Dissecting EEG-Language Models: Token Granularity, Model Size, and Cross-Site Generalization

Anonymous Authors[1]

Abstract

We investigate how token granularity and model size affect EEG-language model performance in both in-distribution and cross-site scenarios. We pretrain a 1D ConvNeXt tokenizer (`Weaver` tokenizer) and use it to adapt `Qwen3` language models (`Weaver` model) with continual pretraining on the Big EEG (BEEG) dataset, the largest corpus of EEG, iEEG, and MEG data to date. Finetuning and evaluating on seizure detection and forecasting, we find that optimal scaling depends heavily on the task and whether evaluation is in-distribution. For seizure detection, finer tokenization consistently improves performance. On CHB-MIT and Siena, our models achieve balanced accuracy comparable to or exceeding state-of-the-art EEG foundation models. In contrast, cross-site seizure forecasting benefits significantly from coarser tokenization, challenging the assumption that higher fidelity is always better. While increasing language model size improves in-distribution detection, it offers no benefit for cross-site generalization. These results establish token granularity as a critical, task-dependent scaling dimension for clinical EEG models.

1. Introduction

Electroencephalography (EEG) is a non-invasive method for monitoring brain activity, and has become an increasingly important tool in real-world clinical settings for the monitoring and diagnosis of various neurological disorders (Schomer & Lopes da Silva, 2018). However, manual review of EEG signals can take away precious time and resources that can be otherwise spent on improving patient care. Reliable and automated assistive EEG signal analysis tools therefore have the potential to improve patient care and reduce physician workload (Smith, 2005).

Cross-site generalization is a blocker to deploying automated EEG analysis tools in clinical practice. Models trained at one hospital often show strong degradation in performance when applied to another hospital due to differences in data collection, such as acquisition equipment, patient demographics, and recording protocols. This is known as distribution shift, and it is one of the primary reasons that promising deep learning-based EEG analysis systems have not achieved widespread clinical adoption (Kostas et al., 2021; Roy et al., 2019).

Foundation models pretrained on diverse data have shown improved robustness to distribution shift in other domains, such as computer vision, medical imaging and speech recognition (Radford et al., 2022; Moor et al., 2023; Kirillov et al., 2023). A growing body of work interfaces pretrained language models with non-text modalities through discrete tokenization, including audio (Défossez et al., 2024) and images (Team, 2024), to produce a strong foundation model.

Previous works have explored this strategy in EEG modeling domain. NeuroLM (Jiang et al., 2024) trains a decoder-only language model with a text-aligned tokenizer. NeuroCogiter (Cong, 2025) uses a similar strategy, while adopting a more sophisticated multi-stream causal pretraining strategy. Both works incorporate text alignment steps for learning the tokenizer.

However, these works fix tokenization granularity early and focus optimization on the downstream language model, leaving the interaction between tokenizer design and model scaling unexplored. *We ask: how do tokenization granularity and language model size jointly affect downstream performance, and do gains achieved in-distribution transfer to cross-site generalization?*

We address this gap through a controlled study using an intentionally transparent design: a 1D ConvNeXt tokenizer trained only with MSE loss, paired with pretrained `Qwen3` language models adapted through standard continual pretraining. This transparency lets us isolate the effects of tokenization granularity and model size on both in-distribution and cross-site performance. Concretely, we train a 1D ConvNeXt-based tokenizer with a finite scalar quantiza-

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

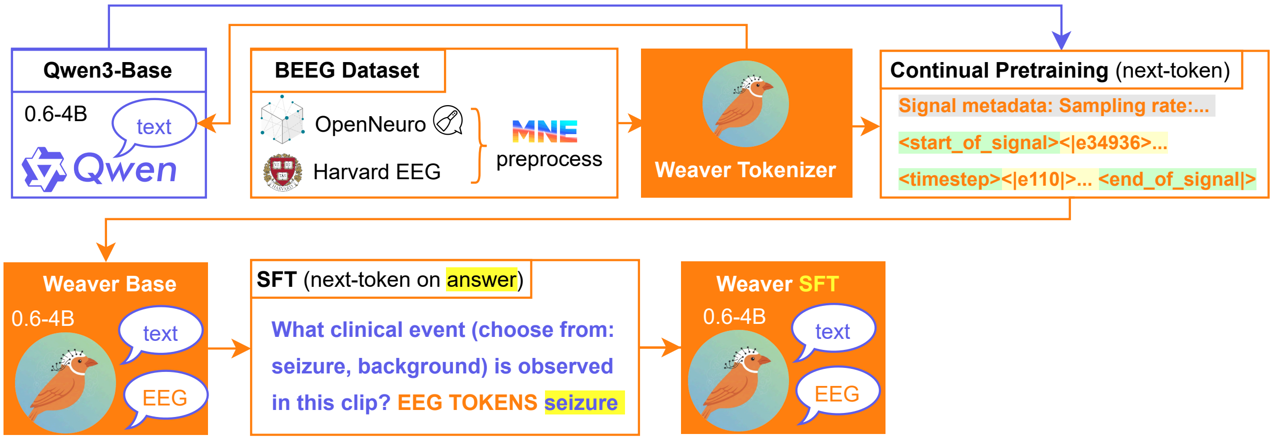
Qwen3-Base
0.6-4B
text
Qwen
BEEG Dataset
OpenNeuro
Harvard EEG
MNE
preprocess
Weaver Tokenizer
Continual Pretraining (next-token)
Signal metadata: Sampling rate:...
<start_of_signal><|e34936>...
<timestep><|e110|>... <end_of_signal|>
Weaver Base
0.6-4B
text
EEG
SFT (next-token on answer)
What clinical event (choose from: seizure, background) is observed in this clip? EEG TOKENS seizure
Weaver SFT
0.6-4B
text
EEG

Figure 1. Overview of our framework (`Weaver`). We construct the BEEG pretraining corpus by combining the Harvard EEG dataset with EEG, iEEG, and MEG recordings from OpenNeuro. The `Weaver` Tokenizer (a 1D ConvNeXt encoder-decoder with FSQ bottleneck) converts multi-channel EEG clips into discrete tokens, which are serialized with metadata and fed to a `Qwen3` language model for continual pretraining via next-token prediction. We finetune the resulting `Weaver` model on seizure detection and forecasting tasks using a question–signal–answer format, then evaluate under in-distribution and leave-one-site-out settings.

tion (FSQ) bottleneck at three temporal granularities (32, 64, and 128 EEG samples). We continually pretrain `Qwen3` base models (0.6B, 1.7B, and 4B parameters) for next token prediction leveraging 50 billion tokens on a large corpus (named BEEG) combining the Harvard EEG dataset (Zafar et al., 2025) and EEG, iEEG, MEG recordings from OpenNeuro. After continual pretraining, we finetune each language model on seizure detection and forecasting tasks by providing prompts and answers to the model using three clinical datasets: CHB-MIT (Guttag, 2010; Shoeb, 2009), Siena (Detti, 2020; Detti et al., 2020), and TUSZ (Shah et al., 2018). See Fig. 1 for the overall training pipeline.

To ensure that our conclusions about scaling trends are valid, we also compare our model's seizure detection performance against state of the art EEG models. We select the following state of the art EEG models as our baselines. BIOT (Yang et al., 2023) uses a transformer with full attention between EEG patches and masked reconstruction pretraining. `LaBraM` (Jiang et al., 2023) learns a tokenizer to produce discretized maksed reconstruction targets for pretraining a full attention transformer model. On the other hand, `CSBrain` (Zhou et al., 2025) utilizes a novel cross-scale attention mechanism and structured sparsity to avoid full attention between all patches. `CBraMod` (Wang et al., 2024) also uses a factorized attention approach (termed "criss-cross" attention in the paper) to decouple the dense attention pattern. `GT-STAFG` (Nafea & Ismail, 2025) on the other hand, uses a graph transformer (Shehzad et al., 2026) to process the spatio-temporal relations between EEG signal.

We evaluate under two regimes: (i) in-distribution, where training and test subjects come from the same sites, and (ii) leave-one-site-out, where models are trained on two sites and tested on the held-out third site. Our main findings are:

- **Detection scaling:** Finer tokenization consistently improves both in-distribution and cross-site seizure detection accuracy. In contrast, scaling model size only improves in-distribution performance. Larger models show no benefit, and sometimes degradation, on held-out sites.
- **Forecasting scaling:** Seizure forecasting shows a distinct pattern: in-distribution performance exhibits no reliable scaling trends with either tokenization or model size, while cross-site generalization benefits from *coarser* tokenization. This contrast suggests that detection and forecasting rely on different temporal scales of EEG features.
- **Representation analysis:** To understand these scaling differences, we analyze learned representations. Despite flattening multi-channel EEG into a 1D token sequence, representations remain robust to sensor permutation, with finer tokenization maintaining higher invariance. Fine-grained tokenization also produces disproportionately large weight updates in mid-network layers. This suggests finer tokenization learns distinct features from coarser ones.
- **High-performance seizure detection:** `Weaver` achieves seizure detection balanced accuracy comparable to or exceeding current state-of-the-art EEG foun-

dation models on CHB-MIT and Siena datasets, though we note differences in evaluation protocols.

