2	10.1038/s41597-021-00870-6
ds002841	10.18112/openneuro.ds002841.v1.0.1	29	169	Intuitive physics with fMRI	10.7554/eLife.46619
ds002995	10.18112/openneuro.ds002995.v1.0.1	18	192	Taste Quality Representation in the Human Brain	10.1523/JNEUROSCI.1751-19.2019
ds003085	10.18112/openneuro.ds003085.v1.0.0	39	156	Temporal Dynamics of Emotional Music	10.1016/j.neuroimage.2019.116512
ds003089	10.18112/openneuro.ds003089.v1.0.1	20	40	Somatosensory phase-encoded bilateral full-body light touch stimulation	10.1016/j.neuroimage.2020.117257
ds003148	10.18112/openneuro.ds003148.v1.0.1	35	412	Neuroimaging evidence for network sampling theory of human intelligence	10.1038/s41467-021-22199-9
ds003242	10.18112/openneuro.ds003242.v1.0.0	95	598	MRI data of 40 adult participants in response to a cue induced craving task following food fasting, social isolation and baseline (within-subject design)	10.1038/s41593-020-00742-z
ds003338	10.18112/openneuro.ds003338.v1.1.0	19	116	Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing	10.1016/j.neuron.2020.10.029
ds003340	10.18112/openneuro.ds003340.v1.0.2	18	142	Tasting Pictures: Viewing Images of Foods Evokes Taste-Quality-Specific Activity in Gustatory Insular Cortex	10.1073/pnas.2010932118
ds003342	10.18112/openneuro.ds003342.v1.0.0	18	187	Hand-selective visual regions represent how to grasp 3D tools for use: brain decoding during real actions	10.1523/JNEUROSCI.0083-21.2021
ds003521	10.18112/openneuro.ds003521.v1.0.0	35	35	Watching Friday Night Lights (Study 2)	10.1126/sciadv.abf7129
ds003524	10.18112/openneuro.ds003524.v1.0.0	12	24	Watching Friday Night Lights (Study 1)	10.1126/sciadv.abf7129

A.2 Downstream mental states overview

We provide a brief overview of the mental states included in both downstream datasets below. For any further details on the experimental procedures of the datasets, we defer the reader to the original publications [33] (HCP) and [39] (MDTB).

HCP. Appendix Table A2 provides an overview of the mental states of each HCP experiment task.

Table A2: HCP mental states. For each task, the mental states and total number of mental states are listed.

Task	Mental states	Count
Working memory	body, faces, places, tools	4
Gambling	win, loss	2
Motor	left / right finger, left / right toe, tongue	5
Language	story, math	2
Social	interaction, no interaction	2
Relational	relational, matching	2
Emotion	fear, neutral	2

MDTB. The MDTB dataset includes the following set of tasks, each representing one mental state in our analyses (as labeled by the original authors): CPRO, GoNoGo, ToM, actionObservation, affective, arithmetic, checkerBoard, emotionProcess, emotional, intervalTiming, landscapeMovie, mentalRotation, motorImagery, motorSequence, nBack, nBackPic, natureMovie, prediction, rest, respAlt, romanceMovie, spatialMap, spatialNavigation, stroop, verbGeneration, and visualSearch.

A.3 FMRI data preprocessing

We preprocessed all fMRI data with fMRIPrep (versions 20.2.0 and 20.2.3; a minimal, automated preprocessing pipeline for fMRI data [5]) using fMRIPrep's default settings without FreeSurfer [73] surface preprocessing. We then applied a sequence of additional minimal processing steps to fMRIPrep's derivatives, which included I) spatial smoothing of the fMRI sequences with a 3mm full-width at half maximum Gaussian Kernel, II) detrending and high-pass filtering (at 0.008 s) of the individual voxel activity time courses, and III) basic confound removal by regressing out noise in the data related to head movement (as indicated by the six basic motion regressors x, y, z, roll, pitch, and yaw) as well as the mean global signal and mean signal for white matter and cerebrospinal fluid masks (as estimated by fMRIPrep). Lastly, we parcellated each preprocessed fMRI run with the DiFuMo atlas (see section 2.2 of the main text) and standardized the resulting individual network time courses to have a mean of 0 and unit variance.

B Experiment details

B.1 Hyper-parameter evaluation run for largest Sequence-BERT model

The largest model variant of the transformer encoder model that we trained in the Sequence-BERT framework during our hyper-parameter evaluation (see section 3.2 of the main text), with 12 hidden layers and an embedding dimension of 768, did not meaningfully learn over the course of its training. For better visibility of the other models' performances, we decided to not include the model's training run in Fig. 3 and are instead showing it in Appendix Fig. B1.

B.2 BOLD reconstruction error in validation data

To evaluate the BOLD reconstruction performance of our pre-trained models (see section 3.3 of the main text) for different parts of the brain, we computed each models' mean reconstruction error (L_{rec}) in the upstream validation data for each network of our BOLD parcellation (see section 2.1 of the main text) and projected the average network reconstruction errors back to the voxel-level (as described in section 2.1 of the main text).

