# NeuroNTP - A Generalizable Multimodal Foundation Model for Epilepsy

**József Kovács**[1] **Amadeus Hauser**[2] **Gudrun Gröppel**[3] **Wolfgang Narzt**[4] **Philipp Seidl**[5]
[1]Institute of Signal Processing [2]University Clinic for Neurology
[3]Department of Paediatrics and Adolescent Medicine
[4]Department of Business Informatics – Software Engineering
[5]Institute for Machine Learning
Johannes Kepler University Linz, Austria
{jozsef.kovacs, philipp.seidl}@jku.at

## Abstract

Epilepsy affects approximately 1% of the global population, where seizure detection and reliable forecasting are critical for patient safety and quality of life. While emerging foundation models increasingly recast EEG analysis as a language modeling task, most remain confined to single-modality inputs and narrow domain generalization. In this work, we present *NeuroNTP*, a modular multimodal foundation model built on an **xLSTM** backbone and pretrained on one of the largest retrospective epilepsy datasets ever collected, comprising $150,000$ **hours** of **neural** (EEG) and **physiological** (ECG, $SpO_2$) signals paired with rich **clinical text**. NeuroNTP uses a flexible tokenizer based on **montage geometry**. It combines **causal next-token prediction**, which allows streaming inference, with **seizure-proximity supervision**: auxiliary regression targets that encode the temporal distance to the preceding and the upcoming seizure. This enables **unified sequence modeling** across modalities. This approach allows the model to capture long-range temporal dependencies and cross-modal interactions that unimodal baselines ignore. In extensive experiments, we show that NeuroNTP outperforms supervised baselines by a substantial margin on held-out clinical cohorts. Furthermore, the model **generalizes to public EEG benchmarks** such as CHB-MIT and Siena Scalp EEG, **achieving state-of-the-art** balanced accuracy and AUROC under standard evaluation settings. We release code and model weights to facilitate further research: https://github.com/ml-jku/neurontp.

## 1 Introduction

Epilepsy is a complex neurological disorder characterized by recurrent seizures, affecting millions globally [1]. Continuous monitoring is critical for diagnosis and treatment, however manual analysis of such extensive recordings is prohibitively labor-intensive and prone to error [2–4]. Automated seizure analysis promises to alleviate this burden, yet current state-of-the-art systems face significant limitations that hinder their clinical deployment [5–7]. First, **data scarcity and heterogeneity** limit robustness; models trained on small, supervised datasets struggle to generalize across the diverse hardware and sensor montages found in real-world clinical settings [8, 7]. Second, the **unimodal focus** of most systems neglects the systemic nature of seizures, failing to leverage critical autonomic biomarkers present in ECG or $SpO_2$ signals [9, 10, 7]. Third, Transformer architectures [11] face **computational constraints** in continuous monitoring: attention compute grows quadratically with context length, and even with memory-efficient attention [12] and key-value caching [13, 14], memory and latency per update keep growing with the length of the history, which limits efficient streaming over long recordings. Finally, there is a **generalization gap** between high-density clinical monitoring

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

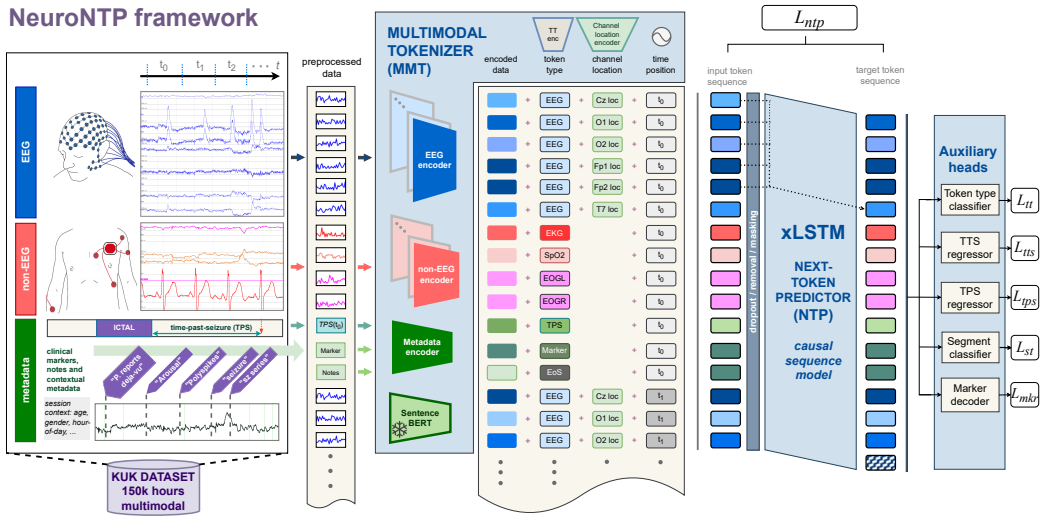


Figure 1: **NeuroNTP Framework.** Multimodal inputs (EEG, ECG, $SpO_2$, and clinical text markers) are processed by modality-specific encoders. Continuous signals are discretized and flattened into a unified sequence of *channel-level tokens*. Each token is enriched with token-type and temporal information, and, in the case of EEG tokens, with 3D spatial electrode location information. The sequence is processed by an xLSTM model optimizing a next-token prediction objective ($L_{ntp}$) alongside auxiliary tasks ($L_{aux}$).

and low-channel or modality-constrained acquisition settings [15, 16].
In this work, we propose *NeuroNTP*, a scalable multimodal foundation model designed to bridge these gaps. Our study focuses on the following main contributions:

- **Causal modelling with seizure-proximity supervision:** Our main methodological component combines causal next-token prediction over a single multimodal token stream with two auxiliary targets: time-to-seizure (TTS) and time-past-seizure (TPS). These targets shape the learned representation around clinically relevant transitions rather than stationary background activity (Table 16). We do not claim a new general multimodal pretraining paradigm; our contribution is the combination of causal sequence modelling, time-aligned clinical text, and seizure-proximity supervision for multimodal epilepsy monitoring.
- **Multimodal architecture:** NeuroNTP models EEG, ECG, $SpO_2$ and clinical text as a single token sequence in a shared representation space. Its xLSTM backbone carries a recurrent state enabling inference with constant memory usage and latency per update, whereas for Transformers both grow with the sequence length (Figure 12).
- **Montage-geometry generalization:** EEG electrode positions are encoded by their 3D positions at tokenization, while electrode-specific signal encodings are refined through learned, spatially initialized embeddings, enabling transfer to unseen montages with a frozen backbone. Related electrode encodings exist for EEG and iEEG [17–19].

## 2 Background

**Epilepsy and its clinical context.** Epilepsy is a chronic neurological disorder affecting approximately 50 million people worldwide [1]. Diagnosis and treatment rely heavily on electroencephalography (EEG) to measure brain activity and identify ictal (seizure) patterns. From a machine learning perspective, EEG analysis presents a unique set of challenges. The data is high-dimensional, non-stationary, and suffers from a low signal-to-noise ratio due to muscle artifacts and environmental noise. Furthermore, seizure manifestations are highly idiosyncratic; a seizure pattern for one patient may resemble background activity for another, leading to significant distribution shifts across subjects [7]. Historically, automated analysis relied on manual feature engineering. Deep learning (DL) shifted this paradigm towards end-to-end learning, yet standard CNNs and RNNs often struggle to generalize across the diverse hardware and sensor montages found in clinical practice. A more detailed clinical background is provided in Appendix A.

**Related Work.** The application of foundation models to neurophysiology represents a paradigm shift towards large-scale representation learning. Inspired by Large Language Models (LLMs), recent approaches treat brain activity as a "language," discretizing continuous signals into tokens to learn the intrinsic grammar of neural dynamics [20, 21]. The dominant architecture has been the Transformer, employing either **masked modeling** for retrospective analysis (e.g., BIOT [21], LaBraM [20], CBraMod [22]) or **causal modeling** for forecasting (e.g., NeuroGPT [23], BrainWave [24]). However, when streaming over long recordings, the memory and latency per update of standard Transformers grow with the length of the history, even with FlashAttention [12] and key-value caching [13, 14] (Appendix I). Furthermore, most existing models are unimodal (EEG-only). This neglects the systemic nature of epilepsy, where autonomic signs (visible in ECG) often precede cortical discharges [9]. Integrating these multimodal signals is critical for robust forecasting. Multimodal physiological pretraining has been explored before, e.g., SleepFM [25], Biosignal MAE [26], and PhysioOmni [27], and so has the general recipe of tokenization, self-supervised pretraining, and task-related supervision. These models use bidirectional or contrastive objectives and are not built for causal streaming over continuous, multi-day epilepsy monitoring (see Appendix B for their applicability to our tasks). NeuroNTP builds on these components. Its contribution is to combine them for epilepsy monitoring: causal modelling with a stateful xLSTM backbone [28], time-aligned clinical text, and seizure-proximity supervision, together with 3D electrode coordinates for montage-agnostic processing. A detailed comparison of recent foundation models is provided in Table 5 in Appendix B.

**Defining the Task Landscape.** To avoid ambiguity, we briefly distinguish the distinct tasks encompassing seizure analysis: **Seizure Detection:** The reactive classification of a time window as "ictal" (seizure) or "interictal", typically for retrospective review. **Early Detection:** The real-time identification of a seizure immediately after electrographic onset but before clinical manifestation, critical for closed-loop intervention. **Seizure Prediction:** The binary determination of whether a seizure will occur within a specific future horizon. While historically prominent, this binary paradigm is increasingly being superseded by probabilistic forecasting due to issues with reproducibility and clinical utility. **Seizure Forecasting:** The probabilistic estimation of seizure risk over a future horizon (e.g., "30% risk in the next hour"), relying on long-term state modelling, including circadian and multidien rhythm, rather than just morphological triggers [29, 30]. Although we evaluate our model on seizure prediction, we inherently view it as a forecasting system; its causal architecture outputs continuous seizure probabilities rather than deterministic triggers, enabling nuanced risk assessment rather than rigid alarms. **Seizure Localization:** The spatial identification of the Seizure Onset Zone (SOZ)—the specific anatomical region where seizure activity originates [24, 31]. **Seizure-Related Annotation Classification:** The categorization of specific events within the recording (e.g., "spike-and-wave," "arousal," or "automatisms"). Reconstruct the semantic content of expert annotations aligned with the biosignal.

## 3 Datasets

**Pretraining Dataset:** We utilize the Kepler Universitätsklinikum (KUK) dataset, a massive retrospective cohort comprising over 150,000 hours of *continuous* monitoring (up to several days) from 1,302 patients via 9,780 sessions (>16 TB). Unlike standard EEG-only benchmarks, KUK is inherently multimodal, featuring synchronized EEG (10-20 system, as well as other montages), single-lead ECG, and $SpO_2$ signals. Crucially, it includes 101 distinct marker types derived from clinical text, capturing specific behavioral manifestations (e.g., automatisms, myoclonus) and environmental context (e.g., "nursing staff enters", "patient sitting in bed") often absent in public datasets. This richness enables NeuroNTP to learn cross-modal dependencies, such as cortical-autonomic coupling. The cohort covers a broad demographic and diverse pathologies, including temporal, extratemporal, and generalized seizure types, providing the high-entropy signal necessary for effective pretraining. A detailed description is provided in Appendix D.

**Downstream datasets** To evaluate generalization across domains and form factors, we employ public datasets:

- **CHB-MIT:** An established pediatric dataset from Boston Children's Hospital [32], characterized by long-term recordings (up to several days).
- **Siena Scalp EEG:** A public dataset comprising scalp EEG recordings from adult patients [33]. We include this dataset to further validate detection performance on standard clinical montages distinct from the pretraining distribution.

# 4 NeuroNTP Framework

**NeuroNTP** is a modular multimodal sequence model for learning transferable representations from biosignals and associated clinical metadata. The architecture consists of three main components: (i) a *Multimodal Tokenizer (MMT)* that maps heterogeneous inputs into a unified token sequence, (ii) a *Next-Token Predictor (NTP)* that models this sequence autoregressively, and (iii) a set of *auxiliary prediction heads* that regularize representation learning through complementary supervisory signals (Figure 1). The framework is designed to be extensible across datasets, sensor configurations, and training regimes, enabling pretraining on large heterogeneous corpora and adaptation to downstream tasks such as seizure detection, prediction, or annotation.

## 4.1 Multimodal Tokenization and Input Representation

We formulate physiological monitoring as a multimodal sequence modeling problem. The input sequence is derived from heterogeneous sources processed by modality-specific tokenizers. NeuroNTP represents a physiological recording as a series of temporally ordered subsequences, where each timestep is encoded by tokens describing the available modalities, channels, and context. These timestep-level groups are flattened into a single causal sequence, enabling unified modeling across modalities and time.

This flattening structure forces the model to capture intra-timestep dependencies between modalities and propagate this information temporally. Combined with auxiliary supervision at every token position, the objective enforces representation learning from partial prefixes and prevents reliance on late cues. This integration of autoregressive modeling and multitask supervision generates spatially and temporally coherent representations.

The MMT projects a temporally segmented sample of $T$ chunks into a shared embedding space $\mathbb{R}^{d_{\text{emb}}}$. Each token is associated with a semantic pair $(m, c)$, where $m$ denotes the modality type (e.g., biosignal, clinical text) and $c$ the channel or structural index. This pair determines the specific encoder branch. The formulation supports both multi-channel modalities (e.g., EEG electrodes) and global tokens (e.g., clinical markers) natively.

For a given timestep $t$, the MMT dynamically constructs a sequence of tokens $\mathbf{x}_t = (x_{t,k})_{k=1}^{K_t}$, where $k$ indexes the slots available at that timestep and $K_t$ depends on the model configuration and available modalities. Each slot $k$ corresponds to a specific modality-channel pair $(m_k, c_k)$ and is encoded by its respective encoder. The complete token sequence, forming the input to the NTP module, is obtained by flattening all timestep-level subsequences: $\mathbf{X} = \mathbf{x}_0 \parallel \mathbf{x}_1 \parallel \ldots \parallel \mathbf{x}_{T-1}$, where $\parallel$ denotes concatenation. The flexible token structure accommodates variable modalities, facilitating transfer across heterogeneous datasets.

**Biosignal encoders** Biosignal channels, including EEG and non-EEG modalities, are encoded using a convolutional architecture adapted from ContraWR [34]. For efficiency and transferability, we employ a shared stem encoder that is reused across all channels, followed by lightweight channel-specific adaptation layers. This design enables parameter sharing across modalities, preserves unique channel identities, and allows reuse of pretrained representations when adapting to new montages or sensor configurations. Each channel is processed independently and without assuming any fixed ordering.