To support reproducibility and future research, we release:

- **BEEG:** code for building the largest electrophysiology pretraining corpus to date (EEG, iEEG, MEG).
- **BEEGBench:** a package for creating reproducible EEG-language datasets for supervised finetuning (SFT).
- **Weaver model weights:** `Weaver` Tokenizer checkpoints at three granularities, `Weaver` continual-pretrained EEG-language models up to 4B parameters.

2. Methods

2.1. BEEG: Continual pretraining dataset

For the continual pretraining of the language model, we draw data from two large open datasets. For the Harvard EEG dataset (Zafar et al., 2025; Sun et al., 2025), to ensure high data quality, we only keep recordings with duration longer than 1 minute. The remaining data are split into train and validation split based on the patient ID, with data from 4 patients chosen randomly as the validation set. Data from remaining patients are used as the training set. Each recording is further split into non-overlapping clips of 10-30 second.

We further leverage all the EEG, iEEG, and MEG data from the OpenNeuro platform (Markiewicz et al., 2021). Since OpenNeuro data consists of multiple individual datasets, we select one subject from each sub-dataset to form the validation set. The remaining subjects are used as the training set. We split each recording into clips of 5-10 seconds.

As will be detailed in section 2.5, we flatten the signals from all electrodes in a clip into a 1D sequence consisting of tokens from all electrodes for the language model to process. Since the recordings from OpenNeuro usually have a larger electrode count, in order to maintain a similar total sequence length (in terms of the number of tokens), we elect to use shorter clips.

The resulting dataset, to be called the Big EEG (BEEG) dataset, consists of clips totaling approximately 2.75 million hours (Table 1). It is the largest and most diverse electrophysiology recording dataset to date. We will release the preprocessing script for generating this dataset from the Harvard and OpenNeuro datasets.

2.2. Evaluation datasets

The evaluation datasets for seizure detection and forecasting are collected from sites strictly outside of BEEG pretraining dataset to prevent data leakage. For each evaluation dataset, we randomly select 20% of the subjects as test subjects, and 10% of all subjects as validation subjects. The remaining subjects are used as the training set. Each recording is split into 5-15 second clips. In order to make evaluation result analysis easier, we apply a balanced sampling strategy that ensures each class (with and without seizure) is represented roughly equally in the training, validation, and test sets. Preprocessing of data is done according to section 2.3.

Table 1. BEEG dataset composition. Duration is reported in hours; channel statistics are per-recording.

	Duration	Num Channels	
Dataset	(hours)	Mean	Std
Harvard EEG	2,710,349	32.0	9.3
OpenNeuro EEG	38,919	56.6	58.1
OpenNeuro iEEG	3,247	107.8	42.6
OpenNeuro MEG	809	281.2	66.5
Total	**2,753,324**	—	—

For seizure forecasting, to produce positive samples (seizure forecasting with different future horizons $[T_1, T_2]$), for each EEG recording in an evaluation dataset, we identify locations with seizures based on seizure annotation, and sample a clip of 5-10 seconds long that is within the T_2 to T_1 minutes window before the seizure onset. We also randomly select non-seizure clips which do not have seizures within the next T_1 to T_2 duration as negative samples. We examine 5 prediction horizons, with $T_1 = 0, 10, 20, 30, 40, 50$ and $T_2 = T_1 + 10$. Preprocessing is identical as seizure detection.

Following the aforementioned process, we produce 50,000 training clips for seizure detection from each site, and 10,000 training clips for seizure forecasting from each site across all clip durations. For testing, each site generates additional 1,000 clips for seizure detection with durations of 5s, 10s, and 15s, respectively, or 3,000 testing clips in total. Similarly, we generate 3000 testing clips of different durations for seizure forcasting.

2.3. EEG data preprocessing

All EEG data are first high-pass filtered at 0.5 Hz to detrend, and then notch filtered at 50 Hz and 60 Hz to remove power line noise. Note that we apply both filters to unify the EEG data from different sites. We then apply channel-wise z-score normalization to standardize the amplitude of the EEG signal clips, with the mean and std derived from the clip itself.

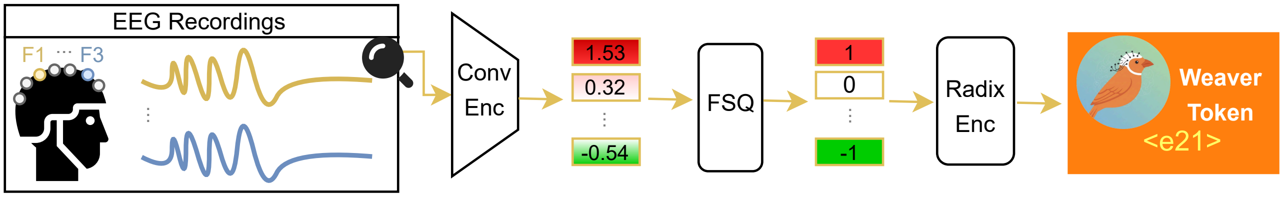

Figure 2. Weaver Tokenizer architecture. A single-channel EEG signal passes through wavelet downsampling, a ConvNeXt encoder, FSQ quantization, and radix encoding to produce discrete tokens. Details of each component are provided in 2.4.

2.4. Weaver Tokenizer

Since different montages for EEG exist, we design the Weaver Tokenizer to process the signal over a token duration from a single EEG channel at a time, thus deferring the duty of fusing information from all electrodes together to the latter language model. The weaver tokenizer converts single channel EEG signal into sequence of tokens. We provide an overview in figure 2.

To keep the tokenizer simple and efficient, the Weaver Tokenizer consists of an encoder, a finite scalar quantization (FSQ) bottleneck (Mentzer et al., 2023), and a decoder. Both the encoder and the decoder are built by stacking 1D ConvNeXt (Woo et al., 2023) blocks. The downsampling and upsampling are performed by an L-level wavelet-inspired transform from the stable audio tools library (Stability AI & Contributors) that achieves $2^L\times$ temporal compression using invertible biorthogonal wavelets (details in Appendix C). The FSQ bottleneck maps the input to a fixed codebook of size 65536 (2^{16}). The whole network is trained with mean squared error (MSE) between the input and reconstructed signal, and a straight-through-estimator (STE) is used to ensure differentiability of the discrete FSQ.

The tokenization process for a single EEG channel proceeds as follows: the raw signal is first downsampled by the wavelet module to reduce temporal resolution, then passed through the ConvNeXt encoder to produce a latent representation of shape $T \times H$, where T is the compressed time dimension and H is the latent dimension. FSQ quantizes each of the H latent dimensions independently into discrete levels. Since this produces multiple quantized values per timestep, we apply radix encoding to combine them into a single integer token per timestep: given quantized values $(q_1, \ldots, q_H)$ with $q_i \in \{0, \ldots, \ell_i - 1\}$ for level counts $(\ell_1, \ldots, \ell_H)$, the token is $\sum_{i=1}^{H} q_i \prod_{j=1}^{i-1} \ell_j$, yielding a vocabulary of size $\prod_{i=1}^{H} \ell_i = 65536$. Thus, each channel yields T tokens from an input window, with each token drawn from a vocabulary of 65536 possible values.

The tokenizer is trained on a uniform mixture of iEEG, EEG, and MEG data from our BEEG pretraining corpus, although in this work we only explored EEG-based tasks. For determining the optimal number of layers and hidden dimensions, we performed a hyperparameter search using NSGA-II multi-objective optimization (Deb et al., 2002) provided by the Optuna library (Akiba et al., 2019). We optimize the following two objectives: (1) the reconstruction loss, and (2) the encoding speed. That is, we aim to find the tokenizer that reconstructs the signal with minimal MSE while maintaining fast encoding speed. The search space we explored is detailed in appendix A and we list the optimal hyperparameters found in B. Each hyperparameter search is given a fixed time budget of 5 days on a single GPU. The tokenizer for each token length is fixed once trained and not finetuned during either continual pretraining or finetuning stage.