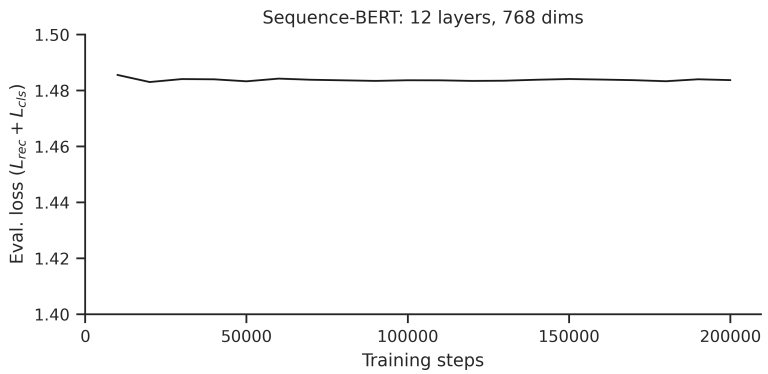

Figure B1: Upstream validation loss of the largest Sequence-BERT model variant with 12 hidden layers and an embedding dimension of 768.

The four pre-trained models exhibit similar distributions of reconstruction error throughout the brain (Appendix Fig. B2), with relatively higher errors in the posterior parietal, occipital, and cingulate cortices as well parts of the limbic system.

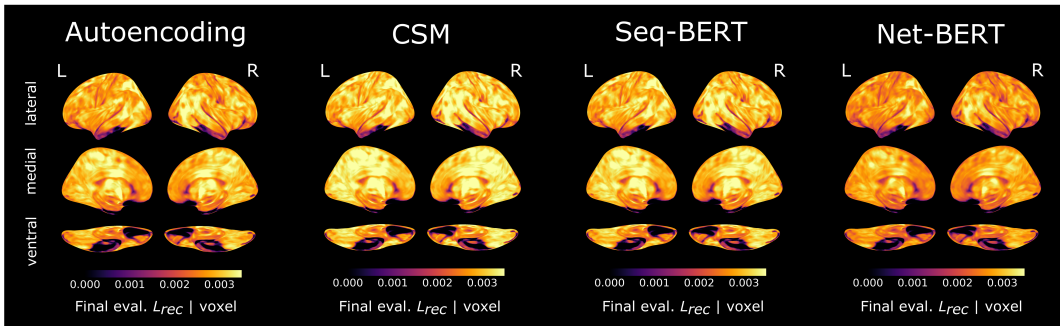

Figure B2: Mean voxel-wise reconstruction error (L_{rec}) of the final pre-trained models. Errors are projected onto the inflated cortical surface of the FsAverage template [73].

B.3 Adaptation of pre-trained language models

We adapted pre-trained language model variants of GPT-2 [41] and BERT [21] (as provided by Hugging Face's Transformer library [42]) to our upstream data in two training phases, using the CSM and Sequence-BERT frameworks, respectively (Appendix Fig. B3): First, we froze all parameters of the two language models and trained them for $20,000$ training steps at a mini-batch size of 512 and a learning rate of $1e^{-4}$, bringing all other parameters of the CSM and Sequence-BERT frameworks into sensible ranges. After this warmup phase, we continued training both models for a total of $150,000$ training steps of the same mini-batch size, using learning rates of $5e^{-4}$ (GPT-2) and $5e^{-5}$ (BERT) respectively, while allowing all model parameters to change freely.

B.4 Downstream adaptation of pre-trained models

For each downstream application of our pre-trained models, we evaluated two learning rates ($1e^{-5}$ and $5e^{-5}$) and different training lengths. Specifically, we scaled the number of training steps N_{step} according to the number of subjects N_{sub} in the training data, such that: $N_{step} = 1000 + \eta N_{sub}$, using two η-values (150 and 300) (see Appendix Fig. B6 and B7). We only report test decoding accuracies for the model variant achieving the highest validation decoding accuracy in this 4-point grid search in the main text (see section 3.5 of the main text).

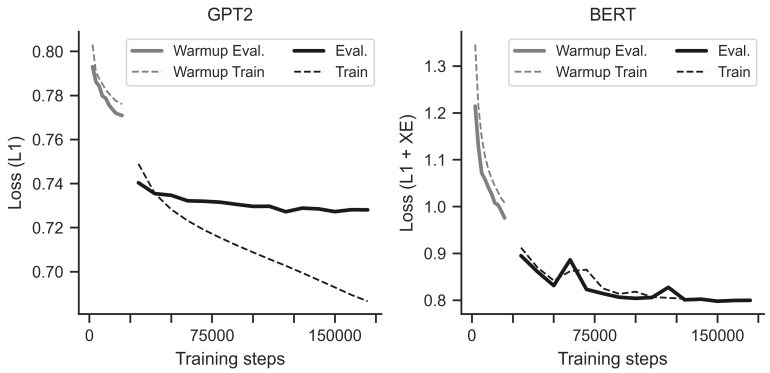

Figure B3: Adapting pre-trained language models to our upstream data in the CSM (GPT-2 [41]) and Sequence-BERT (BERT [21]) frameworks does not yield meaningful performance gains.