The encoder pipeline (Figure 2) consists of a short-time Fourier transform (STFT) that maps the raw signal into a magnitude spectrogram, followed by a stack of residual convolutional blocks with batch normalization and ReLU activations that extract time–frequency features, and a global average pooling operation over time and frequency that produces a fixed-size embedding in $\mathbb{R}^{d_{\text{emb}}}$.
To specialize the shared features for individual channels, we apply a small channel-specific adaptation head based on a bottleneck projection and a low-rank residual update. This mechanism introduces only a small number of additional parameters per channel while allowing fine-grained adaptation to sensor-specific characteristics. The resulting vector constitutes the biosignal token for the corresponding channel and timestep.
This formulation decouples the representation of biosignal content from the spatial identity of channels, which is introduced separately through channel location embeddings, and supports robust transfer to new acquisition setups with minimal architectural changes.

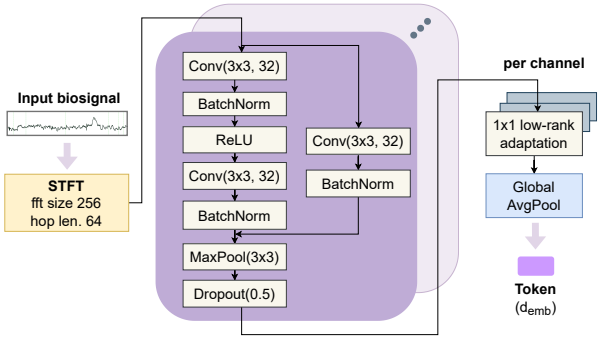


Figure 2: **Biosignal encoder.** Efficient biosignal encoder with low-rank adaptation to channel-specific signal characteristics.

**Clinical Markers, Notes, Temporal Context, and Metadata** Clinical markers and free-text notes are encoded using a frozen medical-BERT model [35], yielding semantic embeddings that are projected into $\mathbb{R}^{d_{\text{emb}}}$. When multiple markers occur within the same timestep, their embeddings are aggregated, resulting in a multi-label semantic target that is later reconstructed by an auxiliary head. This encourages the model to preserve clinically meaningful semantic structure and prevents collapse toward purely segment-level representations.
Temporal context is represented by time-past-seizure (TPS) and time-to-seizure (TTS), computed globally per patient from real-world timestamps across all sessions. While both are used as regression targets, only TPS is provided as input. TTS is withheld from the input and predicted solely from past observations, preventing look-ahead leakage and enforcing a strictly causal modeling setup.
Session-level metadata, including patient age, gender, and the recording start hour, are integrated via the End-of-Segment (EoS) token. Age is encoded via a learnable linear projection of the normalized value, while gender and hour-of-day are mapped to learnable embeddings via lookup tables. These context vectors are summed with the EoS token at every timestep, conditioning the sequence generation on demographic and circadian priors without disrupting the temporal flow of biosignal tokens.

### 4.2 Multimodal next-token prediction

The core of NeuroNTP is the Extended Long Short-Term Memory (xLSTM) [28] backbone. Unlike Transformers, whose attention compute grows quadratically with sequence length and whose memory per inference step grows with the history, xLSTM scales linearly ($\mathcal{O}(N)$) and runs with constant memory per step in recurrent mode, while retaining the ability to model long-range dependencies. xLSTM introduces two main innovations: exponential gating and a matrix memory.

**sLSTM (Scalar LSTM):** To enable the revision of storage decisions, sLSTM employs exponential gating with a scalar memory cell $c_t$ and a normalizer state $n_t$:

$$c_t = f_t c_{t-1} + i_t z_t \tag{1}$$
$$n_t = f_t n_{t-1} + i_t \tag{2}$$

where $i_t = \exp(\tilde{i}_t)$ and $f_t = \sigma(\tilde{f}_t)$ or $\exp(\tilde{f}_t)$ are the exponential input and forget gates. The hidden state is normalized as $h_t = o_t \odot (c_t/n_t)$.

**mLSTM (Matrix LSTM):** To enhance storage capacity, mLSTM utilizes a matrix memory $\mathbf{C}_t \in \mathbb{R}^{d\times d}$ updated via a covariance rule $\mathbf{C}_t = f_t\mathbf{C}_{t-1} + i_t\mathbf{v}_t\mathbf{k}_t^\top$ where $\mathbf{k}_t$ and $\mathbf{v}_t$ are key and value vectors. Retrieval is performed via a query vector $\mathbf{q}_t$, allowing for high-capacity storage with linear complexity. This efficiency is critical for processing the high-frequency, long-duration sequences inherent to continuous seizure monitoring.

Furthermore, both sLSTM and mLSTM can be used in recurrent mode for efficient inference.

Among linear-time sequence models, we prefer xLSTM over state-space models such as Mamba [36, 37]. Both scale linearly with sequence length and support constant-memory recurrent inference; xLSTM, however, includes a variant with non-linear recurrent connections between hidden states

(memory mixing), which supports state tracking, i.e., maintaining and updating a latent state over long sequences, where the linear recurrences of current state-space models are provably limited [38, 28, 39]. This property has been shown to be important for long-horizon time-series forecasting [39], and we expect it to matter for following a slowly evolving patient state, such as a pre-ictal build-up, during continuous monitoring. We do not test this directly here; a controlled comparison with state-space backbones is left for future work.

**Embedding space regularization** To enhance robustness, we employ token dropout and masking. Biosignal channels and input tokens are randomly dropped to prevent sensor over-dependence, while non-signal tokens (e.g., clinical markers, time-past-seizure) are replaced with learnable mask tokens. This adapts masked modeling [40] to an autoregressive setting, encouraging conditional reconstruction and enforcing reliance on cross-modal context.

### 4.3 Pretraining Objective

For the main next-token prediction loss ($\mathcal{L}_{\text{ntp}}$), we use a robust L1-type regression loss instead of the mean squared error, as it is less sensitive to outliers. We implement it as a weighted sum of quantile (pinball) losses [41]; this formulation also allows asymmetric weightings, which we explored during development.

For a target embedding $\mathbf{y} \in \mathbb{R}^{d_{\text{emb}}}$ and a single predicted embedding $\hat{\mathbf{y}} \in \mathbb{R}^{d_{\text{emb}}}$, the loss is

$$\frac{1}{d_{\text{emb}}} \sum_{j=1}^{d_{\text{emb}}} \sum_{\tau \in \mathcal{Q}} \omega_\tau \max(\tau(y_j - \hat{y}_j),\, (\tau - 1)(y_j - \hat{y}_j))\,, \tag{3}$$

where $\mathcal{Q}$ denotes the quantile set and $\omega_\tau$ are the corresponding weights. Since a single prediction $\hat{\mathbf{y}}$ is shared across all quantiles, this is not a distributional (multi-quantile) prediction. With the symmetric quantile set and weights used for pretraining (Appendix L), each pair $\tau, 1-\tau$ of equal weight sums to the absolute error, so the loss reduces exactly to a scaled L1 loss, $\frac{1}{2d_{\text{emb}}}\|\mathbf{y} - \hat{\mathbf{y}}\|_1$.

The overall training objective is a weighted sum of the primary NTP loss and the auxiliary losses, $\mathcal{L} = \alpha\mathcal{L}_{\text{ntp}} + \beta\mathcal{L}_{\text{tt}} + \gamma\mathcal{L}_{\text{tts}} + \delta\mathcal{L}_{\text{tps}} + \varepsilon\mathcal{L}_{\text{st}} + \zeta\mathcal{L}_{\text{mkr}}$, where $\mathcal{L}_{\text{ntp}}$ denotes the L1-type loss above for next-token prediction, $\mathcal{L}_{\text{tt}}$ the cross-entropy loss for token-type classification, $\mathcal{L}_{\text{tts}}$ and $\mathcal{L}_{\text{tps}}$ the Smooth L1 losses for time-to-seizure and time-past-seizure regression, $\mathcal{L}_{\text{st}}$ the binary cross-entropy loss for segment-type classification, and $\mathcal{L}_{\text{mkr}}$ the same L1-type loss for marker embedding reconstruction.

The training configuration and model hyperparameters were selected through targeted manual tuning, informed by prior experience with recurrent architectures and biosignal modeling (Appendix L).

**Auxiliary Tasks** While next-token prediction is the primary objective, we introduce auxiliary tasks that act as regularizers and encourage the emergence of semantically meaningful internal representations, following the auxiliary learning framework [42]. Specifically, we attach prediction heads for token type classification, segment type classification (ictal vs. non-ictal), time-past-seizure (TPS) regression, time-to-seizure (TTS) regression, and marker embedding reconstruction. These tasks are evaluated at each token position, forcing every prefix of the sequence to encode information relevant to seizure state, temporal context, and clinical semantics. The proximity targets (TTS and TPS) require annotated seizure onset times during pretraining. Such annotations are available in common public seizure datasets, such as TUSZ, CHB-MIT, and Siena (Appendix H). At inference time, these targets are not needed: unless an annotation-assisted setting is stated explicitly, downstream detection and prediction use biosignal input only.

### 4.4 Downstream Adaptation and Inference

Before finetuning, we perform *linear probing* by reusing the pretrained segment-type auxiliary head as a single linear layer binary seizure classifier and training only this head while keeping the backbone frozen. This provides a direct measure of the linear separability of the learned representations without any task-specific finetuning.

**Fine-tuning** For downstream tasks, all auxiliary heads are removed and replaced by one small task-specific MLP classifier. Channel-specific adapters for electrodes not seen during pretraining are

initialized from the spatially nearest pretrained electrode (Appendix I) and then trained. Training is performed in two stages. First, the backbone is frozen and the downstream head is trained with a learning rate of $5 \times 10^{-4}$. Second, the entire network is unfrozen and jointly optimized using a learning rate of $1 \times 10^{-5}$. The final model is selected based on the lowest segment-type classification loss on the validation set.
For ictal/non-ictal classification, we employ the focal loss [43], which has been shown to be effective for seizure detection under severe class imbalance [44]. All experiments are repeated five times with different random seeds, and results are reported as mean and standard deviation across runs.

**Inference** Test recordings are treated as continuous streams and predictions are produced for every chunk. As evaluation progresses, the recurrent state of the model is preserved and propagated, allowing each prediction to leverage the accumulated historical context. In practice, this is realized with a fixed context length, consistent with the pretraining setup, while conceptually permitting information flow across the entire recording. This protocol avoids any form of look-ahead, mirrors real-world deployment, and enables the model to exploit temporal dependencies through its internal state rather than isolated, independently processed windows.

# 5 Experiments and Results

## 5.1 Evaluation Methodology and Metrics

We evaluate NeuroNTP on seizure detection and prediction using the private KUK held-out set, and assess transfer to public benchmarks on CHB-MIT and Siena Scalp EEG. We further evaluate clinical marker inference on KUK without task-specific fine-tuning as a proxy for semantic alignment between learned biosignal representations and the clinical vocabulary. We additionally ablate the electrode-location encoding (Appendix J, Table 17).

**Metrics:** We report Balanced Accuracy (BACC), Area Under the Precision-Recall Curve (AUPRC), and Area Under the ROC Curve (AUROC). For seizure prediction, we additionally report Delta Average Precision ($\Delta$AP) (AUPRC minus base-rate; 0 at random). We view AUPRC as the most relevant metric because we are dealing with highly unbalanced data and prioritize the positive class, whereas AUROC can be high if the model is good at ranking negatives low.

**Baselines:** We compare against task-specific models trained from scratch or fine-tuned, including BIOT [21], ContraWR [34], FFCL [45], SPaRCNet [46], CNN-Transformer [47], and EEGNet [48], alongside other state-of-the-art architectures.

**Input settings:** Our evaluation primarily targets deployable settings. Unless stated otherwise, models receive only the recorded biosignals as input (*biosignal-only*). Information unavailable in a live stream, such as clinical markers, free-text notes, or the time past the last seizure, is excluded. For the multimodal KUK dataset, we additionally assess the effect of supplying such auxiliary information as input (*annotation-assisted*). Specifically, for seizure detection we vary the groups of clinical text annotations provided, and for seizure prediction we provide the time-past-seizure (TPS) value.

## 5.2 Performance on the KUK Private Dataset

### 5.2.1 Seizure Detection

We first evaluate NeuroNTP on the held-out portion of our large-scale private pretraining dataset, KUK. A four-fold cross-validation scheme is employed using the recordings of 60 patients, with each fold using 15 patients for evaluation. This provides a direct measure of the model's performance on the same data distribution it was trained on, but with unseen patients. As shown in Table 1, NeuroNTP outperforms all supervised baselines. Notably, the inclusion of clinical annotations as an input modality provides a substantial boost in performance across all metrics, highlighting the value of integrating rich contextual information. Specifically, clinical notes provide critical environmental context (e.g., "patient eating", "nurse entering") that enables the model to disentangle physiological artifacts caused by daily activities from true ictal patterns. With biosignal-only input, i.e., the same input as all baselines, NeuroNTP reaches an AUROC of 84.2%, +2.5% over the best-performing baseline (FFCL [45]; Table 1a). With annotation-assisted input, the AUROC increases further to 88.4% when EEG-phenomenon annotations are added (Table 1b).

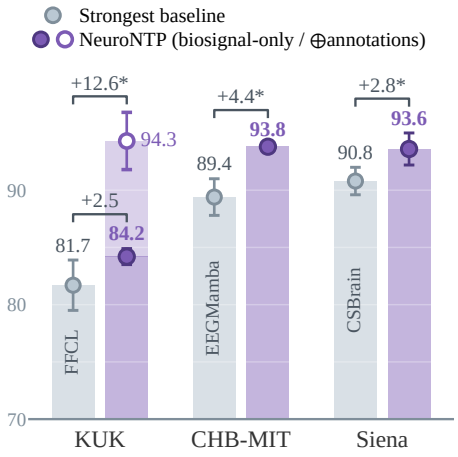


Figure 3: **Seizure detection performance (AUROC).** NeuroNTP vs. the strongest baseline per dataset (name inside the grey bar), biosignal-only input (Tables 1a and 4). Open marker and light bar: KUK with all clinical annotations as input (Table 1b). NeuroNTP outperforms the strongest baseline on all three datasets; the gain is significant on CHB-MIT and Siena, and on KUK once annotations are added. Error bars: $\pm 1$ std; y-axis starts at 70%. Brackets: AUROC gain over the baseline; * $p < 0.05$ (two-sided Welch's $t$-test; $n = 5$ runs, KUK: 4 folds).