2.5. Weaver: Language model continual pretraining

We use the Qwen3 model family (Yang et al., 2025) as our base language model. We continually pretrain the 0.6B, 1.7B, and 4B base models to produce Weaver models.

To feed EEG signals to the language model, we serialize each multi-channel clip into a 1D token sequence. The sequence begins with a metadata header containing the sampling rate, number of sensors, data type (EEG, iEEG, or MEG), and an ordered list of sensor names. Following the header, we emit a <|start_of_signal|> token, then flatten the tokenized signal in time-major order: for each timestep, we concatenate the tokens from all sensors and append a <|timestep|> delimiter before advancing to the next timestep. The sequence concludes with an <|end_of_signal|> token. An example serialization is shown below:

```
Signal metadata:
Sampling rate: 512.0 Hz
Number of sensors: 16
Datatype: eeg
Sensor names: Fpz, T8, C3, ..., F7
<|start_of_signal|><|e34936|><|e296|>...
<|timestep|><|e110|><|e456|>...
<|end_of_signal|>
```

The embeddings for all EEG related special tokens are initialized with a normal distribution of mean=0, std=0.022 and updated during the continual pretraining, along with the original text embeddings and other model parameters.

Continual pretraining uses lr=10^{-4}, global batch size=512, max sequence length of 2000 tokens, and we train for 50 billion tokens in total. We clip the gradient norm to 1.0 to prevent gradient explosion, and apply a 500 step learning rate warmup from 0, and a cosine learning rate schedule that decays to 1e-5. To preserve the text capability of the language model, we interleave EEG with text data from DCLM (Li et al., 2024) in a 1:1 ratio throughout continual pretraining.

2.6. Weaver-SFT: Downstream tasks finetuning

We treat the seizure detection and seizure forecasting tasks (under different prediction horizons) as the next-token-prediction objective on the answer tokens, and finetune the Weaver models jointly for both tasks to produce Weaver-SFT. Each training example follows a question–signal–answer format: we first present a text question (e.g., “What clinical event is observed in this clip?”), followed by the serialized EEG clip as described above, and finally the target answer (e.g., “seizure” or “background”). The model is trained to predict only the answer tokens given the preceding context. The full prompts used for finetuning are detailed in appendix E. We use a global batch size of 8 with max sequence length of 16384 tokens with sequence packing to finetune the language model weight, while keeping the tokenizer frozen. All finetuning uses 1000 gradient update steps. We run 10 rounds of hyperparameter searches for each model to search for the optimal learning rate and weight decay. The hyper-parameter search space for finetuning is detailed in appendix D.

During evaluations, we simplified the forecasting task into a binary classification problem by pooling outcomes: seizure within 30 minutes vs. no seizure within 30 minutes (combining later seizures and background).

3. Results

More granular tokenization shows better reconstruction performance across frequency bands We first evaluate the reconstruction error of the tokenizer in different frequency bands, shown in Figure 3. It shows that 32 samples/token provides the lowest reconstruction error throughout the frequency range.

In-distribution effects of model settings We quantify in-distribution performance using binomial generalized estimating equations (GEE) to regress trial-level correctness on model size, token granularity, pretraining steps, and clip duration (see Tables 3 and 4 in Appendix F for more details). The results are summarized in Figure 4. We report odds ratios (OR), which quantify how the odds of a correct prediction change with each predictor: an OR of 1.02 for a variable means a one-unit increase in that variable is associated with 2% higher odds of correctness, while an OR below 1 indicates lower odds. For log-transformed predictors (model size, samples/token, clip duration), the OR represents the effect of doubling that variable.

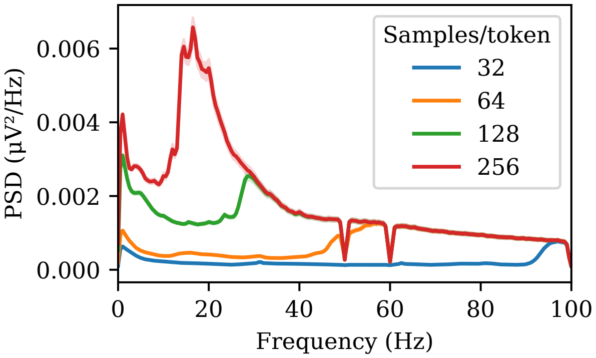

Figure 3. Reconstruction error for different frequency bands for `weaver` tokenizers with different granularities. Lower is better.

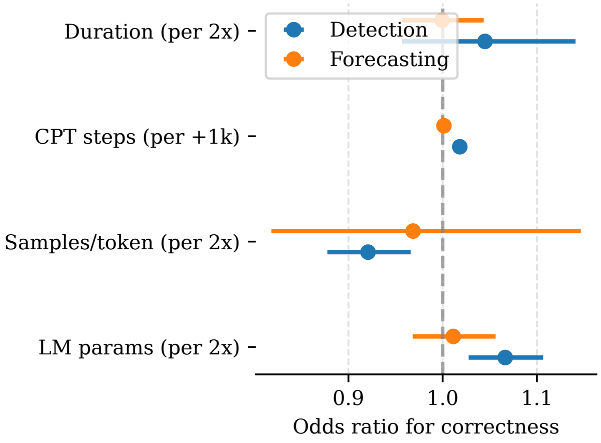

Figure 4. How input duration, continual pretraining step, tokenizer granularity, and language model size affect in-distribution task performances, measured by average odds ratio. Odds ratio > 1 indicates an increase in correct prediction rate, and < 1 indicates a decrease.

Seizure detection Seizure detection (N=311,385) shows clear benefits from scaling. Both larger model size (p=0.001; $\approx$ 1.6% gain per doubling) and finer token granularity (p=0.001; $\approx$ 2.1% gain per halving samples-per-token) significantly improve performance, with no significant difference in their effect magnitudes (Wald test p=0.58). Continual pretraining also yields consistent gains ($p<10^{-5}$; $\approx 0.5\%$ per 1k steps), while clip duration shows no significant effect (p=0.328).

Seizure forecasting In contrast, seizure forecasting (N=314,940) exhibits no reliable scaling trends. Neither model size (p=0.614), token granularity (p=0.711), pretraining steps (p=0.312), nor clip duration (p=0.976) show significant associations with correctness, suggesting a fundamentally different scaling behavior compared to detection.

Cross-site effects of model settings We evaluate cross-site generalization using a leave-one-site-out protocol, fitting a binomial GEE (Figure 5; details in Tables 5 and 6 in Appendix F).

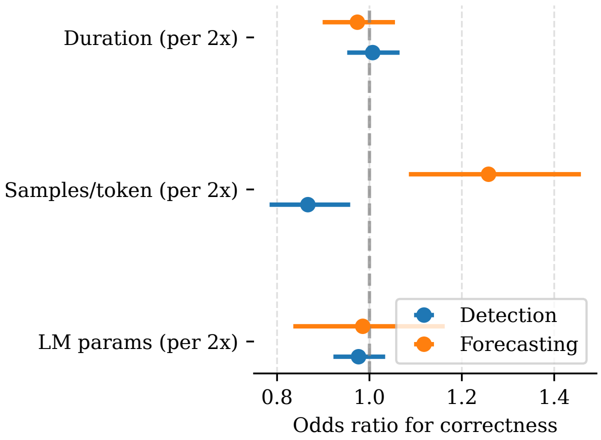

Figure 5. How input duration, continual pretraining step, tokenizer granularity, and model size affect cross-site task performances of variants of `Weaver-SFT`, measured by average odds ratio. Odds ratio >1 indicates an increase in correct prediction rate, and <1 indicates a decrease.

Seizure detection For cross-site seizure detection (N=62,277), tokenizer granularity remains the only significant predictor: more granular tokenization significantly improves performance (p=0.005; $\approx 3.6\%$ gain per halving samples-per-token). Notably, scaling language model size does not improve cross-site accuracy (p=0.415), suggesting that larger models do not necessarily generalize better to unseen hospitals. The clip duration was also not significant (p=0.809).

Seizure forecasting Cross-site seizure forecasting (N=62,988) reveals a reversal of the granularity trend: coarser representations significantly improve performance (p=0.002; $\approx 5.7\%$ gain per doubling samples-per-token). Neither model size (p=0.863) nor clip duration (p=0.517) showed significant effects.