B.5 Baseline models

Linear baseline. The linear baseline model forms a decoding decision in two steps: it first aggregates the time course of each network $n \in X$: $a_n = b_n + \sum_t X_{t,n} w_{t,n}$ (with w and b indicating weights and biases) and then predicts a probability $p(a)_i$ that X belongs to mental state i from the distribution of aggregated time courses $a \in \mathbb{R}^n$: $p(a)_i = \sigma(b_i + \sum_n a_n w_{n,i})$, where σ represents the $softmax$ function and w and b indicate a second set of weights and biases.

Decoding head $p(\cdot)$. We also evaluated the performance of our decoding head $p(\cdot)$ (see section 2.2.1 of the main text) when applied directly to the parcellated BOLD data $X \in \mathbb{R}^{t \times n}$. To allow for the application of $p(\cdot)$ to inputs of varying length, we first averaged the signal of each network n over its time course ($\overline{X}_n = \frac{1}{t} \sum_{i=1}^{t} X_{i,n}$) before forwarding the time-averaged signal $\overline{X}$ to the decoding head to make a decoding decision for each mental state i: $p(\overline{X})_i$.

Training. Similar to the other frameworks, we trained both baseline models to minimize a standard cross-entropy loss with additional $L2$-regularization: $L_{cls} = -\sum_i y_i \log p_i + \lambda \sum_i w_i^2$, where y_i indicates a binary variable that equals 1 if i is the correct mental state and 0 otherwise, while λ scales the $L2$-regularization strength.

Hyper-parameter evaluation. For each application of the baseline models, we evaluated two learning rates ($1e^{-3}$ and $1e^{-4}$), three $L2$-regularisation strengths (0.1, 1, and 10), and two training lengths (1000 and 3000 training steps at a mini-batch size of 512; Note that the baseline models generally required fewer training steps to converge than the pre-trained models; see Appendix Fig. B6 and B7) in a 12-point grid search. As for the pre-trained models, we only report the test decoding accuracy of the model variant achieving the highest validation accuracy in the main text (see section 3.5 of the main text for details on the data split and Appendix Fig. B6 and B7 for an overview of model performances).

B.6 Downstream feature ablation analysis

To understand which parts of the brain were most relevant for the mental decoding decisions of the adapted models, we performed a feature ablation analysis of the test mental state decoding decisions of the best-performing models (that were adapted to the largest training dataset size of both downstream datasets; see Tables 1 and 2 of the main text). Specifically, for each sample of the test datasets, we replaced the signal of individual networks with random Gaussian noise from a standard normal distribution and measured the effect on the model's decoding prediction for the sample's mental state label. Note that we evaluated the predicted logits prior to their $softmax$ scaling. This analysis revealed that the decoding decisions of all pre-trained models are strongly dependent on the signal of the occipital and inferior temporal cortex as well as parts of the pre- and postcentral gyrus in both datasets (see Appendix Fig. B4 and B5).

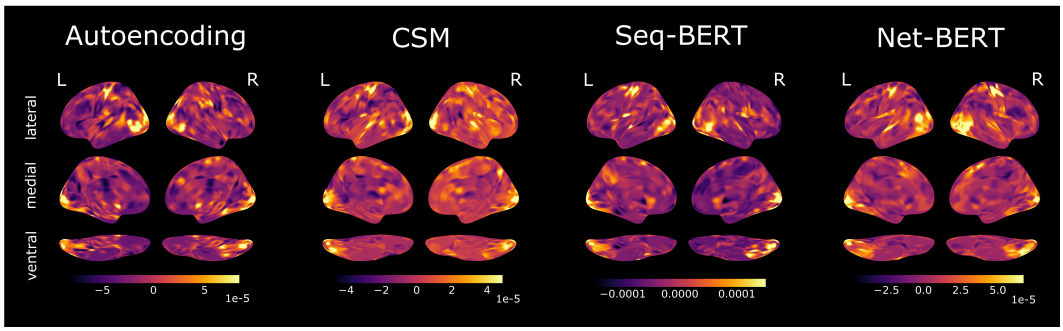

Figure B4: Ablation analysis of the pre-trained models for the HCP test dataset. For each sample, we ablate the time course of individual networks and measure the resulting effect on the models' prediction for the mental state associated with the sample. Average changes in output are projected onto the inflated cortical surface of the FsAverage template [73].