Table 1: **Seizure detection performance on the KUK held-out dataset.** Results use 4-fold patient-wise cross-validation on recordings from 60 unseen patients. Values are mean $\pm$ standard deviation (in %). Purple denotes best and grey within one std of best.

(a) Comparison with task-specific baselines (biosignal-only).

| Model | BACC | AUPRC | AUROC |
|---|---|---|---|
| BIOT | $59.2^{\pm 2.1}$ | $38.3^{\pm 2.4}$ | $74.3^{\pm 1.2}$ |
| ContraWR | $59.7^{\pm 2.0}$ | $48.4^{\pm 3.7}$ | $79.0^{\pm 3.6}$ |
| EEGNet | $50.8^{\pm 0.1}$ | $34.5^{\pm 1.0}$ | $70.4^{\pm 1.3}$ |
| FFCL | $61.7^{\pm 1.9}$ | $51.9^{\pm 1.8}$ | $81.7^{\pm 2.2}$ |
| Pytorch-TCN | $58.3^{\pm 3.1}$ | $38.3^{\pm 0.6}$ | $73.6^{\pm 1.4}$ |
| SPaRCNet | $56.3^{\pm 1.2}$ | $38.5^{\pm 1.2}$ | $65.8^{\pm 2.5}$ |
| STTransformer | $55.5^{\pm 0.9}$ | $29.6^{\pm 11.7}$ | $64.1^{\pm 13.9}$ |
| **NeuroNTP** | $\mathbf{71.5}^{\pm 2.1}$ | $\mathbf{52.7}^{\pm 5.5}$ | $\mathbf{84.2}^{\pm 0.7}$ |

(b) NeuroNTP with annotation-assisted evaluation.

| Marker group† | BACC | AUPRC | AUROC |
|---|---|---|---|
| General / technical | $74.7^{\pm 3.1}$ | $60.7^{\pm 5.0}$ | $87.0^{\pm 0.6}$ |
| ⊕ EEG phenomena | $85.3^{\pm 6.0}$ | $79.3^{\pm 9.2}$ | $88.4^{\pm 5.9}$ |
| ⊕ Seizure-related | $\mathbf{90.7}^{\pm 2.6}$ | $\mathbf{86.7}^{\pm 4.7}$ | $\mathbf{94.3}^{\pm 2.5}$ |

† Temporally aligned clinical annotations provided as additional input; ⊕ marks groups added cumulatively to those in the rows above. Marker groups are detailed in Appendix D.1.

### 5.2.2 Seizure Prediction

We evaluate the model's ability to predict seizures on the KUK dataset across different forecasting horizons. Detailed quantitative results are provided in Appendix E (Table 8, Figure 10). The multimodal, fine-tuned NeuroNTP consistently outperforms all baselines across all horizons. For the 0–15 minute horizon, it achieves an AUROC of 85.0%, demonstrating strong predictive capability. Performance degrades gracefully with increasing horizon length while remaining superior to competing methods. The comparison between multimodal and biosignal-only variants underscores the importance of non-EEG data for robust forecasting. Even the frozen (linear probing) multimodal model remains highly competitive, indicating that the pretraining objective yields representations that are linearly separable for the prediction task. While clinical text and contextual signals such as time-past-seizure may not always be directly available in continuous monitoring settings, they could potentially be approximated through automated detection and annotation.

### 5.2.3 Auxiliary Tasks

Beyond the primary downstream capabilities, NeuroNTP was evaluated on several epiliepsy-relevant auxiliary tasks to assess the semantic and temporal richness of its learned representations:
**Clinical Marker Inference:** This task evaluates the model's ability to reconstruct the semantic content of expert annotations (e.g., "Spike", "Arousal", "Seizure") solely from biosignal input. The model predicts a marker token embedding which is mapped to a frozen Sentence-BERT space, serving as a proxy for how well the model aligns physiological patterns with clinical vocabulary.
**Time-to-Seizure (TTS) Regression:** Distinct from the binary classification used in seizure prediction, this task estimates the continuous time remaining until the next seizure onset. It is evaluated via linear probing to determine if the pre-seizure latent trajectory contains linearly separable information regarding the temporal proximity of an impending event.
**Auxiliary Structural Tasks:** Several internal objectives used during self-supervised pretraining were analyzed to verify structural learning. These include *Time-Past-Seizure (TPS) regression* (encoding retrospective temporal context), *Token Type Classification* (discriminating between heterogeneous modalities like EEG and ECG), and *Segment Type Classification* (distinguishing ictal from non-ictal

activity in the latent space). Linear probing of the segment-type head on frozen representations with biosignal-only input (Table 3) shows that multimodal pretraining yields clearly better representations than biosignal-only pretraining on all three datasets (e.g., KUK AUROC 78.9% vs. 63.1%), even though no text is used at inference; the clinical-context objectives thus regularize the latent space beyond the modalities seen at test time.

Table 2: **Group-level detection performance of clinical marker categories.** Coarse-grained evaluation without task-specific fine-tuning. Support = number of ground-truth instances; values in %. Detailed confusion matrix in Appendix G.

| Group | Support | Precision | Recall | F1 |
|---|---|---|---|---|
| Seizure event | 835 | 56.2 | 68.7 | 61.8 |
| EEG phenomenon | 1169 | 70.4 | 97.0 | 81.6 |
| Behavioural | 688 | 56.0 | 40.0 | 46.7 |
| Other | 606 | 53.8 | 15.4 | 23.9 |

Table 3: **Linear probing** Comparison of biosignal-only and multimodal pretraining on KUK, evaluated using linear probing with frozen backbones and biosignal-only input across three datasets. Values are mean ± standard deviation in %.

| | Biosignal-only pretraining | | Multimodal pretraining | |
|---|---|---|---|---|
| Dataset | AUPRC | AUROC | AUPRC | AUROC |
| KUK | $26.5^{\pm 3.8}$ | $63.1^{\pm 4.8}$ | $43.7^{\pm 1.4}$ | $78.9^{\pm 1.5}$ |
| CHB-MIT | $14.9^{\pm 2.6}$ | $84.2^{\pm 3.4}$ | $42.0^{\pm 5.7}$ | $92.3^{\pm 0.2}$ |
| Siena | $23.5^{\pm 3.5}$ | $89.0^{\pm 0.6}$ | $32.7^{\pm 2.0}$ | $91.4^{\pm 1.0}$ |

### 5.3 Generalization to Public Datasets

#### 5.3.1 Seizure Detection on CHB-MIT

To assess generalization across patient populations and recording setups, we evaluate on the public CHB-MIT pediatric dataset. We follow the CBraMod protocol [22] (see Appendix F). As shown in Table 4, NeuroNTP achieves the highest AUROC (93.8%), improving over the strongest prior model (EEGMamba, 89.4%) by 4.4%. While CSBrain attains the highest AUPRC, NeuroNTP remains competitive and provides the strongest ranking performance. These results, obtained on a pediatric dataset under a substantially different montage than used during pretraining, demonstrate the robustness and transferability of the learned representations.

#### 5.3.2 Seizure Detection on Siena Scalp

We further evaluate generalization on the Siena Scalp EEG dataset, following the protocol of Zhou et al. [49] (see Appendix F). As shown in Table 4, NeuroNTP achieves the highest AUROC (93.6%), compared with 90.8% for the strongest prior model (CSBrain) and 87.2% for the strongest supervised baseline (CNN-Transformer). While CSBrain attains the highest AUPRC, NeuroNTP is within one standard deviation of it. The consistent performance across KUK, CHB-MIT, and Siena underscores the effectiveness of large-scale multimodal pretraining for robust and generalizable seizure detection.

## 6 Discussion

**Conclusions** We have presented NeuroNTP, a multimodal foundation model for epilepsy monitoring. It combines causal next-token prediction, time-aligned clinical text, and seizure-proximity supervision, and uses a stateful xLSTM backbone that keeps memory per update constant over long recordings. By pretraining on the massive KUK dataset, we demonstrate that scaling data and model capacity unlocks superior performance on downstream tasks, even when transferring to modality-constrained settings.

**Limitations** While privacy regulations preclude releasing the KUK dataset, we mitigate this by releasing NeuroNTP's full weights and code, and verifying reproducibility on the public TUH corpus. Furthermore, the model's generative nature allows it to function as a high-fidelity simulator for synthetic data augmentation. Technically, while NeuroNTP transfers to unseen montages with a frozen backbone, it currently lacks temporal invariance, requiring fixed-rate resampling. Additionally, the modality-consistent ordering in flattened token sequences may introduce inductive bias regarding signal simultaneity, necessitating future exploration of permutation-invariant masking. Finally, despite xLSTM's constant-memory stateful inference, edge deployment remains memory-constrained, requiring quantization, while clinical adoption will necessitate further development of interpretability. For clinical use, a system must balance false alarms against missed seizures: too many false alarms cause alarm fatigue, while missed detections put patients at risk. Our evaluation mainly reports

Table 4: **Seizure detection on CHB-MIT and Siena Scalp.** NeuroNTP (multimodally pretrained) is fine-tuned and evaluated with biosignal-only input (no clinical markers, notes, or TPS). Baseline results are taken from Wang et al. [22] (CHB-MIT) and Zhou et al. [49] (Siena); – denotes not reported. Values are mean ± standard deviation in %. Purple denotes best and grey within one std of best.

| | CHB-MIT | | | Siena Scalp | | |
|---|---|---|---|---|---|---|
| **Method** | **BACC** | **AUPRC** | **AUROC** | **BACC** | **AUPRC** | **AUROC** |
| EEGNet | $56.6^{\pm 1.1}$ | $19.1^{\pm 1.8}$ | $80.5^{\pm 1.4}$ | $74.9^{\pm 5.2}$ | $37.5^{\pm 8.7}$ | $86.9^{\pm 5.3}$ |
| EEGConformer | $59.8^{\pm 1.4}$ | $22.1^{\pm 2.1}$ | $82.3^{\pm 1.7}$ | $75.6^{\pm 2.1}$ | $20.9^{\pm 7.9}$ | $81.6^{\pm 2.6}$ |
| SPaRCNet | $58.8^{\pm 1.9}$ | $12.5^{\pm 1.2}$ | $81.4^{\pm 1.5}$ | $65.7^{\pm 3.8}$ | $31.6^{\pm 6.6}$ | $73.3^{\pm 8.6}$ |
| ContraWR | $63.4^{\pm 0.0}$ | $22.6^{\pm 1.7}$ | $81.0^{\pm 1.1}$ | $65.5^{\pm 3.1}$ | $37.1^{\pm 4.1}$ | $78.2^{\pm 6.0}$ |
| CNN-Transformer | $63.9^{\pm 0.7}$ | $24.8^{\pm 2.3}$ | $86.6^{\pm 0.8}$ | $69.8^{\pm 5.6}$ | $38.4^{\pm 7.1}$ | $87.2^{\pm 4.0}$ |
| FFCL | $62.6^{\pm 1.0}$ | $20.5^{\pm 3.5}$ | $82.7^{\pm 0.5}$ | $66.2^{\pm 3.9}$ | $39.4^{\pm 9.0}$ | $81.5^{\pm 11.6}$ |
| ST-Transformer | $59.2^{\pm 1.9}$ | $14.2^{\pm 0.9}$ | $82.4^{\pm 4.9}$ | $75.3^{\pm 3.8}$ | $36.4^{\pm 2.5}$ | $88.8^{\pm 0.9}$ |
| BIOT | $70.7^{\pm 4.6}$ | $32.8^{\pm 4.6}$ | $87.6^{\pm 2.8}$ | $73.5^{\pm 6.7}$ | $38.1^{\pm 8.9}$ | $90.3^{\pm 3.0}$ |
| LaBraM-Base | $70.8^{\pm 3.6}$ | $32.9^{\pm 4.0}$ | $86.8^{\pm 2.0}$ | $70.8^{\pm 3.3}$ | $31.2^{\pm 9.8}$ | $88.1^{\pm 3.3}$ |
| CBraMod | $74.0^{\pm 2.8}$ | $36.9^{\pm 3.8}$ | $88.9^{\pm 1.5}$ | $73.2^{\pm 6.5}$ | $41.1^{\pm 7.2}$ | $90.4^{\pm 2.2}$ |
| EEGMamba | $74.5^{\pm 2.5}$ | $38.9^{\pm 4.2}$ | $89.4^{\pm 1.6}$ | – | – | – |
| CSBrain | $72.6^{\pm 1.2}$ | $\mathbf{51.6}^{\pm 4.5}$ | $89.2^{\pm 3.2}$ | $76.6^{\pm 4.7}$ | $\mathbf{48.7}^{\pm 3.4}$ | $90.8^{\pm 1.2}$ |
| **NeuroNTP** (ours) | $\mathbf{84.7}^{\pm 1.1}$ | $45.4^{\pm 0.8}$ | $\mathbf{93.8}^{\pm 0.2}$ | $\mathbf{87.5}^{\pm 2.1}$ | $46.6^{\pm 7.4}$ | $\mathbf{93.6}^{\pm 1.4}$ |

ranking metrics (AUROC, AUPRC), which do not fix this trade-off; an operating threshold must still be chosen, ideally per care setting and evaluated with event-based measures such as false alarms per hour. Clinical text also generally requires careful handling with respect to privacy. In our case, the text inputs are short clinical annotations made during monitoring that describe EEG phenomena, clinical signs, sleep stages, and technical events (e.g., "spike", "oral automatisms", "REM", "artifact"; Appendix D.1), and they contain no patient-identifying information.

**Outlook** In future work, we plan to validate NeuroNTP on the prospective arm of the KUK study to assess its real-time forecasting capabilities in a live clinical setting. Furthermore, we aim to extend the model's multimodal capabilities to include additional wearable sensors (e.g., photoplethysmography, electrodermal activity), enhancing its applicability to diverse ambulatory monitoring setups. We also envision extending NeuroNTP to a Federated Learning framework, allowing institutions to collaboratively train on decentralized data without compromising patient privacy. Finally, our objective is to explore few-shot adaptation strategies to rapidly personalize the foundation model for individual patients using limited calibration data, further improving forecast precision [50].

**Data and Code Availability** The private KUK dataset used for pretraining cannot be publicly released due to strict patient privacy regulations. The source code for NeuroNTP, including the model architecture, training pipelines, and evaluation scripts, is publicly available at `https://github.com/ml-jku/neurontp`. We also release the pretrained model weights to facilitate further research.