Model representation is robust to sensor permutation One concern with flattening EEG tokens to a 1D sequence is that the order in which sensors are flattened may cause the model to produce representations that are highly permutation-dependent. We evaluate the degree to which the models are permutation-invariant by shuffling the order of sensors in the input and measuring similarity to a prototype representation. Specifically, for each EEG sample, we compute a prototype as the mean embedding across multiple random sensor orderings. We then measure permutation robustness as the average cosine similarity between each permuted sample's representation and this prototype; high similarity indicates the model produces consistent representations regardless of sensor order. The results are shown in figure 6. It shows that for all token granularities, the correlation between permuted and unpermuted representations is high (all ρ>0.999), indicating that although in principle the order of sensors in the flattened sequence may affect the model's output, the effect is minimal in practice. Specifically, the more granular tokenizers, 32 samples/token and 64, show a consistently high permutation invariance throughout the continual pretraining, whereas the coarser 128 samples/token tokenizer shows a gradual decline in permutation invariance as the model is trained for more steps.

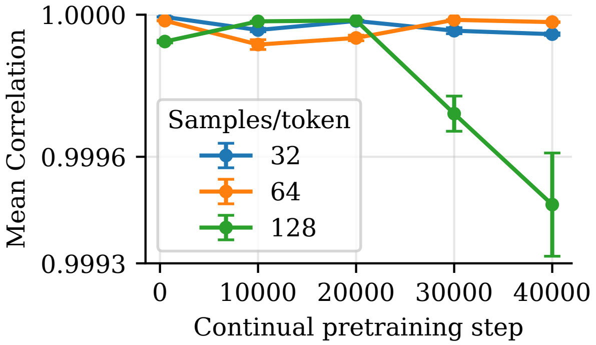

Figure 6. Permutation variance of `Weaver-base-0.6B` of different token granularity as measured by mean correlation between mean-pooled representations of samples with permuted sensor order. Higher is more invariant.

Fine-grained tokenization leads to larger mid-network layer updates Next, we would like to understand whether different tokenization granularities would show different learning dynamics. We evaluate the relative change in attention weights in each layer of the 0.6B language model after continual pretraining and plot the results in figure 7. It shows that for layers 0-10 and layers 20-28, the three types of tokenization show qualitatively similar patterns. However, for the finest tokenization, 32 samples/token, a prominent difference is visible from layer 10-20. Specifically, the attention weights from layer 10-20 show a larger relative change than the other two tokenization schemes.

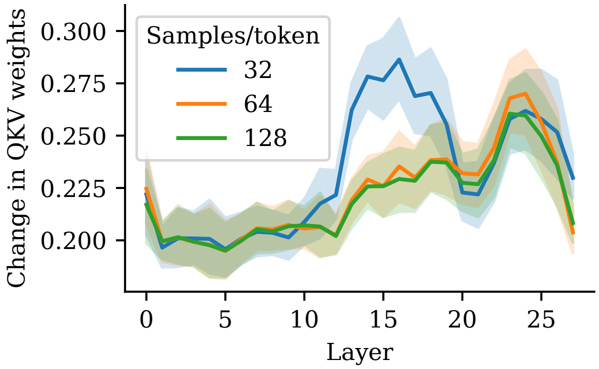

Figure 7. Relative weight change of attention weights in each layer of the 0.6B model after continual pretraining (`Weaver-Base`) compared to the initial base weights (`Qwen`).

Weaver achieves comparable or better seizure detection performance than state-of-the-art EEG foundation models We compare our model with other state-of-the-art EEG foundation models on the seizure detection task. The results are shown in Table 2. Our `Weaver` models (under the best parameter setting for in-distribution) achieve state-of-the-art results on CHB-MIT and Siena datasets as measured by balanced accuracy (BAcc), significantly outperforming existing EEG foundation models performance as reported in the CSBrain benchmark. Specifically, our 4B parameter model achieves 0.893 BAcc on CHB-MIT and our 1.7B model achieves 0.839 BAcc on Siena, surpassing the previous bests of 0.740 and 0.766 respectively. On TUSZ, `Weaver` demonstrates competitive performance (0.765 BAcc), comparable to strong baselines like BIOT (0.784), though trailing behind specialized models like GT-STAFG (0.835). To the best of our knowledge, there are no public reports of seizure forecasting models that we can compare with.

4. Discussion

Our study investigates the scaling laws of EEG foundation model size and tokenizer granularity, revealing a surprising decoupling between representation fidelity and model capacity. We establish that token granularity, not model size, is the primary driver of cross-site generalization for seizure detection. This challenges the direct transfer of text-based scaling laws to continuous modalities, suggesting that for biological signals, a more nuanced scaling law might be in place. While larger models fit in-distribution data better, they fail to generalize to new hospitals absent a high-fidelity representation of the underlying signal dynamics.

Table 2. Cross-subject segment-level seizure detection balanced accuracy (BAcc; higher is better. Best peformance bolded). Ours use 15s clips, while CHB-MIT and Siena baselines follow the CSBrain benchmark split (10s windows). TUSZ baselines follow GT-STAFG's evaluation setup.

Model	CHB-MIT	Siena	TUSZ
Ours			
`Weaver` (64k_64, 0.6B)	0.764	0.802	0.686
`Weaver` (64k_64, 1.7B)	0.741	**0.839**	0.689
`Weaver` (64k_64, 4B)	**0.893**	0.796	0.765
Baselines			
CSBrain	0.726	0.766	–
CBraMod	0.740	0.732	–
LaBraM	0.708	0.708	–
BIOT	0.707	0.735	0.784
GT-STAFG	–	–	**0.835**

In our cross-site generalization experiments, we find that higher token granularity improves detection performance, but it will decrease forecasting performance. We hypothesize that seizure detection requires features that are higher frequency in nature, whereas forecasting might rely on slower state dynamics that are difficult to capture from token to token and thus benefit from using a coarser tokenizer.

Mid-network layer updates are particularly pronounced for fine-grained tokenization, which is a unique pattern among the different tokenization granularity we considered. This could explain why scaling tokenization has a reliable effect on downstream performance even when evaluated cross-site. It causes more fundamental changes in model weights (and in turn representations), whereas coarser tokenization might be learning shallower features.

We note that our performance comparison with previous state of the art models (Table 2) is not directly comparable, as factors such as clip length and test set selection are different. Thus, they only serve to show that our model can produce seizure detection performance comparable to previous state-of-the-art models, without claiming our model is strictly better. This also highlights the importance of a unified platform for evaluating EEG models.

`Weaver` still trails behind the state-of-the-art `GT-STAFG` on TUSZ performance, despite having a significantly larger model size (refer to Table 2). A plausible explanation of this is that `Weaver` is trained on both seizure detection and seizure forecasting tasks across 3 different datasets, while `GT-STAFG`'s TUSZ performance is obtained by training only on TUSZ for seizure detection. Future studies on how model expressivity interacts with multi-task training might help improve our understanding of this gap.

We established that for in-distribution and cross-site seizure detection, token granularity is an important factor that can be used to improve model performance, and sometimes even stronger than scaling the language model size. Similar conclusions for tasks such as image generation in computer vision have been reached where a strong image tokenizer can be used to match performance of continuous input models (Yu et al., 2023; You et al., 2025). One future direction is to explore ways to learn strong representation in the tokenizer beyond token granularity. Methods such as masked reconstruction (He et al., 2021; Fu et al., 2024) and self-distillation has been very successful methods for learning strong representations in computer vision (Siméoni et al., 2025; Balestriero & LeCun, 2025), and EEG tokenizer learning can incorporate representation steps inspired by them.

Limitations Our study has several limitations that suggest directions for future work:

1. We evaluate cross-site generalization using only three clinical sites (CHB-MIT, Siena, TUSZ); the observed scaling patterns may not hold across a more diverse set of hospitals with greater variation in equipment, protocols, and patient populations.
2. In order to study cross-site generalization, we limited our scope to seizure detection and forecasting tasks. It is possible that other downstream tasks may not benefit from tokenization granularity scaling.
3. Although our tokenizer is trained on EEG, iEEG, and MEG data, we only evaluate downstream performance on scalp EEG tasks. How well our findings transfer to intracranial or magnetoencephalography is unclear.
4. To keep the computation cost manageable, we fixed the tokenizer vocabulary size to 64k. Future studies should also explore how tokenizer vocabulary size interacts with model size and task performance.
5. Our language model experiments use the `Qwen3` family exclusively. Other architectures, such as mixture of experts (MoE) or model families may show different scaling behaviors.
6. We do not explore hybrid or adaptive tokenization strategies that could potentially balance the competing demands of detection and forecasting tasks.

Impact Statement

In this work, we showed that tokenization granularity can be a more reliable way to scale the cross-site performance of EEG analysis models. This provides rigorous evidence that smaller, well-tokenized models can match the efficacy of larger ones, which can significantly reduce the carbon footprint of 24/7 seizure monitoring and democratizing access to high-performance AI in resource-constrained clinical settings.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330701.