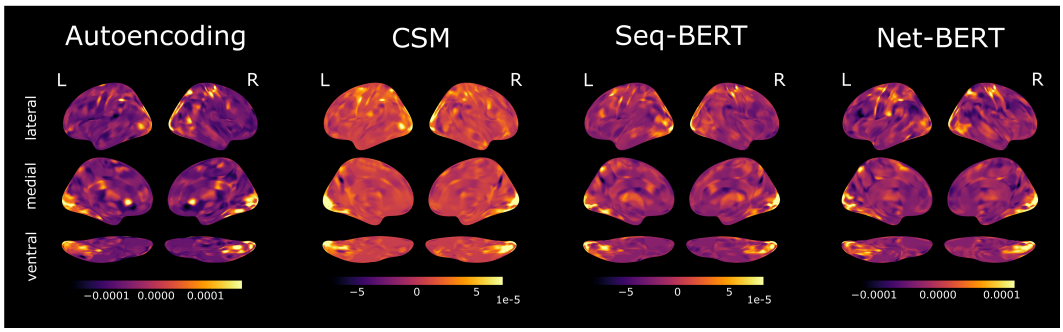

Figure B5: Ablation analysis of the pre-trained models for the MDTB test dataset. Conventions as in Appendix Fig. B4.

B.7 Replication of downstream adaptation analysis

To test for the stability of the reported model performances over the non-deterministic aspects of their training (such as different random weight initializations or random shufflings of the data during training), we replicated our downstream adaptation analysis (see section 3.5 of the main text) with different random splits of the data and different random seeds per split. This replication confirmed our initial results (compare Tables 1 and 2 of the main text with Appendix Tables A3 and A5).

We also tested whether the final test decoding accuracy of each model training run of the analysis was meaningfully different between our initial analysis (see Appendix Fig. B6 and B7) and the replication (see Appendix Fig. B8 and B9) by computing the difference in final test decoding accuracy between each initial training run and its replication and testing the resulting distribution of test decoding accuracy differences against a mean of 0 in a two-sided t-test (Appendix Tables A4 and A6). The models' final test decoding accuracies were not meaningfully different between our initial analysis and the replication, indicating strong stability of the models' performance over the various non-deterministic aspects of their training.

Table A3: HCP test performances in replication analysis. Conventions as in Table 1.

Framework	Metrics	$N = 1$	$N = 3$	$N = 6$	$N = 12$	$N = 24$	$N = 48$
Linear	Acc, F1	33.1(±.74), 28.8	36.8(±.76), 35.3	44.7(±.79), 45.7	49.1(±.79), 50.2	50.0(±.79), 50.7	55.6(±.79), 53.7
$p(X)$	Acc, F1	53.0(±.79), 58.9	50.1(±.79), 58.8	53.8(±.78), 61.2	55.8(±.79), 61.7	35.2(±.76), 40.1	51.5(±.79), 59.0
Autoencoding	Acc, F1	59.1(±.78), 44.9	67.2(±.74), 55.7	72.4(±.71), 64.8	81.5(±.61), 76.6	85.8(±.55), 80.8	89.5(±.48), 84.5
CSM	Acc, F1	**73.4(±.70)∗, 61.6**	**78.2(±.65)∗, 69.9**	**85.9(±.55)∗, 80.9**	**90.6(±.46)∗, 86.1**	**91.2(±.45)∗, 87.7**	**94.4(±.36)∗, 91.6**
Seq-BERT	Acc, F1	37.1(±.76), 17.3	43.3(±.78), 16.1	52.2(±.79), 37.8	66.9(±.74), 56.2	81.2(±.62), 73.5	88.6(±.50), 84.1
Net-BERT	Acc, F1	36.6(±.76), 13.4	49.4(±.79), 27.4	57.9(±.78), 46.6	68.9(±.73), 61.3	77.9(±.66), 70.0	87.1(±.53), 81.4

Table A4: Statistical comparison of HCP test decoding performances between the initial analysis and replication. Two-sided t-tests compare the the distribution of differences in final test decoding accuracy between each fitting run of the initial analysis and its replication against 0.