**Ethics** All activities within the scope of the EPILEPSIA project were legitimized by a positive ethics vote (EK 1021/2023) by the responsible ethics committee.

**Broader Impact** This work contributes to automated seizure detection and forecasting systems that have the potential to improve quality of life for people with epilepsy by enabling proactive management, reducing injury risk, and alleviating uncertainty. At the same time, deploying large-scale foundation models in healthcare raises important considerations. Clinical data require strict attention to privacy and security; we address this by releasing pretrained weights rather than raw data and by exploring federated learning approaches. Because NeuroNTP was pretrained on data from a single center, it may inherit local biases, and its performance may not transfer to other hospitals, patient populations, or ambulatory settings without local validation. Finally, unclassified seizures could go unattended and false alarms may cause alarm fatigue, so operating thresholds should be set together for each patient with clinicians and validated prospectively.

**Acknowledgments and Disclosures**

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria, carried out by the Austrian Research Promotion Agency FFG, in the frame of the initiative "Digital Health - The Digital Patient Journey". Approved by the Ethics Committee of the Johannes Kepler University Linz (EK 1021/2023). We thank the Kepler University Hospital and the Institute of Signal Processing, JKU. We further thank the projects FWF AIRI FG 9-N (10.55776/FG9), AI4GreenHeatingGrids (FFG- 899943), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01), FWF Bilateral Artificial Intelligence (10.55776/COE12) as well as NXAI GmbH, Audi AG, Silicon Austria Labs (SAL), Merck Healthcare KGaA, GLS (Univ. Waterloo), TÜV Holding GmbH, Software Competence Center Hagenberg GmbH, dSPACE GmbH, TRUMPF SE + Co. KG and FiveSquare GmbH.

**Competing interests:** The authors declare no competing interests.

## References

[1] World Health Organization. Epilepsy, February 2024. URL https://www.who.int/news-room/fact-sheets/detail/epilepsy. (Accessed: March 01, 2026).

[2] Simon Shorvon, Renzo Guerrini, Mark Cook, and Samden Lhatoo, editors. *Oxford Textbook of Epilepsy and Epileptic Seizures*. Oxford University Press, 2012. ISBN 9780199659043. doi: 10.1093/med/9780199659043.001.0001. URL https://doi.org/10.1093/med/9780199659043.001.0001.

[3] David E Friedman and Lawrence J Hirsch. How long does it take to make an accurate diagnosis in an epilepsy monitoring unit? *Journal of Clinical Neurophysiology*, 26(4):213–217, 2009. doi: 10.1097/WNP.0b013e3181b2f2da.

[4] Christa B Swisher, Corey R White, Brian E Mace, Keith E Dombrowski, Aatif M Husain, Bradley J Kolls, Rodney R Radtke, Tung T Tran, and Saurabh R Sinha. Diagnostic accuracy of electrographic seizure detection by neurophysiologists and non-neurophysiologists in the adult ICU using a panel of quantitative EEG trends. *Journal of clinical neurophysiology*, 32(4): 324–330, 2015. doi: 10.1097/WNP.0000000000000144.

[5] Nina Moutonnet, Steven White, Benjamin P Campbell, Saeid Sanei, Toshihisa Tanaka, Hong Ji, Danilo Mandic, and Gregory Scott. Clinical translation of machine learning algorithms for seizure detection in scalp electroencephalography: systematic review, 2024. URL https://arxiv.org/abs/2404.15332.

[6] Sheng Wong, Anj Simmons, Jessica Rivera-Villicana, Scott Barnett, Shobi Sivathamboo, P. Perucca, Patrick Kwan, L. Kuhlmann, Rajesh Vasa, and Terence J. O'Brien. EEG based automated seizure detection–A survey of medical professionals. *Epilepsy & Behavior*, 149: 109518, 2023. doi: 10.1016/j.yebeh.2023.109518.

[7] Ijaz Ahmad, Faizan Ahmad, Sunday Timothy Aboyeji, Yongtao Zhang, Peng Yang, Javed Ali Khan, Rab Nawaz, and Baiying Lei. Advanced Multimodal Learning for Seizure Detection and Prediction: Concept, Challenges, and Future Directions, 2026. URL https://arxiv.org/abs/2601.05095.

[8] Sheng Wong, Anj Simmons, Jessica Rivera-Villicana, Scott Barnett, Shobi Sivathamboo, P. Perucca, Z. Ge, P. Kwan, L. Kuhlmann, Rajesh Vasa, K. Mouzakis, and T. O'Brien. EEG datasets for seizure detection and prediction—A review. *Epilepsia Open*, 8(2):252–267, 2023. doi: 10.1002/epi4.12704.

[9] Roland D Thijs, Philippe Ryvlin, and Rainer Surges. Autonomic manifestations of epilepsy: emerging pathways to sudden death? *Nature Reviews Neurology*, 17(12):774–788, 2021. doi: 10.1038/s41582-021-00574-w.

[10] K. Vandecasteele, Thomas De Cooman, C. Chatzichristos, E. Cleeren, Lauren Swinnen, Jaiver Macea Ortiz, S. Van Huffel, M. Dümpelmann, A. Schulze-Bonhage, M. de Vos, W. van

Paesschen, and B. Hunyadi. The power of ECG in multimodal patient-specific seizure monitoring: added value to an EEG-based detector using limited channels. *Epilepsia*, 62(10): 2333–2343, 2021. doi: 10.1111/epi.16990.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[13] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

[14] Chunan Shi, Yilei Chen, Yilin Chen, Xupeng Miao, and Bin Cui. Multi-segment attention: Enabling efficient kv-cache management for faster large language model serving. *arXiv preprint arXiv:2606.02964*, 2026.

[15] Miguel Bhagubai, Christos Chatzichristos, Lauren Swinnen, Jaiver Macea, Jingwei Zhang, Lieven Lagae, Katrien Jansen, Andreas Schulze-Bonhage, Francisco Sales, Benno Mahler, Yvonne Weber, Wim Van Paesschen, and Maarten De Vos. BIDS SeizeIT2. *OpenNeuro*, 2025. doi: 10.18112/openneuro.ds005873.v1.0.0. URL https://openneuro.org/datasets/ds005873/versions/1.0.0.

[16] Ilona Hubbard, Sandor Beniczky, and Philippe Ryvlin. The challenging path to developing a mobile health device for epilepsy: the current landscape and where we go from here. *Frontiers in neurology*, 12:740743, 2021.

[17] Georgios Mentzelopoulos, Evangelos Chatzipantazis, Ashwin G Ramayya, Michelle J Hedlund, Vivek P Buch, Kostas Daniilidis, Konrad P Kording, and Flavia Vitale. Neural decoding from stereotactic eeg: accounting for electrode variability across subjects. *Advances in Neural Information Processing Systems*, 37:108600–108624, 2024.

[18] Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. In *International Conference on Learning Representations*, volume 2025, pages 69100–69121, 2025.

[19] Lucine L Oganesian, Saba Hashemi, and Maryam Shanechi. Barista: Brain scale informed spatiotemporal representation of human intracranial neural activity. *Advances in Neural Information Processing Systems*, 38:167161–167194, 2026.

[20] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI, 2024. URL https://arxiv.org/abs/2405.18765.

[21] Chaoqi Yang, M Brandon Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. In *Neural Information Processing Systems*, pages 78240–78260. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-3420. URL https://openreview.net/forum?id=c2LZyTyddi.

[22] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding, 2025. URL https://arxiv.org/abs/2412.07236.

[23] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-GPT: Towards a Foundation Model for EEG. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[24] Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications, 2025. URL https://arxiv.org/abs/2402.10251.

[25] Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals, 2024. URL https://arxiv.org/abs/2405.17766.

[26] Ching Fang, Christopher Sandino, Behrooz Mahasseni, Juri Minxha, Hadi Pouransari, Erdrin Azemi, Ali Moin, and Ellen Zippi. Promoting cross-modal representations to improve multimodal foundation models for physiological signals. *arXiv preprint arXiv:2410.16424*, 2024.

[27] Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. *arXiv preprint arXiv:2504.19596*, 2025.

[28] Maximilian Beck, Korbinian Poppel, M. Spanring, Andreas Auer, Oleksandra Prudnikova, Michael K Kopp, G. Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended Long Short-Term Memory. *Neural Information Processing Systems*, pages 107547–107603, 2024. doi: 10.52202/079017-3417.

[29] Benjamin H Brinkmann, Edward E Patterson, Charles Vite, Vincent M Vasoli, Daniel Crepeau, Matt Stead, J Jeffry Howbert, Vladimir Cherkassky, Joost B Wagenaar, Brian Litt, et al. Forecasting seizures using intracranial EEG measures and SVM in naturally occurring canine epilepsy. *PloS one*, 10(8):e0133900, 2015.

[30] Timothée Proix, Wilson Truccolo, Marc G Leguia, Thomas K Tcheng, David King-Stephens, Vikram R Rao, and Maxime O Baud. Forecasting seizure risk in adults with focal epilepsy: a development and validation study. *The Lancet Neurology*, 20(2):127–135, February 2021. ISSN 1474-4422. doi: 10.1016/s1474-4422(20)30396-3. URL http://dx.doi.org/10.1016/s1474-4422(20)30396-3.

[31] Bright Worlanyo Aklamanu. A review of neuroimaging in epilepsy: Diagnostic strategies and clinical decision framework. *Brain Disorders*, 19:100261, 2025. doi: 10.1016/j.dscb.2025.100261.

[32] Ali H Shoeb. Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment. *PhD Thesis, Massachusetts Institute of Technology*, 2009.

[33] P. Detti, G. Vatti, and Garazi Zabalo Manrique de Lara. EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7):846, 2020. doi: 10.3390/pr8070846.

[34] Chaoqi Yang, Danica Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised EEG Representation Learning for Automatic Sleep Staging, 2023. URL https://arxiv.org/abs/2110.15278.

[35] Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J.W.L. Aerts, and Alexander Löser. medBERT.de: A comprehensive German BERT model for the medical domain. *Expert Systems with Applications*, 237:121598, March 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.121598. URL http://dx.doi.org/10.1016/j.eswa.2023.121598.

[36] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

[37] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL https://arxiv.org/abs/2405.21060.

[38] William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35492–35506, 21–27 Jul 2024.

[39] Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.23719.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/N19-1423.

[41] Roger Koenker and Gilbert Bassett. Regression Quantiles. *Econometrica*, 46(1):33, January 1978. ISSN 0012-9682. doi: 10.2307/1913643. URL http://dx.doi.org/10.2307/1913643.

[42] Shikun Liu, Andrew Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. *Neural Information Processing Systems*, 32, 2019.

[43] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection, 2017. URL https://arxiv.org/abs/1708.02002.

[44] Yanna Zhao, Changxu Dong, Gaobo Zhang, Yaru Wang, Xin Chen, Weikuan Jia, Qi Yuan, Fangzhou Xu, and Yuanjie Zheng. EEG-Based Seizure detection using linear graph convolution network with focal loss. *Comput. Methods Programs Biomed.*, 208:106277, 2021. doi: 10.1016/j.cmpb.2021.106277.

[45] Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. *Biomedical signal processing and control*, 72:103342, 2022. doi: 10.1016/j.bspc.2021.103342.

[46] Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17):e1750–e1762, 2023.

[47] Yaxin Ma, Yonghao Song, and Fei Gao. A novel hybrid CNN-transformer model for EEG motor imagery classification. In *2022 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2022.

[48] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, July 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c. URL http://dx.doi.org/10.1088/1741-2552/aace8c.

[49] Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=agcXjEHmyW.

[50] Ekaterina Sysoykova, Bernhard Anzengruber-Tanase, Michael Haslgrubler, Philipp Seidl, and Alois Ferscha. Federated Few-Shot Learning for Epileptic Seizure Detection Under Privacy Constraints. *arXiv preprint arXiv:2512.13717*, 2025. doi: 10.48550/arXiv.2512.13717.

[51] Robert S. Fisher. The New Classification of Seizures by the International League Against Epilepsy 2017. *Current Neurology and Neuroscience Reports*, 17:1–6, 2017. URL https://api.semanticscholar.org/CorpusID:31647719.

[52] S. Beniczky, E. Trinka, E. Wirrell, F. Abdulla, Raidah Al Baradie, M. Alonso Vanegas, S. Auvin, M. B. Singh, H. Blumenfeld, Alicia Bogacz Fressola, R. Caraballo, Mar Carreño, F. Cendes, Augustina Charway, Mark Cook, Dana Craiu, B. Ezeala-Adikaibe, B. Frauscher, Jacqueline A French, M. Gule, N. Higurashi, Akio Ikeda, F. Jansen, Barbara Jobst, Philippe Kahane, N. Kishk, C. S. Khoo, K. Vinayan, L. Lagae, K. Lim, A. Lizcano, A. McGonigal, Katerina Tanya Perez-Gosiengfiao, P. Ryvlin, N. Specchio, Michael R Sperling, Hermann Stefan, William O. Tatum, M. Tripathi, E. M. Yacubian, S. Wiebe, J. Wilmshurst, Dong Zhou, and J. Cross.

Updated classification of epileptic seizures: Position paper of the International League Against Epilepsy. *Epilepsia*, 66(6):1804 – 1823, 2025. doi: 10.1111/epi.18338. URL https://api.semanticscholar.org/CorpusID:277995387.

[53] Kirsten M Fiest, Khara M Sauro, Samuel Wiebe, Scott B Patten, Churl-Su Kwon, Jonathan Dykeman, Tamara Pringsheim, Diane L Lorenzetti, and Nathalie Jetté. Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. *Neurology*, 88(3): 296–303, 2017.

[54] Dale C. Hesdorffer, Giancarlo Logroscino, Emma K. T. Benn, N Katri, Gregory D. Cascino, and W. A. Hauser. Estimating risk for developing epilepsy. *Neurology*, 76:23 – 27, 2011. URL https://api.semanticscholar.org/CorpusID:31851591.

[55] Christian Meisel and Tobias Loddenkemper. Seizure prediction and intervention. *Neuropharmacology*, 172:107898, 2020. ISSN 0028-3908. doi: https://doi.org/10.1016/j.neuropharm.2019.107898. URL https://www.sciencedirect.com/science/article/pii/S0028390819304691.

[56] M. Maturana, C. Meisel, Katrina L. Dell, Philippa J. Karoly, W. D'Souza, D. Grayden, A. Burkitt, P. Jiruška, J. Kudláček, J. Hlinka, M. Cook, L. Kuhlmann, and D. Freestone. Critical slowing down as a biomarker for seizure susceptibility. *Nature Communications*, 11(1):2172, 2020. doi: 10.1038/s41467-020-15908-3. URL https://doi.org/10.1038/s41467-020-15908-3.