Balestriero, R. and LeCun, Y. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv [cs.LG]*, November 2025. doi: 10.48550/arXiv.2511.08544.

Cong, M. NeuroCognitor: Unified EEG-language framework for cognitive load analysis via instruction-tuned multi-task learning. *IEEE Access*, 13(99):201645–201665, January 2025. ISSN 2169-3536. doi: 10.1109/access.2025.3637315.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2):182–197, April 2002. ISSN 1089-778X,1941-0026. doi: 10.1109/4235.996017.

Detti, P. Siena scalp EEG database. *PhysioNet*, August 2020. doi: 10.13026/5d4a-j060.

Detti, P., Vatti, G., and Zabalo Manrique de Lara, G. EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes (Basel)*, 8 (7):846, July 2020. ISSN 2227-9717,2227-9717. doi: 10.3390/pr8070846.

Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. Moshi: a speech-text foundation model for real-time dialogue. *arXiv [eess.AS]*, September 2024.

Fu, L., Lian, L., Wang, R., Shi, B., Wang, X., Yala, A., Darrell, T., Efros, A. A., and Goldberg, K. Rethinking patch dependence for masked autoencoders. *arXiv [cs.CV]*, January 2024.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., and Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, June 2000. ISSN 1524-4539,0009-7322. doi: 10.1161/01.cir.101.23.e215.

Guttag, J. CHB-MIT scalp EEG database. *PhysioNet*, June 2010. doi: 10.13026/C2K01R.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv [cs.CV]*, November 2021.

Jiang, W., Zhao, L., and Lu, B.-L. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, October 2023.

Jiang, W., Wang, Y., Lu, B.-L., and Li, D. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. *arXiv [cs.CV]*, April 2023.

Kostas, D., Aroca-Ouellette, S., and Rudzicz, F. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Front. Hum. Neurosci.*, 15:653659, June 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.653659.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., Bitton, Y., Nezhurina, M., Abbas, A., Hsieh, C.-Y., Ghosh, D., Gardner, J., Kilian, M., Zhang, H., Shao, R., Pratt, S., Sanyal, S., Ilharco, G., Daras, G., Marathe, K., Gokaslan, A., Zhang, J., Chandu, K., Nguyen, T., Vasiljevic, I., Kakade, S., Song, S., Sanghavi, S., Faghri, F., Oh, S., Zettlemoyer, L., Lo, K., El-Nouby, A., Pouransari, H., Toshev, A., Wang, S., Groeneveld, D., Soldaini, L., Koh, P. W., Jitsev, J., Kollar, T., Dimakis, A. G., Carmon, Y., Dave, A., Schmidt, L., and Shankar, V. DataComp-LM: In search of the next generation of training sets for language models. *arXiv [cs.LG]*, June 2024.

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncalves, M., Jwa, A., and Poldrack, R. A. OpenNeuro: An open resource for sharing of neuroimaging data. *bioRxiv*, June 2021. doi: 10.1101/2021.06.28.450168.

Mentzer, F., Minnen, D., Agustsson, E., and Tschannen, M. Finite scalar quantization: VQ-VAE made simple. *arXiv [cs.CV]*, September 2023.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023. ISSN 0028-0836,1476-4687. doi: 10.1038/s41586-023-05881-4.

Nafea, M. S. and Ismail, Z. H. GT-STAFG: Graph transformer with spatiotemporal attention fusion gate for epileptic seizure detection in imbalanced EEG data. *AI (Basel)*, 6(6):120, June 2025. ISSN 2673-2688,2673-2688. doi: 10.3390/ai6060120.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. *arXiv [eess.AS]*, December 2022.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. Deep learning-based electroencephalography analysis: a systematic review. *arXiv [cs.LG]*, January 2019.

Schomer, D. L. and Lopes da Silva, F. H. (eds.). *Niedermeyer's electroencephalography: Basic principles, clinical applications, and related fields*. Oxford University Press, New York, NY, 7 edition, March 2018. ISBN 9780190228484,9780190228484. doi: 10.1093/med/9780190228484.001.0001.

Shah, V., von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., Obeid, I., and Picone, J. The temple university hospital seizure detection corpus. *Front. Neuroinform.*, 12:83, November 2018. ISSN 1662-5196. doi: 10.3389/fninf.2018.00083.

Shehzad, A., Xia, F., Abid, S., Peng, C., Yu, S., Zhang, D., and Verspoor, K. Graph transformers: A survey. *IEEE Trans. Neural Netw. Learn. Syst.*, PP, January 2026. ISSN 2162-237X,2162-2388. doi: 10.1109/TNNLS.2025.3646122.

Shoeb, A. H. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. DINOv3. *arXiv [cs.CV]*, August 2025. doi: 10.48550/arXiv.2508.10104.

Smith, S. J. M. EEG in the diagnosis, classification, and management of patients with epilepsy. *J. Neurol. Neurosurg. Psychiatry*, 76 Suppl 2(suppl_2):ii2–7, June 2005. ISSN 0022-3050,1468-330X. doi: 10.1136/jnnp.2005.069245.

Stability AI and Contributors. stable-audio-tools: Generative models for conditional audio generation. GitHub repository.

Sun, C., Jing, J., Turley, N., Alcott, C., Kang, W.-Y., Cole, A. J., Goldenholz, D. M., Lam, A., Amorim, E., Chu, C., Cash, S., Junior, V. M., Gupta, A., Ghanta, M., Nearing, B., Nascimento, F. A., Struck, A., Kim, J., Sartipi, S., Tauton, A.-M., Fernandes, M., Sun, H., Bayas, G., Gallagher, K., Wagenaar, J. B., Sinha, N., Lee-Messer, C., Silvers, C. T., Gunapati, B., Rosand, J., Peters, J., Loddenkemper, T., Lee, J. W., Zafar, S., and Westover, M. B. Harvard electroencephalography database: A comprehensive clinical electroencephalographic resource from four boston hospitals. *Epilepsia*, 66(9):3411–3425, September 2025. ISSN 1528-1167,0013-9580. doi: 10.1111/epi.18487.

Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405. 09818*, 2024. doi: 10.48550/arXiv.2405.09818.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., and Pan, G. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.01548.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. *arXiv [cs.CL]*, May 2025. doi: 10.48550/arXiv.2505.09388.

Yang, C., Brandon Westover, M., and Sun, J. BIOT: Biosignal transformer for cross-data learning in the wild. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023.

You, Z., Ou, J., Zhang, X., Hu, J., Zhou, J., and Li, C. Effective and efficient masked image generation models. *arXiv [cs.CV]*, March 2025.

Yu, L., Lezama, J., Gundavarapu, N. B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., Hauptmann, A. G., Gong, B., Yang, M.-H., Essa, I., Ross, D. A., and Jiang, L. Language model beats diffusion – tokenizer is key to visual generation. *arXiv [cs.CV]*, October 2023.

Zafar, S., Loddenkemper, T., Lee, J. W., Cole, A., Goldenholz, D., Peters, J., Lam, A., Amorim, E., Chu, C., Cash, S., Moura Junior, V., Gupta, A., Ghanta, M., Fernandes, M., Sun, H., Jing, J., and Westover, M. B. Harvard electroencephalography database, 2025.

Zhou, Y., Wu, J., Ren, Z., Yao, Z., Lu, W., Peng, K., Zheng, Q., Song, C., Ouyang, W., and Gou, C. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. *arXiv [cs.HC]*, June 2025.

A. Tokenizer Hyperparameter Search Space

1. Hidden dimension of the encoder: {32, 64, 128, . . . , 512}
2. Hidden dimension of the decoder: {32, 64, 128, . . . , 512}
3. Number of ConvNeXt layers in the encoder: {1, . . . , 8}
4. Number of ConvNeXt layers in the decoder: {1, . . . , 8}
5. ConvNeXt block kernel size in the encoder: {3, 5, 7}
6. ConvNeXt block kernel size in the decoder: {3, 5, 7}
7. Pre-FSQ layernorm: {True, False}
8. MLP ratio in the encoder: uniform$(1, 4)$
9. MLP ratio in the decoder: uniform$(1, 4)$
10. Wavelet kernel type: {bior2.2, bior2.4, bior2.6, bior2.8, bior4.4, bior6.8}
11. Whether the wavelet kernel can be updated: {True, False}
12. How FSQ's 64k codebook is decomposed: One example is (4, 4, 4, 4, 4, 4, 4, 4), which means the codebook is decomposed into 8 groups, each with 4 levels.