	$N = 1$	$N = 3$	$N = 6$	$N = 12$	$N = 24$	$N = 48$
Linear	$t(11) = -1.34, p = 0.21$	$t(11) = -1.54, p = 0.15$	$t(11) = -4.12, p = 0.002$	$t(11) = -2.88, p = 0.02$	$t(11) = -4.22, p = 0.001$	$t(11) = -3.58, p = 0.004$
$p(X)$	$t(11) = 1.08, p = 0.30$	$t(11) = 0.73, p = 0.48$	$t(11) = -0.36, p = 0.73$	$t(11) = 1.35, p = 0.21$	$t(11) = 0.33, p = 0.75$	$t(11) = -1.16, p = 0.27$
Autoencoding	$t(3) = 6.67, p = 0.007$	$t(3) = 0.92, p = 0.43$	$t(3) = 0.78, p = 0.49$	$t(3) = 1.75, p = 0.18$	$t(3) = 4.39, p = 0.02$	$t(3) = 0.72, p = 0.52$
CSM	$t(3) = -0.86, p = 0.45$	$t(3) = 4.32, p = 0.02$	$t(3) = 1.55, p = 0.22$	$t(3) = 1.28, p = 0.29$	$t(3) = 3.67, p = 0.03$	$t(3) = 1.55, p = 0.22$
Seq-BERT	$t(3) = -0.18, p = 0.87$	$t(3) = 0.1, p = 0.93$	$t(3) = -1.56, p = 0.22$	$t(3) = -0.72, p = 0.53$	$t(3) = -3.17, p = 0.05$	$t(3) = -1.28, p = 0.29$
Net-BERT	$t(3) = 0.19, p = 0.86$	$t(3) = 1.28, p = 0.29$	$t(3) = -0.41, p = 0.71$	$t(3) = 0.44, p = 0.69$	$t(3) = 0.9, p = 0.44$	$t(3) = 0.23, p = 0.83$

Table A5: MDTB test performances in replication analysis. Conventions as in Table 1.

Framework	Metrics	$N = 1$	$N = 3$	$N = 6$	$N = 11$
Linear	Acc, F1	59.2(±.47), 45.7	69.6(±.44), 63.3	73.5(±.42), 71.0	79.3(±.39), 77.7
$p(X)$	Acc, F1	71.6(±.43), **65.6**	70.9(±.43), 67.6	77.2(±.40), 76.9	63.2(±.46), 58.3
Autoencoding	Acc, F1	67.9(±.44), 50.8	80.9(±.37), 72.2	83.8(±.35), 77.3	83.8(±.35), 80.4
CSM	Acc, F1	**73.5(±.42)**, 60.3	**84.3(±.35)∗, 83.5**	**87.2(±.32)∗, 85.6**	**90.0(±.27)∗, 89.7**
Seq-BERT	Acc, F1	63.9(±.46), 35.6	83.0(±.32), 81.2	85.8(±.31), 82.2	88.8(±.30), 86.9
Net-BERT	Acc, F1	72.0(±.42), 58.6	82.8(±.36), 75.0	86.0(±.34), 79.5	85.2(±.34), 78.4

Table A6: Statistical comparison of MDTB test decoding performances between the initial analysis and replication. Conventions as in Table A4.

	$N = 1$	$N = 3$	$N = 6$	$N = 11$
Linear	$t(11) = 0.81, p = 0.44$	$t(11) = 0.95, p = 0.36$	$t(11) = 1.09, p = 0.30$	$t(11) = -0.98, p = 0.35$
$p(X)$	$t(11) = -2.55, p = 0.03$	$t(11) = -0.3, p = 0.77$	$t(11) = 0.87, p = 0.40$	$t(11) = -2.58, p = 0.03$
Autoencoding	$t(3) = 2.1, p = 0.13$	$t(3) = 0.35, p = 0.75$	$t(3) = 0.15, p = 0.89$	$t(3) = 0.47, p = 0.67$
CSM	$t(3) = -1.22, p = 0.31$	$t(3) = -1.29, p = 0.29$	$t(3) = -3.29, p = 0.05$	$t(3) = -1.61, p = 0.21$
Seq-BERT	$t(3) = -2.0, p = 0.14$	$t(3) = -0.95, p = 0.41$	$t(3) = -0.56, p = 0.62$	$t(3) = -1.32, p = 0.28$
Net-BERT	$t(3) = -3.75, p = 0.03$	$t(3) = -4.02, p = 0.028$	$t(3) = -4.62, p = 0.02$	$t(3) = -2.09, p = 0.13$