[57] Nhan Duy Truong, Anh Duy Nguyen, Levin Kuhlmann, Mohammad Reza Bonyadi, Jiawei Yang, Samuel Ippolito, and Omid Kavehei. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural networks*, 105:104–111, 2018.

[58] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Neural Information Processing Systems*, 36: 26304–26321, 2023. doi: 10.52202/075280-1144.

[59] Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Neural Information Processing Systems*, volume 37, pages 39249–39280. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1239.

[60] Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2302.14367.

[61] Tongtian Yue, Xuange Gao, Shuning Xue, Yepeng Tang, Longteng Guo, Jie Jiang, and Jing Liu. BrainGPT: Unleashing the Potential of EEG Generalist Foundation Model by Autoregressive Pre-training, 2024. URL https://arxiv.org/abs/2410.19779.

[62] Tianning Feng, Juntong Ni, Wei Jin, and Ezequiel Gleichgerrcht. SeizureFormer: A Multi-Scale Transformer for Seizure Risk Forecasting from RNS-Derived Biomarkers. In *Biocomputing 2026*, page 85–98. WORLD SCIENTIFIC, December 2025. doi: 10.1142/9789819824755_0007. URL http://dx.doi.org/10.1142/9789819824755_0007.

[63] Yanlong Chen, Mattia Orlandi, Pierangelo Maria Rapa, Simone Benatti, Luca Benini, and Yawei Li. Physiowave: A multi-scale wavelet-transformer for physiological signal representation, 2025. URL https://arxiv.org/abs/2506.10351.

[64] Ran Liu, Ellen L Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. *arXiv preprint arXiv:2309.05927*, 2023.

[65] Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals. *ACM Transactions on Computing for Healthcare*, 7(3):1–43, 2026.

[66] Hyungjun Yoon, Seungjoo Lee, Yu Wu, Xiaomeng Chen, Taiting Lu, Freddy Liu, Taeckyung Lee, Hyeongheon Cha, Haochen Zhao, Gaoteng Zhao, et al. Beyond hearing: Learning task-agnostic exg representations from earphones via physiology-informed tokenization. In *International Conference on Learning Representations*, volume 2026, pages 34511–34534, 2026.

[67] Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *Advances in Neural Information Processing Systems*, 38:41179–41212, 2026.

[68] Na Lee, Alexandros Koliousis, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A comprehensive review of biosignal foundation models. *Journal of Neural Engineering*, 2026.

[69] Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG - adapting to any setup with large-scale pretraining on 25,000 subjects. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ZeFMtRBy4Z.

[70] Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: Efficient and topology-agnostic foundation model for eeg signal analysis. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 70682–70708. Curran Associates, Inc., 2025. doi: 10.52202/085713-2377. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/66969a9e6bd7a26dfeccea7227178ca7-Paper-Conference.pdf.

[71] Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-NTFM: A Unified Foundation Model for EEG Signal Representation Learning, 2026. URL https://arxiv.org/abs/2509.24222.

[72] Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between Language and EEG Signals, 2025. URL https://arxiv.org/abs/2409.00101.

[73] Sam Gijsen and Kerstin Ritter. EEG-Language Pretraining for Highly Label-Efficient Clinical Phenotyping, 2025. URL https://arxiv.org/abs/2409.07480.

[74] Iyad Obeid and Joseph Picone. The Temple University Hospital EEG Data Corpus. In *Frontiers in Neuroscience*, volume 10, page 196. Frontiers Media SA, 2016. doi: 10.3389/fnins.2016.00196.

[75] Paolo Detti. Siena Scalp EEG Database. *PhysioNet*, August 2020. doi: 10.13026/5d4a-j060. URL https://doi.org/10.13026/5d4a-j060. Version 1.0.0.

[76] Mariana Abreu, Ana Sofia Carmo, A. Peralta, Francisca Sá, Hugo Plácido da Silva, C. Bentes, and Ana Fred. PreEpiSeizures: description and outcomes of physiological data acquisition using wearable devices during video-EEG monitoring in people with epilepsy. *Frontiers in physiology*, 14:1248899, 2023. doi: 10.3389/fphys.2023.1248899.

[77] Sonya B Dumanis, Jaqueline A French, Christophe Bernard, Gregory A Worrell, and Brandy E Fureman. Seizure forecasting from idea to reality. Outcomes of the my seizure gauge epilepsy innovation institute workshop. *eneuro*, 4(6), 2017.

[78] John M Bernabei, Adam Li, Andrew Y Revell, Rachel J Smith, Kristin M Gunnarsdottir, Ian Z Ong, Kathryn A Davis, Nishant Sinha, Sridevi Sarma, and Brian Litt. HUP iEEG Epilepsy Dataset. *OpenNeuro*, 2022. doi: 10.18112/openneuro.ds004100.v1.1.1.

[79] J Kuo and C Lee-Messer. The Stanford EEG Corpus: A Large Open Dataset of Electroencephalograms from Children and Adults to Support Machine Learning Technology. In *IEEE Signal Processing in Medicine and Biology Symposium*, pages 1–2. IEEE, IEEE, 2017. doi: 10.1109/SPMB.2017.8257034.

[80] M. Ihle, H. Feldwisch-Drentrup, C. Teixeira, A. Witon, B. Schelter, J. Timmer, and A. Schulze-Bonhage. EPILEPSIAE – A European Epilepsy Database. *Comput. Methods Programs Biomed.*, 106(3):127–138, 2012. doi: 10.1016/j.cmpb.2010.08.011.

[81] Mark J Cook, Terence J O'Brien, Samuel F Berkovic, Michael Murphy, Andrew Morokoff, Gavin Fabinyi, Wendyl D'Souza, Raju Yerra, John Archer, Lucas Litewka, et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *The Lancet Neurology*, 12(6):563–571, 2013.

[82] Wikimedia Commons. File: Lobes of the brain NL.svg. https://commons.wikimedia.org/wiki/File:Lobes_of_the_brain_NL.svg, 2026. SVG derivative by user Mysid after H. V. Carter, *Gray's Anatomy* (1918), plate 728; public domain. Accessed September 8, 2026.

[83] Laurens R. Krol. EEG 10-10 System with Additional Information, 2020. URL https://commons.wikimedia.org/wiki/File:EEG_10-10_system_with_additional_information.svg. Licensed under CC BY-SA 4.0. Accessed April 11, 2025.

[84] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23), June 2000. ISSN 1524-4539. doi: 10.1161/01.cir.101.23.e215. URL http://dx.doi.org/10.1161/01.cir.101.23.e215.

[85] Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. Knowledge-Based Construction of Confusion Matrices for Multi-Label Classification Algorithms using Semantic Similarity Measures, 2021. URL https://arxiv.org/abs/2011.00109.

[86] Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital Seizure Detection Corpus. *Frontiers in Neuroinformatics*, Volume 12 - 2018, 2018. ISSN 1662-5196. doi: 10.3389/fninf.2018.00083. URL https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2018.00083.

# A Clinical Background on Epilepsy

Epileptic seizures are caused by simultaneous uncontrolled electrical discharges from millions of nerve cells in the brain. Depending on where these start and which regions of the brain are involved, seizures can vary greatly. A basic distinction is made between focal seizures, which originate in a specific area of the brain, and generalized seizures, which affect both hemispheres of the brain [51]. Focal seizures account for around two-thirds of all seizures and manifest themselves in the form of perceptual disturbances or only subjectively perceived symptoms (auras), unilateral twitching of the face, arm, or leg while fully conscious (focal motor seizures) and/or unconscious behavioral changes, such as smacking or chewing (focal complex seizures). The most well-known seizures (grand mal or tonic-clonic seizures) with loss of consciousness and convulsions throughout the body are classified as generalized seizures. However, there are also milder forms in this category with brief mental lapses (absences) or slight twitching (myoclonic seizures). The International League Against Epilepsy (ILAE) provides a complete list and categorization of all possible types of seizures. [52].

Approximately 50 million people worldwide suffer from epilepsy [1]. The annual incidence in developed countries is up to 50/100,000 new cases and twice as high in developing countries [53]. Epilepsy is therefore the most common chronic neurological disease and occurs just as often as type 1 diabetes, which requires treatment.

Epilepsy occurs at any age. The risk is highest in children and adolescents or in people over 65 years of age. In children, the causes of epilepsy are brain malformations, oxygen deprivation as a result of birth complications, and genetic or metabolic reasons. In older people, stroke, tumors, or brain injuries are triggers of epilepsy [54].

Epilepsy is diagnosed based on a clinical description of seizures provided by patients and their relatives (external anamnesis). In addition, diagnosis is made by measuring brain activity using electroencephalography (EEG), in which electrical impulses are recorded using electrodes attached to the head. Experts can visually distinguish between epilepsy-typical interictal potentials (occurring outside of seizures) and ictal EEG patterns (during seizures). In video EEG monitoring, patients are monitored over several days in an epilepsy monitoring unit (EMU) using continuous EEG and simultaneous video recording to measure seizures and determine the region of the brain from which the seizures originate.

EEG data can be used to predict seizures [55]. A seizure is preceded by minor disturbances in the brain caused by neuronal activation that quickly subside. Once a critical threshold is reached, a whole cascade of neurons is activated, causing the recovery phase to last measurably longer (critical slowing down) and potentially indicating an impending seizure [56].

Measuring the critical slowing down recovery phase as a marker for impending seizures has been the basis for scientific contributions aimed at improving the identification of this phase using AI approaches. It turned out that early deep learning methods (CNNs) were already capable of distinguishing between pre- and interictal states [57].

# B EEG foundation models

The application of foundation models to neurophysiology represents a paradigm shift from manual feature engineering to large-scale representation learning. Inspired by the success of Large Language Models (LLMs), recent approaches treat brain activity as a "language," discretizing continuous signals into tokens to learn the intrinsic grammar of neural dynamics. This transition moves the field from supervised training on small, task-specific datasets to self-supervised pretraining on vast repositories of unlabeled physiological data.

**Architectures and Objectives:** The dominant architecture in this domain has been the Transformer. Models employing **Plain Masking** strategies include Brant [58], CBraMod [22], BIOT [21], LaBraM [20], EEGPT [59], and BrainBERT [60]. These typically use bidirectional attention to reconstruct masked segments. While effective for retrospective analysis, they are computationally expensive ($\mathcal{O}(N^2)$) and less inherently suited for forecasting. CBraMod uses Axial Attention: $\mathcal{O}(N\sqrt{NC})$.

**Forecasting and Generative Modeling:** Recognizing the need for proactive seizure management, the field is shifting towards **Causal Masking** and probabilistic forecasting. This category includes NeuroGPT [23], BrainWave (Brant-2) [24], BrainGPT [61], and SeizureFormer [62]. A key distinc-

tion among these is their inference strategy: NeuroGPT utilizes the token sequence history during evaluation. BrainWave incorporates both look-back and future tokens in its evaluation protocol. By modeling the transition from interictal to preictal states, these systems aim to provide continuous risk assessments. However, when streaming over long histories (e.g., to capture multidien cycles), the memory and latency per update of standard Transformers grow with the context length, even with FlashAttention and key-value caching (Appendix I).

**Heterogeneity and Generalization:** Clinical EEG data is highly heterogeneous, varying in channel layouts and sampling rates. Models like BIOT [21] and LaBraM [20] address this by tokenizing channels independently, allowing for flexible input configurations. However, they typically rely on learned embeddings for a fixed set of channel labels (e.g., standard 10-20 system), which limits their ability to generalize zero-shot to novel or arbitrary sensor placements. NeuroNTP instead encodes electrodes by their 3D coordinates, a direction also taken by recent intracranial models [17–19] (Appendix I). Despite these advances in flexibility, existing foundation models remain largely unimodal. They neglect the Central Autonomic Network (CAN)—the physiological coupling between the brain and heart (ECG)—which often provides early warning signs of seizures.

**NeuroNTP Positioning:** NeuroNTP uses an xLSTM backbone, which scales linearly with sequence length and supports stateful inference with constant memory and latency per update; when streaming over long recordings, this avoids the growing key-value cache of Transformers (Appendix I, Figure 12). Furthermore, NeuroNTP combines causal streaming, temporally aligned clinical text, and seizure-proximity supervision in a single multimodal model (EEG, ECG, $SpO_2$, and clinical text).

**Multimodal Physiological Foundation Models:** Beyond EEG, several recent models pretrain jointly on multiple physiological signals, including SleepFM [25] on polysomnography, PhysioOmni [27] and Biosignal MAE [26] on EEG, ECG, EOG, and EMG, PhysioWave [63] on ECG and EMG, bioFAME [64] on EEG and EOG, NormWear [65] on wearable PPG, ECG, EEG, GSR, and IMU signals, an earphone-based ExG model [66] on ear-EEG, EMG, and EOG, and BrainOmni [67] on EEG and MEG (Table 5); Lee et al. [68] give a broader overview of biosignal foundation models. In principle, their pretrained encoders can be applied to seizure detection by linear probing or finetuning on the EEG channels of CHB-MIT or Siena. Three differences, however, limit a direct comparison with NeuroNTP. First, they are designed for offline analysis of fixed-length segments or recordings with bidirectional or contrastive objectives, rather than for causal streaming inference with a state that carries across hours of monitoring, which seizure prediction relies on. For seizure prediction in particular, they would therefore additionally require a mechanism to carry context across segments over long horizons. Second, several assume a fixed channel set (Table 5, column *Montage Agnostic*), so montages such as the bipolar CHB-MIT montage require remapping or retraining of the input layer. Third, none of them include $SpO_2$ or clinical text, so a comparison is only possible in the biosignal-only setting. We consider SleepFM and PhysioOmni the most relevant candidates, as they model EEG and ECG jointly and provide public weights.

Table 5 provides a comparative overview of recent foundation models in the domain.