B. Optimal hyperparameter for tokenizers

B.1. 32 Samples/token

```
{
  "fsq_levels": [4, 4, 4, 4, 4, 8, 8],
  "enc_dim": 416,
  "enc_layers": 7,
  "dec_dim": 192,
  "dec_layers": 4,
  "wavelet": "bior2.2",
  "wavelet_levels": 5,
  "enc_conv_kernel": 5,
  "enc_expansion": 3.758881181594571,
  "dec_conv_kernel": 5,
  "dec_expansion": 2.3783779598934514,
  "learnable_wavelet_kernel": true
}
```

B.2. 64 Samples/token

```
{
  "levels": [4, 4, 4, 4, 4, 8, 8],
  "enc_dim": 352,
  "enc_layers": 4,
  "dec_dim": 896,
  "dec_layers": 3,
  "wavelet": "bior4.4",
  "wavelet_levels": 6,
  "enc_conv_kernel": 3,
  "enc_expansion": 1.7106679127356494,
  "dec_conv_kernel": 7,
  "dec_expansion": 2.5110991475686695,
  "pre_fsq_norm": true,
  "learnable_wavelet_kernel": true
}
```

B.3. 128 samples/token

```
{
  "levels": [4, 4, 4, 4, 4, 4, 4, 4],
  "enc_dim": 256,
  "enc_layers": 4,
  "dec_dim": 512,
  "dec_layers": 3,
  "wavelet": "bior2.4",
  "wavelet_levels": 7,
  "enc_conv_kernel": 3,
  "enc_expansion": 3.3398946112629897,
  "dec_conv_kernel": 5,
  "dec_expansion": 1.2843996614857947,
  "pre_fsq_norm": true,
  "learnable_wavelet_kernel": true
}
```

B.4. 256 samples/token

```
{
  "levels": [4, 4, 4, 4, 4, 8, 8],
  "enc_dim": 128,
  "enc_layers": 1,
  "dec_dim": 576,
  "dec_layers": 4,
  "wavelet": "bior4.4",
  "wavelet_levels": 8,
  "enc_conv_kernel": 3,
  "enc_expansion": 2.832793423837936,
  "dec_conv_kernel": 5,
  "dec_expansion": 2.2663301595389265,
  "pre_fsq_norm": false,
  "learnable_wavelet_kernel": true
}
```

C. Wavelet Downsampling and Upsampling

We use an L-level wavelet-inspired transform for downsampling that achieves $2^L\times$ temporal compression. Let $x \in \mathbb{R}^{C\times T}$ denote an input with C channels and T time steps. At each level ℓ, we partition x into the first channel $x_1 \in \mathbb{R}^{1\times T}$ and the remaining channels $x_{\text{rest}} \in \mathbb{R}^{(C-1)\times T}$. Let $\tilde{h}$ and $\tilde{g}$ denote the analysis low-pass and high-pass filters of a biorthogonal wavelet. The transform computes:

$$a = (x_1 * \tilde{h}) \downarrow 2, \quad d = (x_1 * \tilde{g}) \downarrow 2 \tag{1}$$

$$x'_{\text{rest}} = \text{reshape}(x_{\text{rest}}, [2(C-1), T/2]) \tag{2}$$

where $*$ denotes convolution and $\downarrow 2$ denotes stride-2 subsampling. The first channel is decomposed into approximation (a) and detail (d) coefficients via filtering, while the remaining channels are reshaped to halve the temporal dimension and double the channel count without filtering. The output $[a; d; x'_{\text{rest}}] \in \mathbb{R}^{2C\times T/2}$ becomes the input for the next level, with a treated as x_1. The process starts with the entire input signal as x_1.

For upsampling, the inverse transform reconstructs the signal using the synthesis filters h and g:

$$x_1 = (a \uparrow 2) * h + (d \uparrow 2) * g \tag{3}$$

$$x_{\text{rest}} = \text{reshape}(x'_{\text{rest}}, [(C-1), T]) \tag{4}$$

where $\uparrow 2$ denotes upsampling by zero-insertion. We start with the first two input channels as h and g, and merge them into the first channel of the next stage, with remaining channels come from x_{rest}. We then repeat on the first two channels in the next stage. We use biorthogonal wavelets (e.g., bior4.4), which satisfy the perfect reconstruction property. The filter coefficients are optionally learnable during training.

D. Finetuning Hyperparameter Search Space

1. Learning rate: loguniform(1e-6, 1e-4)
2. Weight decay: uniform(0, 0.3)

E. Finetuning Prompt

E.1. Seizure detection

```
prompt: "What clinical event (choose from: seizure, background) is observed in this clip?"

completion: "seizure" or "background"
```

E.2. Seizure forecasting

```
prompt: "Forecast the next seizure (choose from: {{unique_events_str | join(’, ’)}})."

completion: "{{event_name}}"
```

The possible events considered for finetuning in our work is: {“0-10 minute(s)”, “10 minute(s)-20 minute(s)”, “20 minute(s)-30 minute(s)”, “30 minute(s)-40 minute(s)”, “40 minute(s)-50 minute(s)”, “50 minute(s)-1 hour(s)”, “$\geq$ 1 hour(s) or no future event”}.

F. Detailed Statistical Results

We fitted binomial Generalized Estimating Equations (GEE) with an exchangeable working correlation structure and robust standard errors to assess the effects of model settings on correctness. The following tables present the detailed regression results.

Table 3. In-distribution Seizure Detection GEE Results ($N = 311,385$, 118 clusters (GEE treats each subject as a cluster to correct scoring biases caused by correlation from multiple samples from the same subject.)). Log-transformed variables (Model size, Samples/token, clip duration) represent the change in log-odds per doubling of the predictor.

PREDICTOR	COEF.	STD. ERR.	z	$P > \|z\|$	[0.025	0.975]
INTERCEPT	-0.9415	0.540	-1.744	0.081	-2.000	0.117
$\log_2$(PARAMETER COUNT)	0.0642	0.019	3.395	0.001	0.027	0.101
$\log_2$(SAMPLES/TOKEN)	-0.0825	0.025	-3.367	0.001	-0.131	-0.034
CPT STEPS (K)	0.0180	0.003	5.853	0.000	0.012	0.024
$\log_2$(CLIP DURATION)	0.0439	0.045	0.978	0.328	-0.044	0.132

G. Model performance under different settings

Table 4. In-distribution Seizure Forecasting GEE Results ($N = 314,940$, 70 clusters).

PREDICTOR	COEF.	STD. ERR.	z	$P > \|z\|$	[0.025	0.975]
INTERCEPT	0.2460	0.656	0.375	0.708	-1.040	1.532
$\log_2$(PARAMETER COUNT)	0.0112	0.022	0.504	0.614	-0.032	0.055
$\log_2$(SAMPLES/TOKEN)	-0.0320	0.086	-0.371	0.711	-0.201	0.137
CPT STEPS (K)	0.0013	0.001	1.011	0.312	-0.001	0.004
$\log_2$(CLIP DURATION)	-0.0007	0.022	-0.030	0.976	-0.044	0.043

Table 5. Cross-site Seizure Detection GEE Results ($N = 62,277$, 118 clusters).

PREDICTOR	COEF.	STD. ERR.	z	$P > \|z\|$	[0.025	0.975]
INTERCEPT	1.9279	1.102	1.750	0.080	-0.232	4.088
$\log_2$(PARAMETER COUNT)	-0.0239	0.029	-0.816	0.415	-0.081	0.033
$\log_2$(SAMPLES/TOKEN)	-0.1430	0.051	-2.789	0.005	-0.244	-0.043
$\log_2$(CLIP DURATION)	0.0069	0.029	0.241	0.809	-0.049	0.063

Table 6. Cross-site Seizure Forecasting GEE Results ($N = 62,988$, 70 clusters).