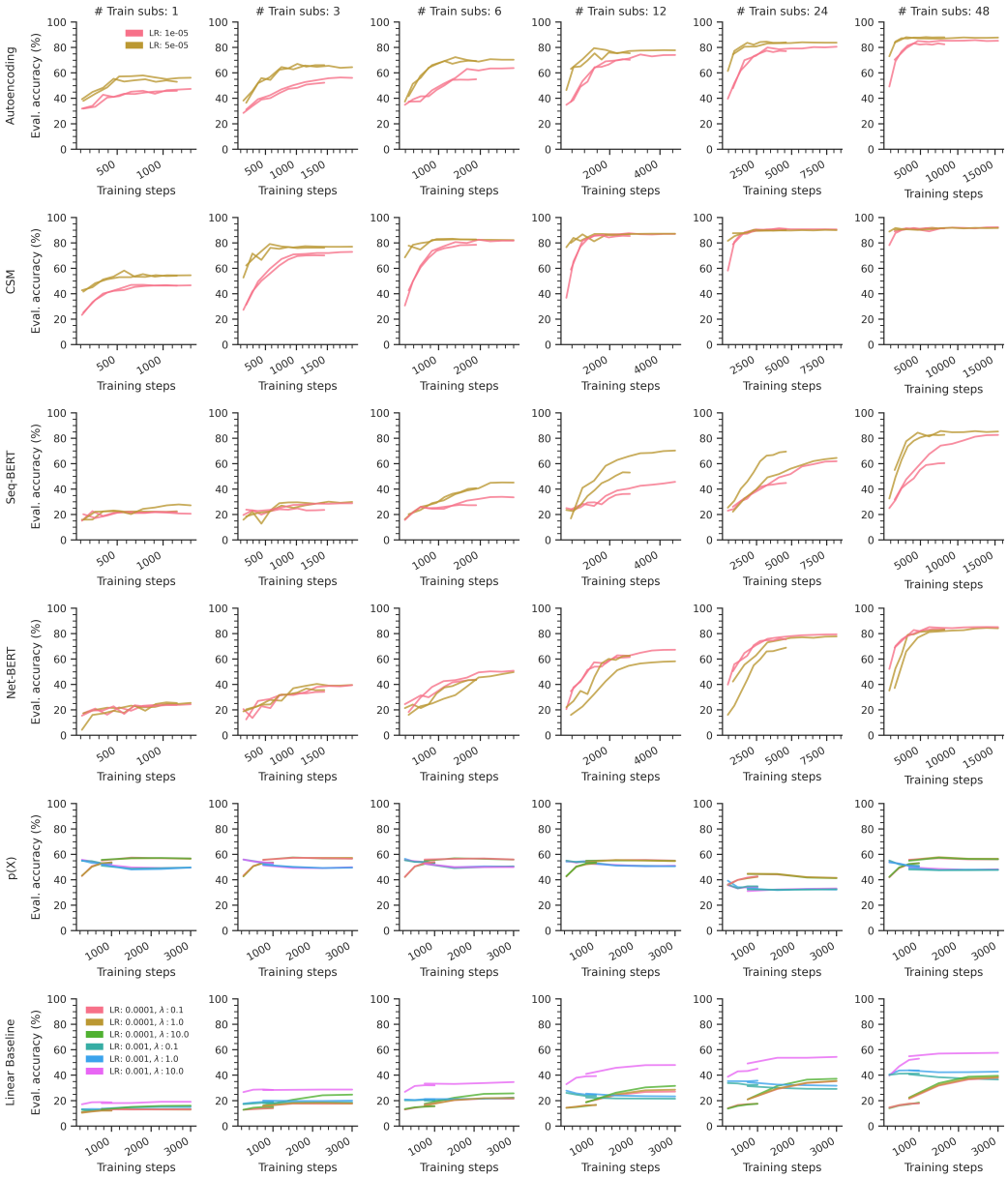

Figure B6: Model validation decoding accuracies during downstream adaptation to HCP data.