Table 5: Comparison of recent Brain Signal Foundation Models. **Montage Agnostic** indicates the ability to handle variable channel layouts zero-shot. Data size is approximate hours (h) of pretraining data.

| Model | Backbone (Complexity) | Objective | Modality | Montage Agnostic | Pretraining Data | Weights |
|---|---|---|---|---|---|---|
| **Brant** [58] | Transformer ($\mathcal{O}(N^2)$) | Masked Modeling | Intracranial | × | Private iEEG (≈2.5k h) | ✓ |
| **CBraMod** [22] | Criss-Cross Trans. ($\mathcal{O}(N\sqrt{NC})$) | Masked Recon. | EEG | ✓ | TUEG (>9k h) | Code |
| **BIOT** [21] | Linformer ($\mathcal{O}(N)$) | Contrastive Learning | EEG | Partly | TUH + Public | ✓ |
| **LaBraM** [20] | VQ-VAE Trans. ($\mathcal{O}(N^2)$) | Spectrum Pred. | EEG | ✓ | TUH (≈2.5k h) | ✓ |
| **EEGPT** [59] | Transformer ($\mathcal{O}(N^2)$) | Masked Recon. + Align. | EEG | × | Mixed Public (≈100 h)* | Code |
| **BrainBERT** [60] | Transformer ($\mathcal{O}(N^2)$) | Masked Spect. | Intracranial | × | Private iEEG (≈44 h) | ✓ |
| **NeuroGPT** [23] | CNN-GPT ($\mathcal{O}(N^2)$) | Masked Recon. | EEG | × | TUH | × |
| **BrainWave** [24] | Transformer ($\mathcal{O}(N^2)$) | Masked Modeling | EEG+iEEG | ✓ | Public + Private (≈41k h) | Upon Pub. |
| **SeizureFormer** [62] | Multi-scale CNN + Trans. ($\mathcal{O}(N^2)$) | Supervised Risk Fcst. | iEEG (RNS) | × | Private RNS counts (≈19–31k days)† | × |
| **REVE** [69] | Transformer ($\mathcal{O}(N^2)$) | Masked Autoencoding | EEG | ✓ | 92 datasets (incl. TUH+Siena) (61k h) | ✓ |
| **LUNA** [70] | Transformer w/ Learned Queries ($\mathcal{O}(S^2)$)‡ | Masked-Patch Recon. | EEG | ✓ | TUEG+Siena (>21k h) | ✓ |
| **Uni-NTFM** [71] | MoE Transformer ($\mathcal{O}(N^2)$) | Multi-view SSL | EEG | ✓ | TUH+Siena+SEED-V (>12k h) | - |
| **NeuroLM** [72] | LLM (GPT-2) ($\mathcal{O}(N^2)$) + VQ | Autoregression + IT | EEG+Text | ✓ | Mixed (incl. TUEG+Siena) (≈25k h) | Code |
| **ELMs** [73] | Dual Enc. (CNN ($\mathcal{O}(N)$) + Trans. ($\mathcal{O}(N^2)$)) | MIL-InfoNCE | EEG+Text | × | TUEG (15k h) | ✓ |
| **SleepFM** [25] | 1D CNN ($\mathcal{O}(N)$) | Contrastive Learning | EEG+EOG+EMG+ECG+Resp. | × | Private PSG (≈100k h) | ✓ |
| **BrainOmni** [67] | Criss-Cross Trans. ($\mathcal{O}(N\sqrt{NC})$) | Masked Code Pred. | EEG+MEG | ✓ | Public EMEG (≈2.6k h) | ✓ |
| **Biosignal MAE** [26] | Transformer ($\mathcal{O}(N^2)$) | MultiMAE | EEG+ECG+EOG+EMG | × | PhysioNet18 (≈15k h) | × |
| **PhysioOmni** [27] | VQ-VAE Trans. ($\mathcal{O}(N^2)$) | Recon. + Masked Code Pred. | EEG+ECG+EOG+EMG | × | Public + Private (≈25k h) | ✓ |
| **NeuroNTP (Ours)** | **xLSTM** ($\mathcal{O}(N)$) | **Causal NTP + Aux.** | **EEG+ECG+SpO$_2$+clinical text** | ✓ | **Private KUK (>150k h)** | ✓ |

* Five task-based datasets: PhysioMI (motor imagery/execution), HGD (High Gamma Dataset), TSU (SSVEP), SEED (emotion) and M3CV (multi-paradigm). The total size is not reported; our estimate from the trial counts and 4 s windows is ≈100 h.

† Not pretrained: supervised forecasting of daily seizure risk (1–14 days ahead) from RNS-detected interictal activity counts (A+B patterns) of 5 patients with 3,030–6,953 days each; labels are derived from Long Episodes (LEs).

‡ scales quadratically with the number of temporal patches $S$

# C Epilepsy datasets

Table 6: **Global Landscape of Epilepsy and Brain Signal Datasets.** Overview of public and private datasets comparing volume, patient cohorts, and available modalities. **KUK** (Ours) represents a magnitude shift in multimodal volume compared to existing benchmarks.

| Dataset | Modalities | Volume | #Pat. | Age | #Sz | Setting / Notes |
|---|---|---|---|---|---|---|
| ***Public Benchmarks*** | | | | | | |
| **TUH (TUEG)** [74] | EEG | ∼27k h | ∼15k | $51 \pm 56$ | - | Hospital (diverse) |
| **TUH (TUSZ)** [74] | EEG | ∼5k h | 642 | $51 \pm 56$ | ∼3k | Hospital (Seizure Corpus) |
| **CHB-MIT** [32] | EEG | ∼980 h | 23 | $10 \pm 5$ | 198 | Hospital (Pediatric) |
| **SeizeIT2** [15] | Wearable (EEG/ECG/EMG) | >11k h | 125 | - | 883 | Ambulatory / vEEG |
| **Siena Scalp** [75] | EEG | ∼141 h | 14 | $44 \pm 14$ | 47 | Hospital long-term |
| **PreEpiSeizures** [76] | Wearable (ECG/Resp/EDA) | ∼5.8k h | 59 | [16, 60] | 348* | Wearable / vEEG |
| **My Seizure Gauge** [77] | Wearable (EDA/PPG/ACC) | ∼2.6k h | 27 | - | 226* | Wearable / Ambulatory |
| **HUP iEEG** [78] | iEEG | ∼1.3k h | 58 | - | - | Hospital (DRE) |
| ***Restricted / Private Cohorts*** | | | | | | |
| **Stanford EEG** [79] | EEG | ∼105k h | ∼12k | - | ∼19k | Hospital |
| **EPILEPSIAE** [80] | iEEG / Scalp EEG | ∼6.5k h | 30 | $41 \pm 14$ | 276 | Hospital long-term |
| **NeuroVista** [81] | iEEG | ∼130k h¶ | 15 | $44 \pm 13$ | - | Ambulatory (Implant) |
| **Emory RNS** [62] | RNS (iEEG) | ∼460–740k h§ | 5 | - | LEs † | Electrographic |
| **Brant iEEG** [58] | iEEG | ∼2.5k h | 26 | - | - | Hospital (DRE) |
| **KUK (Ours)** | **EEG / ECG / $SpO_2$ / Text** | **>153k h** | **1,302** | **$39 \pm 18$** | **∼6k‡** | **Hospital (vEEG)** |

* Includes subclinical seizures.
† Uses Long Episodes (LEs) as seizure surrogates.
‡ EEG-confirmed seizures co-occurring with clinical markers.
§ Range of RNS monitoring time for 5 patients with 3,030–6,953 days each (≈19–31k patient-days); the data are daily detection counts, not continuous iEEG.
¶ Estimated from 15 patients with >1 year per patient.

Table 6 shows an overview of pretraining datasets commonly used. Existing foundation models predominantly rely on the TUH Corpus,

For a more extensive review of the landscape of current datasets we refer the reader to Wong et al. [8].

# D The KUK Pretraining Dataset

The Kepler Universitätsklinikum (KUK) dataset is a massive, private, multimodal clinical dataset used for pretraining the NeuroNTP foundation model. It was collected retrospectively at the Department of Neurology, JKU Linz.

**Scale and Scope.** The dataset represents one of the largest continuous epilepsy cohorts available for research, comprising over **153,000 hours** (≈18 years) of recording and exceeding **16 TB** of raw data (∼11 TB compressed). It includes **1,302 patients** (664 male, 638 female) spanning a demographic range from 3 to 89 years old across **9,780 recording sessions** (Figure 4 and 7).

**Multimodal Data.** Unlike standard EEG-only datasets, KUK is inherently multimodal. **Neural data** consists of 29 EEG channels (international 10–20 system with extended frontotemporal coverage, Figure 8). **Physiological signals** include single-lead ECG, $SpO_2$ (peripheral oxygen saturation), and 2 EOG (electrooculogram) channels. **Textual data** comprises temporally aligned clinical text annotations and unstructured medical reports, enabling context-aware learning.

**Clinical Annotations and Text.** The dataset is enriched with **680,312 text annotations** structured as `<marker>; <state>; <notes>` tuples. These include **101 distinct marker types** describing events such as seizure delimiters (e.g., *Anfall*, *EEG-Beginn*), electrophysiological findings (e.g., *Spike*, *Polyspikes*), and localization (e.g., *frontal*, *rechts*). Additionally, there are **22,141 unique free-text notes** capturing context (e.g., "*Pflege kommt*") or patient behavior.

**Seizure Characteristics.** Reflecting real-world clinical imbalance, seizure activity constitutes less than **0.22%** of total recording time. Only 297 patients (22.8%) had EEG-confirmed seizures, totaling approximately **10,000 events**. While TUH rivals KUK in broad diversity, KUK offers a uniquely detailed representation of extratemporal (Parietal, Occipital, Frontal) and generalized semiologies, exceeding the scope of focal-only (SeizeIT2) or pediatric-only (CHB-MIT) datasets.

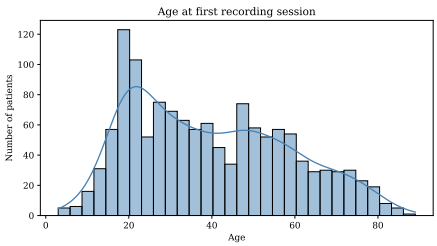


Figure 4: Distribution of the age of patients at the time of the first recording session

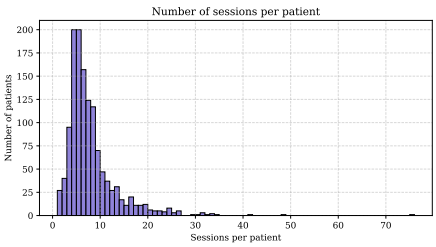


Figure 5: Distribution of the number of sessions per patient

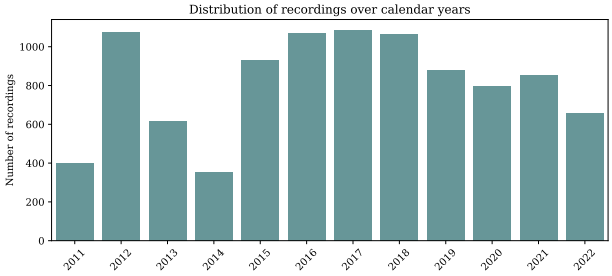


Figure 6: Number of recorded sessions over years

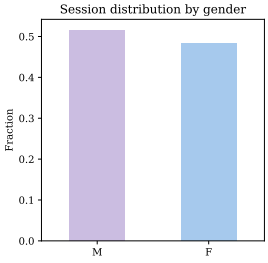


Figure 7: Gender distribution

**Evaluation Subset.** For evaluation, we construct a held-out subset of 60 patients using uniform sampling across the 11-year acquisition period. This ensures temporal representativeness and captures variability in equipment, clinical protocols, and annotation practices.

**Technical Specifications.** All signals are resampled at **256 Hz** (original sampling rates up to 1000 Hz). Due to privacy regulations, the raw dataset is private; however, the model weights (NeuroNTP) are released.

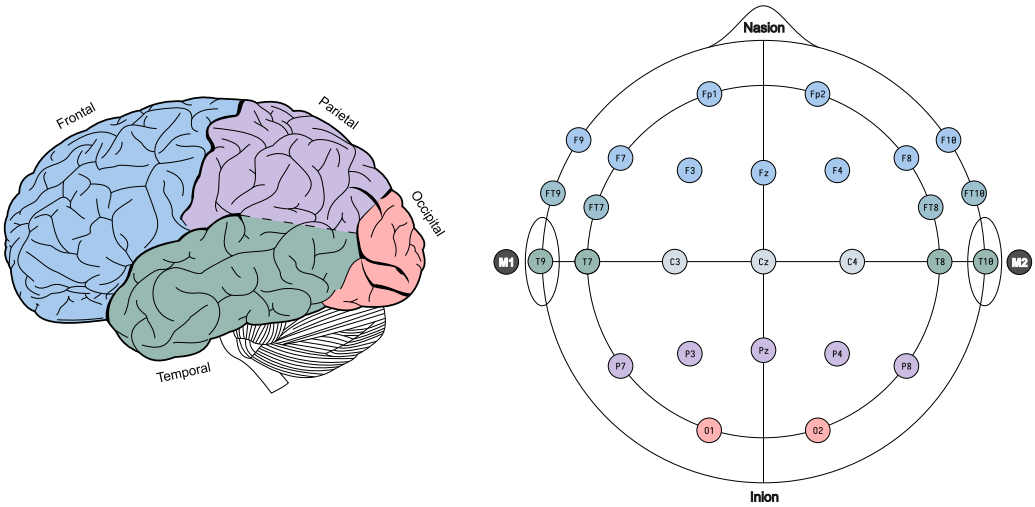


Figure 8: Cortical brain lobes (left) and the extended 10–20 EEG montage used in acquisition of the KUK dataset (right). Correspondence between brain regions and electrode positions is color-coded. The extension includes additional fronto-temporal electrodes FT7, FT8, FT9, FT10, F9, F10, T9, T10. Illustration adapted from *Gray's Anatomy* [82]. Montage adapted from [83] under CC BY-SA 4.0.

### D.1 Clinical annotations in the KUK dataset

The clinical annotations comprise two complementary forms. Structured markers encode standardized diagnostic and procedural cues, including (i) temporal delimiters of events (e.g., `EEG-Beginn`, `Anfall`, `EEG-Ende`), (ii) laterality or localization descriptors (e.g., `frontal`, `rechts`, `temporal`), (iii) electrophysiological findings and EEG-related procedural or state markers (e.g., `Spike`, `Polyspikes`, `HF-Anstieg`, `HV`, `REM`, `Artefakt`), and (iv) other (e.g., `REC`, `VID`, `MON`).

In contrast, free-text notes are unstructured and exhibit substantially higher lexical and semantic variability, capturing spontaneous clinical observations or technical remarks. Typical examples include *Pflege kommt* ("nursing staff enters"), *liegt im Bett und spielt mit Handy* ("lying in bed, playing with phone"), or medication records such as *Keppra 500mg*.

Each annotation comprises three core fields: `marker_type`, specifying the event category (e.g., `EEG-Beginn`, `Artefakt`); `rec_state`, indicating recording status (e.g., `on`, `off`); and `notes`, containing optional free-text descriptions.