PREDICTOR	COEF.	STD. ERR.	z	$P > \|z\|$	[0.025	0.975]
INTERCEPT	-1.1350	2.453	-0.463	0.644	-5.943	3.674
$\log_2$(PARAMETER COUNT)	-0.0146	0.084	-0.173	0.863	-0.180	0.151
$\log_2$(SAMPLES/TOKEN)	0.2294	0.075	3.048	0.002	0.082	0.377
$\log_2$(CLIP DURATION)	-0.0265	0.041	-0.648	0.517	-0.107	0.054

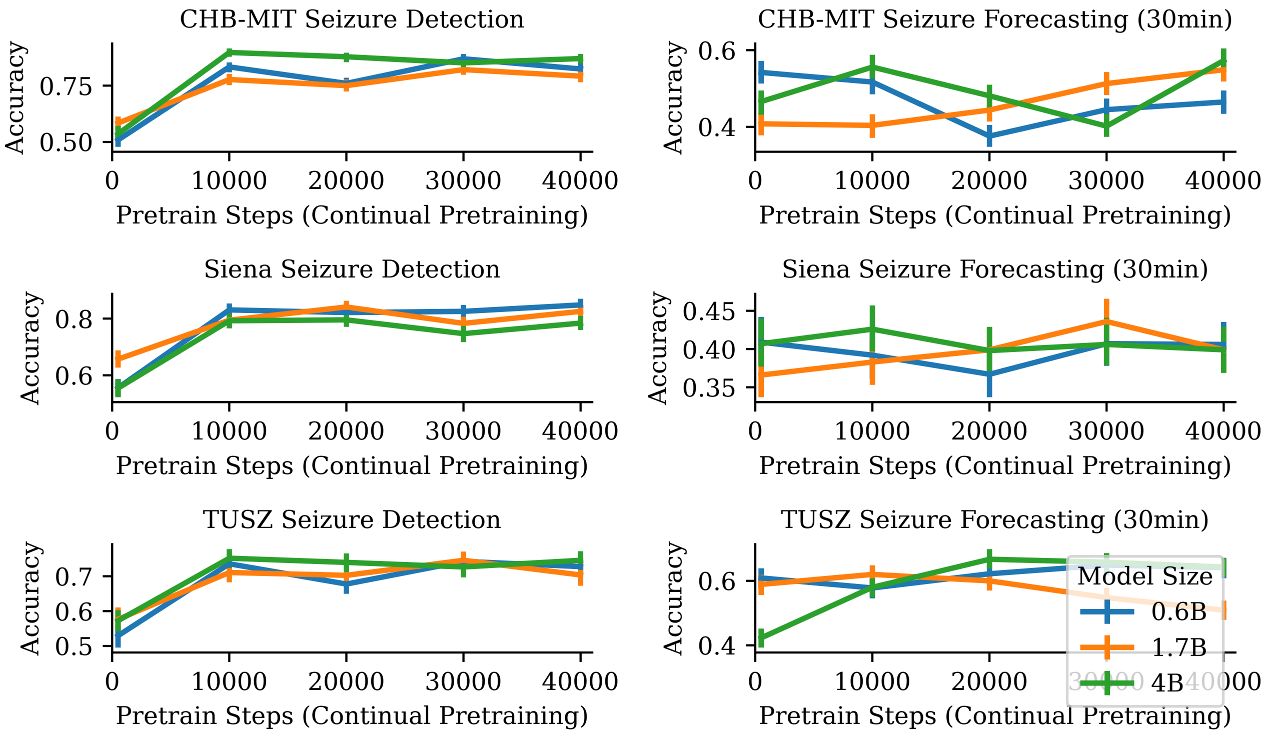

Figure 8. In-distribution model performance for 32 samples/token

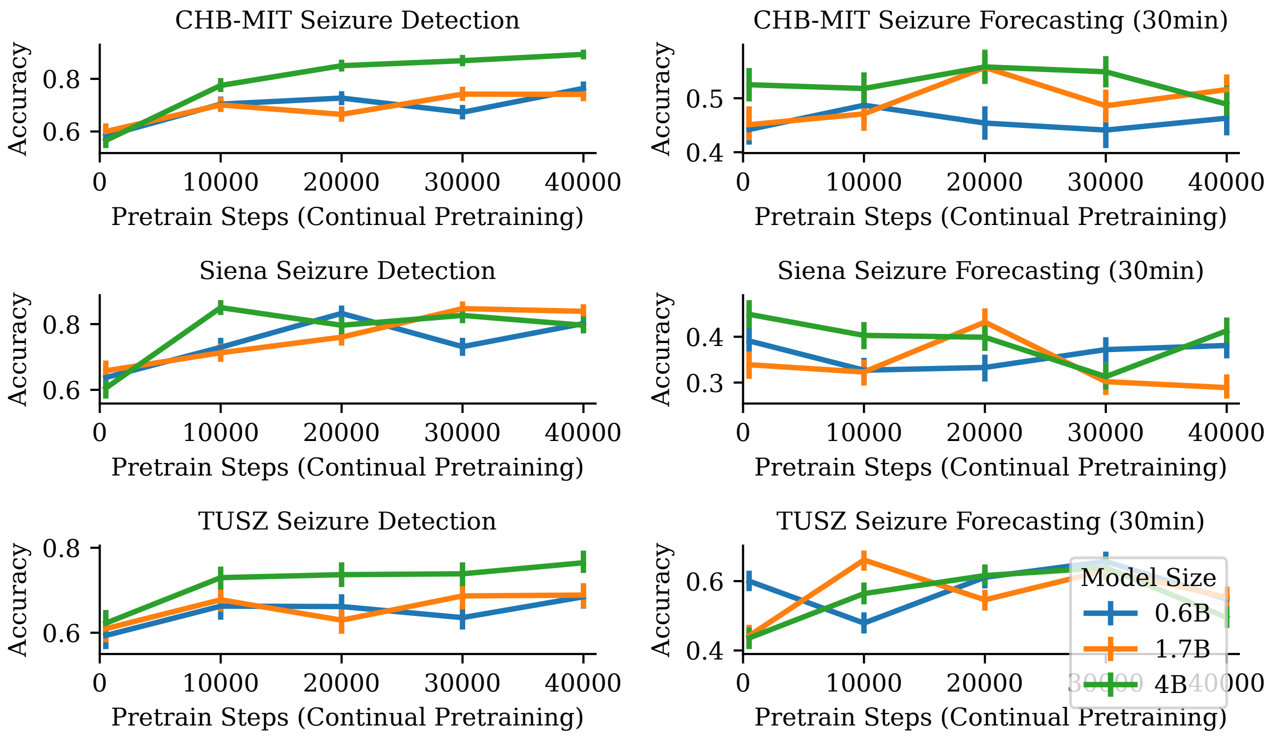