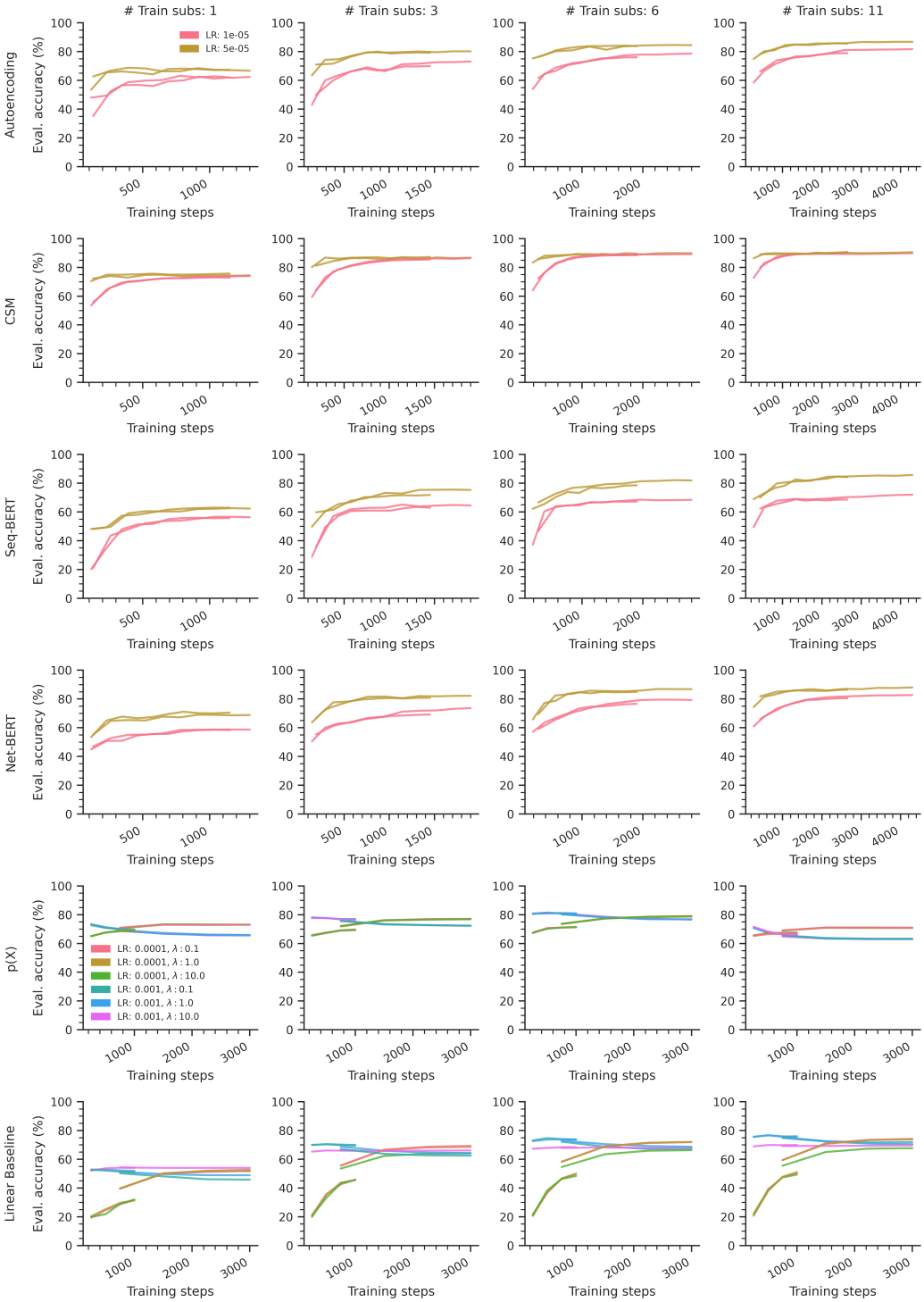

Figure B7: Model validation decoding accuracies during downstream adaptation to MDTB data.