Overall, the dataset comprises 101 distinct structured marker types and 22,141 unique free-text note variants, highlighting the breadth of standardized clinical events and the diversity of unstructured clinical documentation.

Importantly, individual seizure-related markers do not unambiguously identify ictal segments. Following clinician-provided annotation guidelines, onset delimiters (e.g., `EEG-Beginn` for EEG or `Anfallsbeginn` for video) are considered valid only when preceded by a clinician-entered `Anfall` marker within three seconds. Thus, ictal/interictal segmentation is derived from clinician-confirmed annotation sequences rather than individual markers.

**Marker groups for annotation-assisted evaluation.** For the annotation-assisted seizure detection experiments on the KUK dataset (Table 1), the annotations provided as input are organized into three groups that regroup the marker categories (i)-(iv) above (Table 7). *Seizure-related* combine the temporal delimiters of events (i) and laterality/localization descriptors (ii) with the associated free-text notes, which may indicate ictal segments. *EEG phenomena* comprise the electrophysiological findings and EEG-related markers (iii). *General/technical* covers the session-level and technical markers (iv), such as recording start and calibration, as well as general, non-seizure behavioral observations. A comprehensive mapping of all marker types to groups is provided with the released code.

Table 7: Annotation groups used as additional input in the annotation-assisted seizure detection experiments. Representative markers are shown; the complete mapping is provided in the supplementary materials.

| Group | Categories | Markers |
|---|---|---|
| Seizure-related | (i), (ii), notes | `Anfall`, `Anfallsbeginn`, `EEG-Beginn`, `EEG-Ende`, `Beginn`, `Ende`, `links`, `rechts`, `beidseits`, `frontal`, `temporal`, ... |
| EEG phenomena | (iii) | `Spike`, `Polyspikes`, `HF-Anstieg`, `IMP`, `REM`, `non-REM`, `FLA`, `HV`, `PAT`, `Artefakt`, ... |
| General/technical | (iv) | `REC`, `VID`, `MON`, `Aa` (eyes open), `Az` (eyes closed), ... |

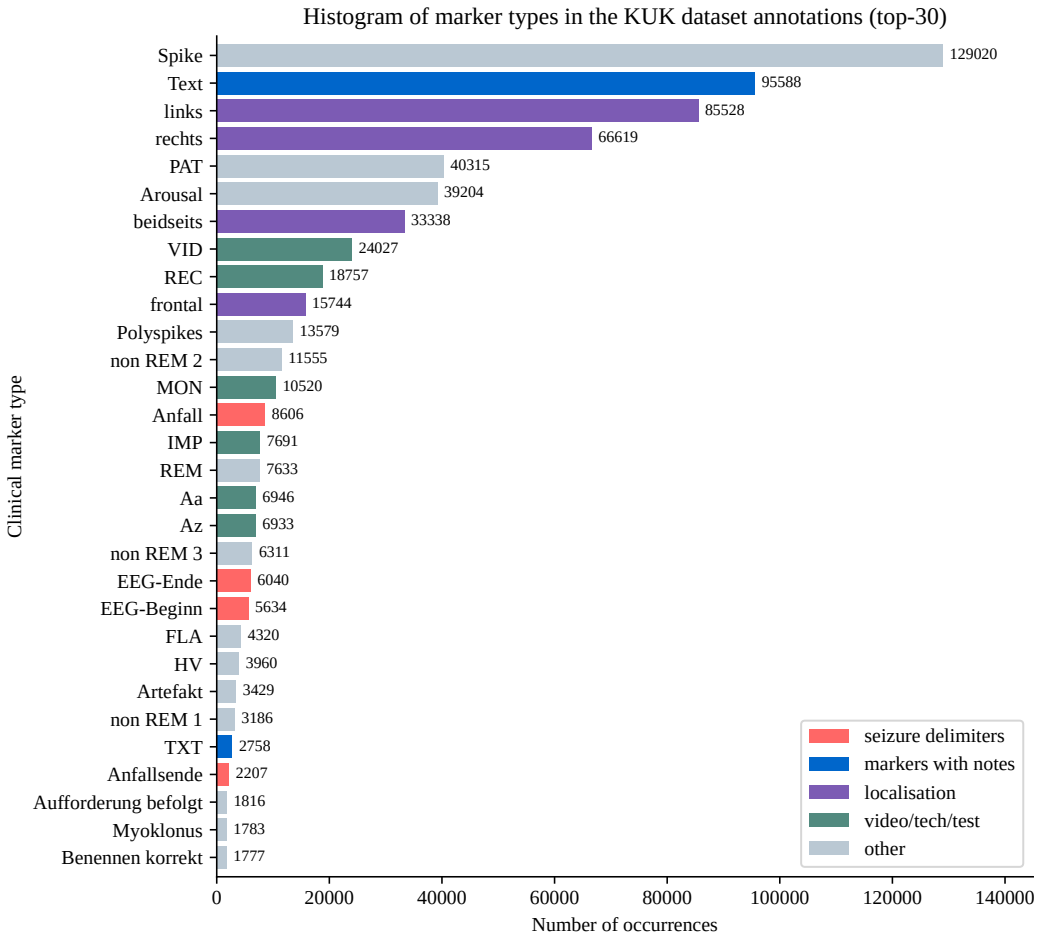


Figure 9: Histogram of the top-30 marker types in the KUK multimodal seizure corpus. Markers of ’Text’ and ’TXT’ types always contain further clinical notes.

## E Seizure prediction results

**Setup.** We frame seizure prediction as binary classification: for each input sequence, the model predicts whether a seizure onset occurs within a future horizon of 0–15, 15–30 or 30–45 min. Models are evaluated on the KUK held-out patients with the same patient-wise cross-validation as for seizure detection (Table 1). We compare two input configurations of NeuroNTP: *biosignal-only* (EEG, ECG and $SpO_2$) and *annotation-assisted*, which additionally receives the time-past-seizure (TPS) derived from seizure annotations. The backbone is either finetuned end-to-end or kept frozen (linear probing). All baselines are trained on biosignal-only input. Besides BACC, AUPRC and AUROC, we report $\Delta$AP, the AUPRC minus the positive rate (chance level of AUPRC: 24.1%, 26.9% and 28.7% for the three horizons).

**Results.** Table 8 lists all results; Figure 10 compares NeuroNTP with the strongest baseline per horizon and metric. With biosignal-only input, NeuroNTP is on par with or below the strongest baseline (AUPRC within one std at all horizons; AUROC 9.1 points below BIOT at 0–15 min). With annotation-assisted input, the finetuned NeuroNTP achieves the best value in every metric and horizon, and its margin over the strongest baseline grows with the horizon (AUROC: $+0.2$, $+5.6$ and $+13.0$ points; AUPRC: $+21.0$, $+23.4$ and $+28.1$ points). With a frozen backbone, the annotation-assisted model remains within one std of the finetuned model at 0–15 min (78.8% vs. 81.7% AUPRC) but falls behind at 30–45 min. Since the baselines do not receive TPS, part of this gain may stem from the TPS input itself rather than from the learned representation; the self-predicted TPS check below shows that the gains require TPS of sufficient quality.

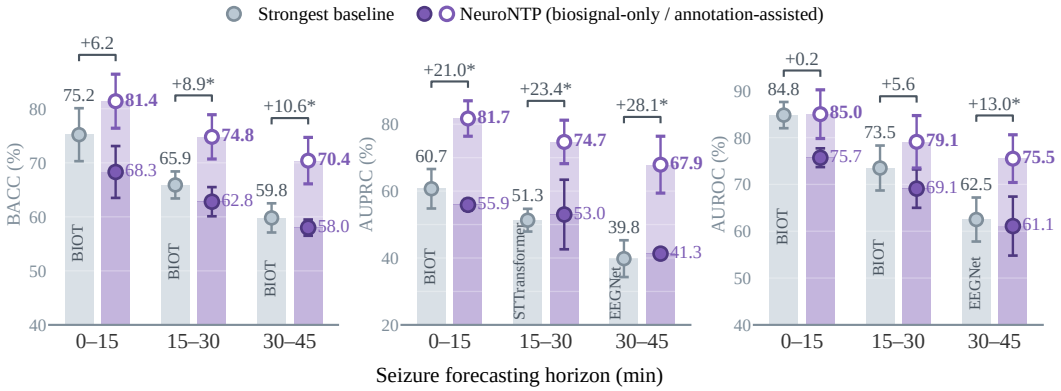


Figure 10: **Seizure prediction on KUK.** BACC, AUPRC and AUROC per forecasting horizon: NeuroNTP vs. the strongest baseline for each horizon and metric (name inside the grey bar). Filled marker: biosignal-only; open marker and light bar: annotation-assisted, i.e., with time-past-seizure (TPS) as additional input. Both NeuroNTP variants are finetuned; baselines receive biosignals only. Chance level: 50% for BACC and AUROC, 24–29% (positive rate) for AUPRC. Error bars: $\pm 1$ std. Brackets: gain of the annotation-assisted model over the baseline; * $p < 0.05$ (two-sided Welch's $t$-test; $n = 4$ folds). Full results in Table 8.

**Self-predicted TPS (preliminary).** The annotation-assisted prediction results (Table 8, Figure 10) use time-past-seizure (TPS) values derived from expert seizure annotations. As a preliminary check of whether these can be replaced by model estimates, we ran a two-pass inference: the model first estimates TPS from the biosignals, and this estimate is then fed back as input for seizure prediction. The first TPS value of each sequence is initialized from the annotations to provide a temporal origin. The check used a closely related checkpoint rather than the exact model of Table 8 and a single run. At 0–15 min, performance dropped to 74.1% BACC, 64.0% AUPRC and 77.7% AUROC (vs. 81.4%, 81.7% and 85.0% with annotated TPS; 68.3%, 55.9% and 75.7% biosignal-only). At 15–30 and 30–45 min it fell to the level of the biosignal-only model or below. The annotation-assisted gains thus require TPS of sufficient quality. In clinical practice, the time of the last seizure is often documented (e.g., in the epilepsy monitoring unit), and it could also be derived from NeuroNTP's own seizure detection (Table 1), a dedicated seizure detector, or timer-based tracking of detected ictal events; evaluating these sources is a natural next step.

Table 8: **Seizure Prediction Performance.** Evaluation across forecasting horizons (15, 30, and 45 minutes). Comparison of NeuroNTP against baselines. The annotation-assisted input configuration includes TPS (time-past-seizure) values, and the biosignal-only configuration only relies on EEG, ECG and SpO2 input modalities; finetuned/frozen designates backbone finetuning for the prediction task. Values are reported in % (mean ± std). Purple denotes max and grey within one std of max.

| **Method** | **Forecasting Horizon** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **0–15 min** | | | | **15–30 min** | | | | **30–45 min** | | | |
| | **BACC** | **AUPRC** | **ΔAP** | **AUROC** | **BACC** | **AUPRC** | **ΔAP** | **AUROC** | **BACC** | **AUPRC** | **ΔAP** | **AUROC** |
| BIOT | $75.2^{\pm 4.9}$ | $60.7^{\pm 5.9}$ | 36.6 | $\mathbf{84.8}^{\pm 2.8}$ | $65.9^{\pm 2.5}$ | $46.7^{\pm 2.7}$ | 19.8 | $\mathbf{73.5}^{\pm 4.8}$ | $59.8^{\pm 2.7}$ | $39.2^{\pm 2.3}$ | 10.5 | $62.1^{\pm 2.7}$ |
| STTransformer | $71.8^{\pm 1.1}$ | $51.9^{\pm 4.2}$ | 27.8 | $79.3^{\pm 1.9}$ | $64.3^{\pm 1.2}$ | $51.3^{\pm 3.4}$ | 24.5 | $70.4^{\pm 1.7}$ | $58.0^{\pm 5.8}$ | $39.4^{\pm 5.7}$ | 10.6 | $61.9^{\pm 7.7}$ |
| FFCL | $70.0^{\pm 4.7}$ | $46.4^{\pm 6.3}$ | 22.4 | $76.4^{\pm 5.5}$ | $59.9^{\pm 2.1}$ | $42.7^{\pm 12.4}$ | 15.9 | $65.7^{\pm 6.1}$ | $59.6^{\pm 4.0}$ | $35.9^{\pm 3.4}$ | 7.2 | $59.9^{\pm 5.1}$ |
| SPaRCNet | $69.4^{\pm 3.5}$ | $45.6^{\pm 10.3}$ | 21.6 | $77.1^{\pm 5.3}$ | $63.2^{\pm 3.9}$ | $40.9^{\pm 6.6}$ | 14.0 | $68.2^{\pm 4.5}$ | $58.5^{\pm 1.5}$ | $36.5^{\pm 5.9}$ | 7.8 | $60.8^{\pm 4.3}$ |
| Pytorch-TCN | $68.8^{\pm 4.2}$ | $43.7^{\pm 5.7}$ | 19.7 | $73.3^{\pm 5.0}$ | $58.7^{\pm 5.2}$ | $41.4^{\pm 6.5}$ | 14.5 | $64.8^{\pm 5.6}$ | $56.2^{\pm 5.9}$ | $38.2^{\pm 8.2}$ | 9.5 | $58.3^{\pm 8.8}$ |
| EEGNet | $67.8^{\pm 5.2}$ | $43.7^{\pm 3.9}$ | 19.6 | $76.0^{\pm 3.2}$ | $62.4^{\pm 6.4}$ | $43.6^{\pm 9.1}$ | 16.7 | $67.9^{\pm 2.9}$ | $59.0^{\pm 3.0}$ | $39.8^{\pm 5.5}$ | 11.1 | $62.5^{\pm 4.7}$ |
| CNNTransformer | $66.1^{\pm 3.6}$ | $41.5^{\pm 9.5}$ | 17.5 | $70.4^{\pm 2.5}$ | $57.3^{\pm 2.9}$ | $35.3^{\pm 6.9}$ | 8.4 | $59.4^{\pm 6.4}$ | $55.5^{\pm 6.1}$ | $36.5^{\pm 10.6}$ | 7.8 | $59.1^{\pm 10.3}$ |
| ContraWR | $65.4^{\pm 9.0}$ | $40.8^{\pm 7.0}$ | 16.7 | $72.7^{\pm 7.0}$ | $59.2^{\pm 4.3}$ | $49.6^{\pm 7.3}$ | 22.7 | $71.4^{\pm 0.6}$ | $56.9^{\pm 6.0}$ | $36.8^{\pm 5.9}$ | 8.1 | $59.2^{\pm 7.2}$ |
| **NeuroNTP (ours)*** | | | | | | | | | | | | |
| *annotation-assisted, finetuned* | $\mathbf{81.4}^{\pm 5.0}$ | $\mathbf{81.7}^{\pm 5.3}$ | **57.7** | $\mathbf{85.0}^{\pm 5.2}$ | $\mathbf{74.8}^{\pm 4.1}$ | $\mathbf{74.7}^{\pm 6.5}$ | **47.9** | $\mathbf{79.1}^{\pm 5.6}$ | $\mathbf{70.4}^{\pm 4.3}$ | $\mathbf{67.9}^{\pm 8.5}$ | **39.1** | $\mathbf{75.5}^{\pm 5.1}$ |
| *annotation-assisted, frozen* | $\mathbf{81.2}^{\pm 2.7}$ | $\mathbf{78.8}^{\pm 3.9}$ | 54.8 | $\mathbf{82.3}^{\pm 3.6}$ | $63.8^{\pm 11.6}$ | $\mathbf{70.5}^{\pm 7.2}$ | 43.6 | $\mathbf{76.9}^{\pm 6.1}$ | $54.7^{\pm 7.9}$ | $52.3^{\pm 15.7}$ | 23.6 | $65.4^{\pm 9.3}$ |
| *biosignal-only, finetuned* | $68.3^{\pm 4.8}$ | $55.9^{\pm 1.7}$ | 31.8 | $75.7^{\pm 2.0}$ | $62.8^{\pm 2.7}$ | $53.0^{\pm 10.4}$ | 26.1 | $69.1^{\pm 4.1}$ | $58.0^{\pm 1.5}$ | $41.3^{\pm 1.3}$ | 12.6 | $61.1^{\pm 6.3}$ |

* Results for NeuroNTP are based on a preliminary hyperparameter configuration.