Figure 9. In-distribution model performance for 64 samples/token

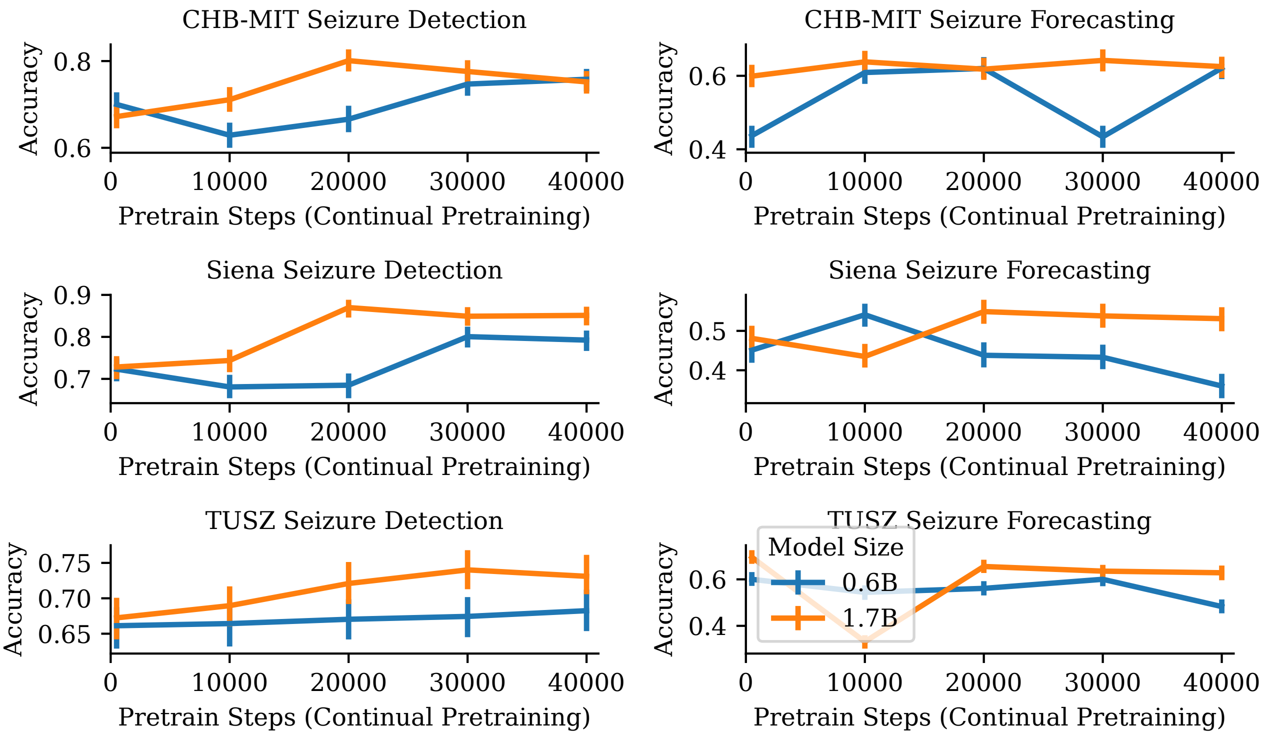

Figure 10. In-distribution model performance for 128 samples/token

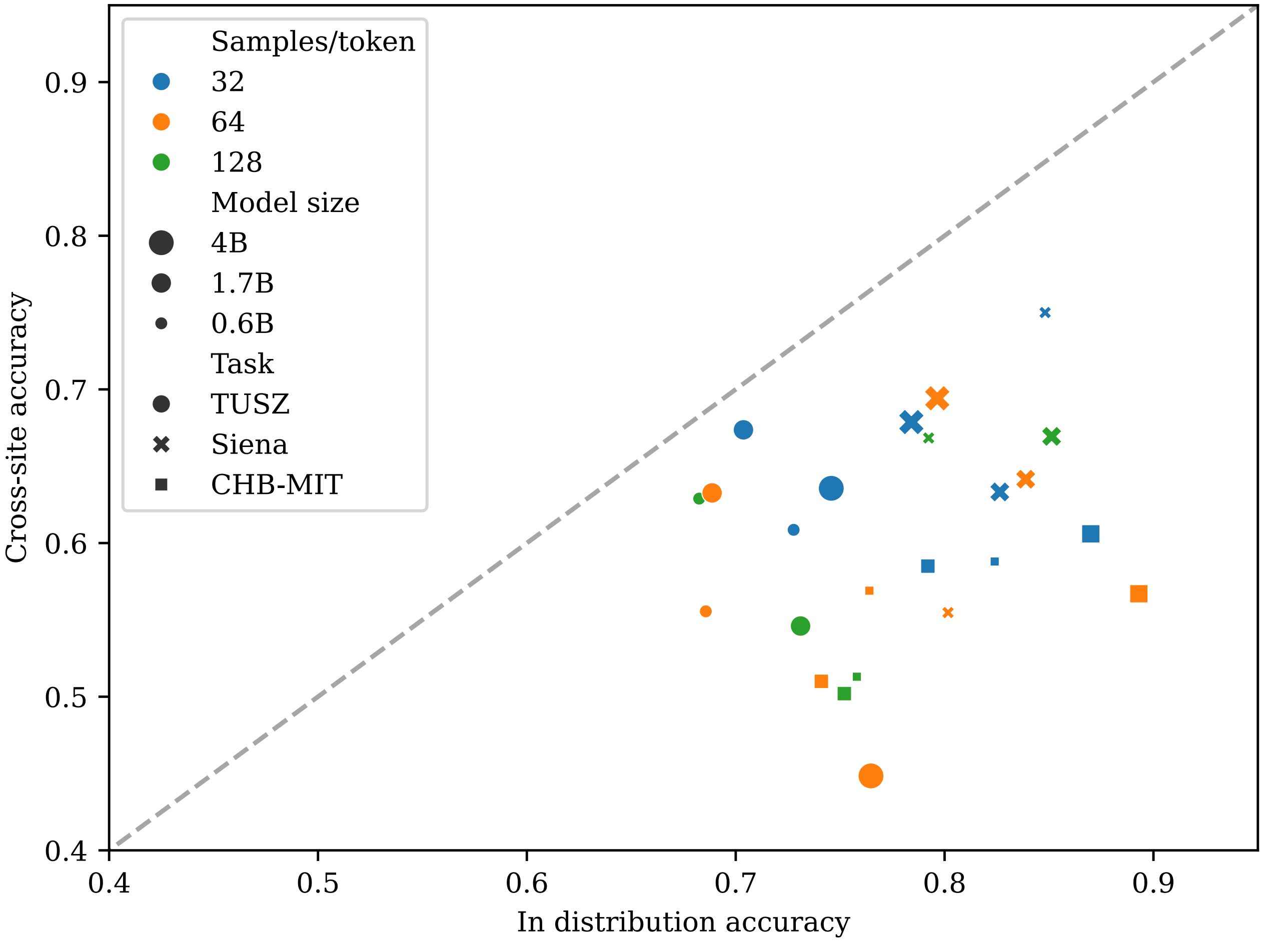

Figure 11. Comparison of in-distribution and cross-dataset model performance for seizure detection

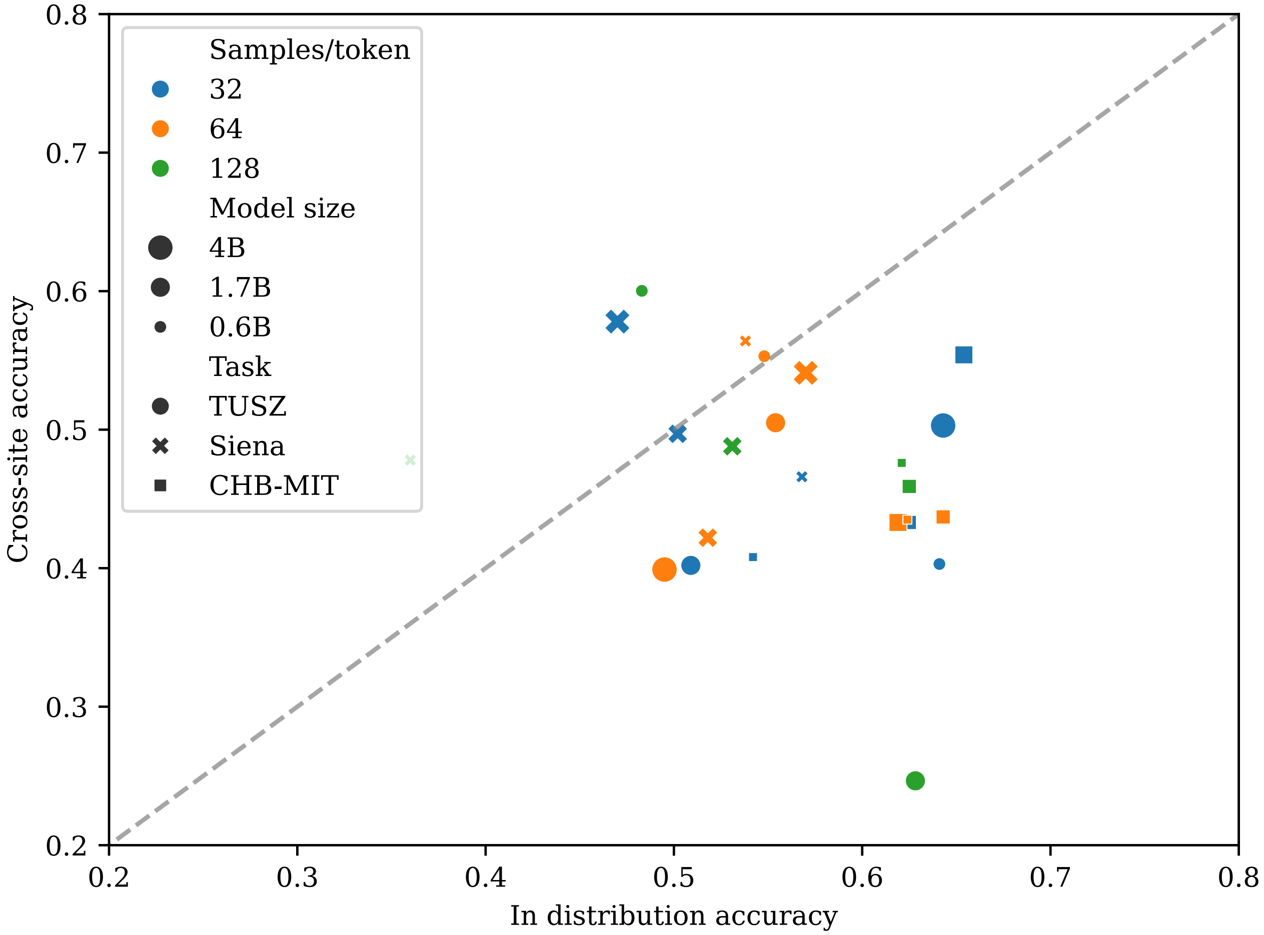

Figure 12. Comparison of in-distribution and cross-dataset model performance for seizure forecasting