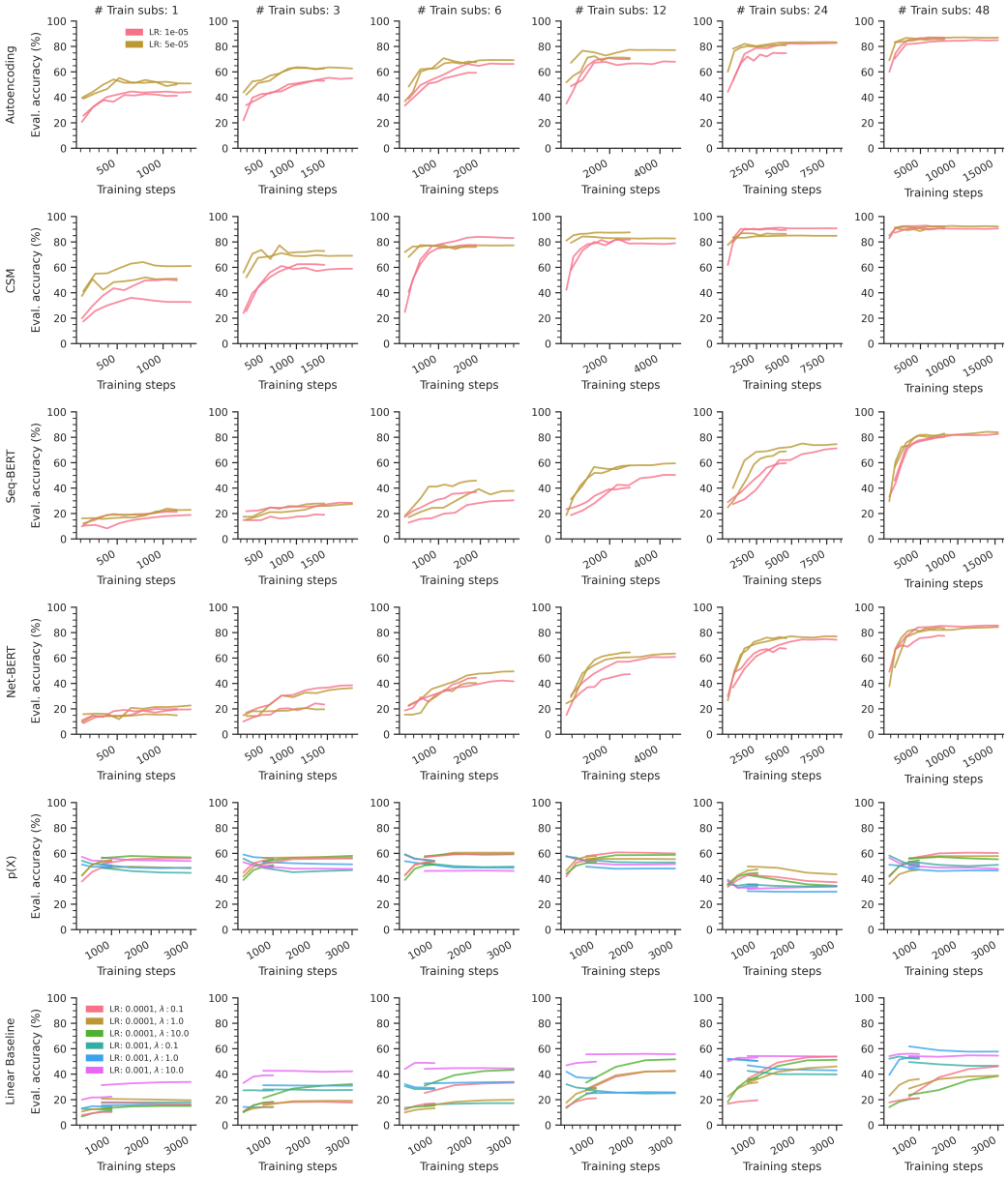

Figure B8: Replication of Appendix Fig. B6 with a different set of random seeds.

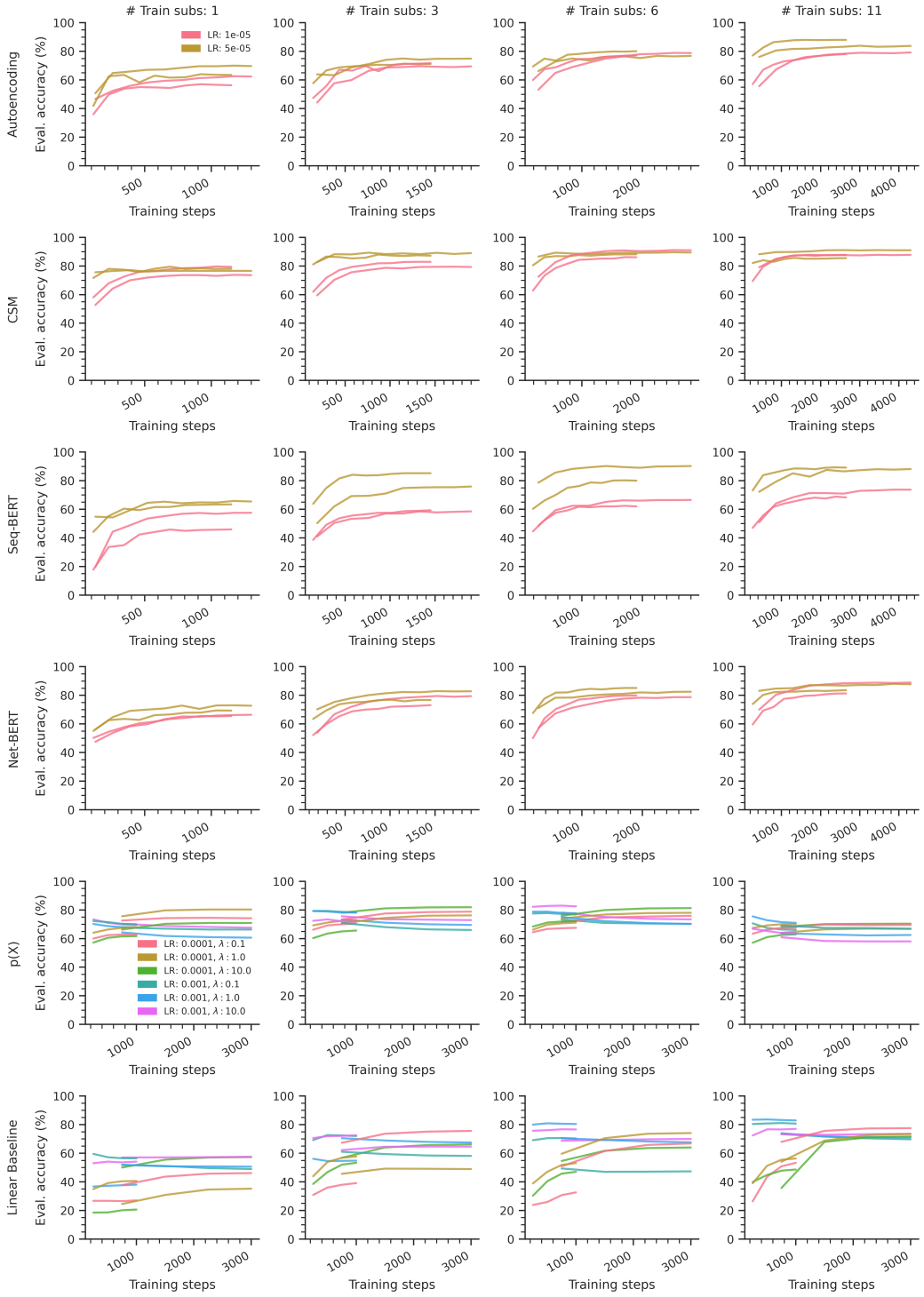

Figure B9: Replication of Appendix Fig. B7 with a different set of random seeds.