# F Downstream datasets, finetuning and evaluation

## F.1 CHB-MIT dataset and evaluation protocol

**Dataset.** We evaluate seizure detection on the publicly available CHB-MIT dataset [32, 84], which comprises long-term scalp EEG recordings from 23 pediatric subjects with intractable seizures (24 cases: chb01–chb24, with cases chb01 and chb21 obtained from the same pediatric female patient with a 1.5 year gap between the sessions[1]). Recordings span up to several days and follow the international 10–20 electrode placement system, sampled at 256 Hz.

**Reference protocol.** For comparability with prior work, we reproduce the preprocessing and evaluation protocol of CBraMod [22], based on the publicly available implementation originally introduced in BIOT [21]. Specifically, we adopt the standard 16-channel bipolar montage and segment recordings into 10-second windows, resulting in $N = 326{,}993$ samples. We verify both the sample count and the ictal/non-ictal distribution of the test set, and use the same subject-wise split as the reference implementation, with subjects 22–23 (cases chb23–chb24) forming the test set.

**Finetuning.** Both finetuning and evaluation are performed with biosignal-only input configuration, without clinical markers, notes or TPS. During finetuning, we randomly sample from the training subset with emphasis on 15-minute preictal–ictal and ictal–postictal regions. The pretrained backbone is first kept frozen while training a lightweight 3-layer MLP head (learning rate $5 \times 10^{-4}$), followed by end-to-end optimization of the full network with a reduced learning rate ($10^{-5}$).

**Preprocessing considerations.** The preprocessing pipeline adopted in BIOT [21] builds on earlier work and existing implementations[2], which use a subset of recordings from subjects 1–20. Seizure regions are augmented using overlapping windows (5-second stride), a common strategy to mitigate class imbalance. In this scheme, windows containing seizure onset or offset are labeled as positive, while windows fully contained within ictal intervals may be labeled as negative. While this augmentation improves class balance, it affects all subsets uniformly and modifies the original acquisition-time distribution. These design choices are well-established in prior work and are effective for benchmarking, but may interact differently with models that explicitly leverage long-range temporal structure.

**Extended evaluation.** While we follow the reference protocol for strict comparability, we additionally assess the impact of these design choices by modifying the training data, labeling scheme, and evaluation distribution while keeping the reference test split fixed. Specifically, we consider: (i) training on all available recordings from subjects 1–19, (ii) labeling all ictal windows as positive, (iii) combining both modifications, and (iv) evaluating on a non-augmented test set without overlapping windows to preserve the original acquisition-time distribution.

These variants allow us to quantify the influence of data selection, labeling assumptions, and augmentation strategies on model performance under controlled conditions. This is particularly relevant in our setting, where the pretrained model leverages long-range temporal structure and may benefit from more consistent labeling of ictal segments and realistic evaluation distributions.

**Extended results.** We observe in Table 9 that modifying the training data and labeling scheme has a substantial impact on performance. While adjusting the labeling primarily improves AUPRC, expanding the training set to all available recordings yields large gains across all metrics. Combining both modifications results in the strongest performance, achieving $97.3\%$ AUROC, $75.3\%$ AUPRC, and $87.7\%$ BACC.

When evaluating on a non-augmented test set that preserves the original acquisition-time distribution, the model achieved $96.1 \pm 0.1\%$ AUROC, $58.4 \pm 0.7\%$ AUPRC, and $85.2 \pm 0.2\%$ BACC. Performance remained strong, with only moderate reductions in AUPRC and BACC, indicating that the model retains robust generalization without relying on augmentation-induced distribution shifts. These findings suggest that both data coverage and consistent ictal labeling are important factors for effective model training, while augmentation primarily affects the evaluation distribution rather than the underlying representation quality.

[1] https://physionet.org/content/chbmit/1.0.0/

[2] https://github.com/bernia/chb-mit-scalp

Table 9: Extended evaluation of seizure detection on the CHB-MIT dataset using the multimodally pretrained NeuroNTP model, fine-tuned and evaluated with biosignal-only input (no clinical markers, notes, or TPS). Baseline results are taken from Wang et al. [22]. All values are mean ± standard deviation (in %). Purple denotes best and grey within one std of best.

| Method | BACC | AUPRC | AUROC |
|---|---|---|---|
| **Task-specific models** | | | |
| EEGNet | $56.6^{\pm 1.1}$ | $19.1^{\pm 1.8}$ | $80.5^{\pm 1.4}$ |
| EEGConformer | $59.8^{\pm 1.4}$ | $22.1^{\pm 2.1}$ | $82.3^{\pm 1.7}$ |
| SPaRCNet | $58.8^{\pm 1.9}$ | $12.5^{\pm 1.2}$ | $81.4^{\pm 1.5}$ |
| ContraWR | $63.4^{\pm 0.0}$ | $22.6^{\pm 1.7}$ | $81.0^{\pm 1.1}$ |
| CNN-Transformer | $63.9^{\pm 0.7}$ | $24.8^{\pm 2.3}$ | $86.6^{\pm 0.8}$ |
| FFCL | $62.6^{\pm 1.0}$ | $20.5^{\pm 3.5}$ | $82.7^{\pm 0.5}$ |
| ST-Transformer | $59.2^{\pm 1.9}$ | $14.2^{\pm 0.9}$ | $82.4^{\pm 4.9}$ |
| **Foundation models** | | | |
| BIOT | $70.7^{\pm 4.6}$ | $32.8^{\pm 4.6}$ | $87.6^{\pm 2.8}$ |
| LaBraM-Base | $70.8^{\pm 3.6}$ | $32.9^{\pm 4.0}$ | $86.8^{\pm 2.0}$ |
| CBraMod | $74.0^{\pm 2.8}$ | $36.9^{\pm 3.8}$ | $88.9^{\pm 1.5}$ |
| EEGMamba | $74.5^{\pm 2.5}$ | $38.9^{\pm 4.2}$ | $89.4^{\pm 1.6}$ |
| CSBrain | $72.6^{\pm 1.2}$ | $51.6^{\pm 4.5}$ | $89.2^{\pm 3.2}$ |
| **NeuroNTP (ours)** | | | |
| Reference protocol | $84.7^{\pm 1.1}$ | $45.4^{\pm 0.8}$ | $93.8^{\pm 0.2}$ |
| Adjusted labeling | $86.3^{\pm 0.8}$ | $65.9^{\pm 1.1}$ | $94.4^{\pm 0.1}$ |
| Trained on all recordings (subjects 1–19) | $86.3^{\pm 0.2}$ | $56.2^{\pm 0.7}$ | $96.5^{\pm 0.1}$ |
| Trained on all recordings + Adjusted labeling | $\mathbf{87.7}^{\pm 0.1}$ | $\mathbf{75.3}^{\pm 0.3}$ | $\mathbf{97.3}^{\pm 0.1}$ |

### F.2 Siena dataset and evaluation protocol

**Dataset.** We evaluate on the Siena Scalp EEG dataset [75], a clinical video-EEG corpus collected at the University of Siena, comprising recordings from adult patients with epilepsy. Signals follow the international 10–20 electrode placement system and are annotated by expert clinicians according to ILAE guidelines.

**Preprocessing and split.** For comparability with prior work, we follow the evaluation protocol of CSBrain [49]. The EEG signals from the patient recording sessions are segmented into 10-second windows. We adopt a subject-independent split, where the last two subjects (PN16 and PN17) form the test set, and the remaining subjects are split into training and validation subsets using an 8:2 ratio.

**Channel selection.** To ensure consistency across recordings, we restrict the input to a common subset of EEG channels shared across subjects, resulting in a reduced montage of 19 channels covering frontal, central, parietal, occipital, and temporal regions (FP1, FP2, F3, F4, F7, F8, FZ, C3, C4, CZ, P3, P4, PZ, O1, O2, T3–T6), which ensures consistent input dimensionality across subjects.

**Finetuning.** Both finetuning and evaluation are performed with biosignal-only input configuration, using the channel selection specified above. No clinical markers, notes or TPS values added to the input during this process. The pretrained NTP backbone is first kept frozen while adding a lightweight 3-layer MLP head (learning rate $5 \times 10^{-4}$) and training it. In the second stage, the entire network is optimized with a reduced learning rate ($5 \times 10^{-6}$).

# G Clinical marker inference

To obtain a more informative view of agreement and confusion between predicted and reference clinical markers, we employ a semantic-aware multi-label confusion analysis inspired by Turki et al. [85], adapted to our SBERT-based marker representation. Instead of relying on exact token matching or one-vs-rest binarization—which tends to overestimate errors in multi-label settings with semantically overlapping labels—we measure semantic proximity directly in the embedding space. Concretely, cosine similarity between the frozen `GerMedBERT` embeddings of predicted and target marker sentences is used as a continuous measure of semantic agreement.

For a given timestep, let $P = \{p_1, \ldots, p_{N_P}\} \subset \mathbb{R}^d$ denote the set of predicted marker embeddings and $T = \{t_1, \ldots, t_{N_T}\} \subset \mathbb{R}^d$ the corresponding reference set. Their alignment is computed as follows:

A1. **Semantic matching:** The smaller set is matched to the larger one by assigning each element to its most similar counterpart in the opposite set, maximizing cosine similarity.

A2. **Similarity filtering:** A match is accepted only if the similarity exceeds a threshold $\tau$, i.e., $\mathrm{Sim}(p_i, t_j) \geq \tau$.

A3. **One-to-one constraint:** Each prediction and target can participate in at most one match. Conflicts are resolved greedily by retaining the pair with highest similarity.

A4. **Residual assignments:** Unmatched predictions are treated as *semantic false positives*, while unmatched targets are treated as *semantic false negatives*.

This procedure yields a confusion-like matrix over individual marker tokens, aggregated across all timesteps. It naturally extends to top-$k$ decoding, where the $k$ nearest SBERT-retrieved marker sentences are merged into a predicted set before semantic alignment, allowing a controlled trade-off between precision and recall.

**Similarity threshold.** Cosine similarity serves as a proxy for semantic relatedness between clinical marker tokens. Following the formulation of Turki et al., we define the acceptance threshold as

$$\tau = \tfrac{1}{2}\left(F_M + F_N\right),$$

where $F_M$ and $F_N$ denote empirical upper and lower bounds on similarity for highly related and unrelated marker pairs, respectively. In practice, we set $F_M = 1.0$ and $F_N = 0.0$, yielding $\tau = 0.5$. This choice suppresses spurious matches while preserving semantically meaningful substitutions. Higher values of $\tau$ enforce stricter equivalence, whereas lower values allow more permissive semantic alignment.

## G.1 Interpretation of the multi-label confusion analysis

Clinical marker inference evaluated using the pretrained model and the auxiliary marker reconstruction task, without task-specific finetuning, exhibits a highly structured but strongly imbalanced behavior that reflects both the semantic organization of the marker space and the statistical dominance of certain event types. At the global level, performance is driven primarily by frequent and physiologically salient markers. EEG phenomena are recovered most reliably: `spike` and `polyspikes` dominate the confusion matrix, and are often interchanged, which is consistent with their close clinical semantics and frequent co-occurrence in EEG recordings.

Seizure delimiters form another coherent group, with `eeg-ende`, `anfall`, and `anfallsende` being predicted interchangeably. Notably, seizure endings are detected more robustly than seizure beginnings, suggesting that termination patterns provide stronger or more consistent biosignal cues than onset transitions. The broad distribution of `eeg-ende` predictions further indicates that this marker acts as a dominant contextual anchor, frequently co-occurring with behavioural and laterality-related annotations.

In contrast, laterality and localization markers are rarely recovered in isolation. Their prediction is strongly suppressed by the dominant EEG-phenomena and seizure-delimiter classes, indicating that the model prioritizes globally salient electrophysiological events over spatial modifiers when decoding marker tokens.

Behavioural markers exhibit moderate group-level detection performance but very low exact class accuracy. This suggests that the model recognizes the presence of a behavioural event without reliably distinguishing its precise subtype. Interestingly, markers such as `vokalisation` are predicted correctly when not overshadowed by high-frequency delimiters like `eeg-ende`, while semantically adjacent pairs such as `aufforderung befolgt` and `aufforderung nicht befolgt` are frequently confused, with a systematic bias toward the latter.

Overall, these observations indicate that NeuroNTP learns a clinically meaningful hierarchy of marker relevance: highly stereotyped electrophysiological patterns and seizure delimiters are captured robustly, whereas contextual, spatial, and behavioural descriptors are represented more weakly and are often overridden by dominant event markers. This behavior aligns with both the physiological salience and annotation frequency of the respective categories, and supports the interpretation that clinical text improves performance by anchoring biosignal representations to semantically structured, event-centric concepts rather than to fine-grained descriptive modifiers.

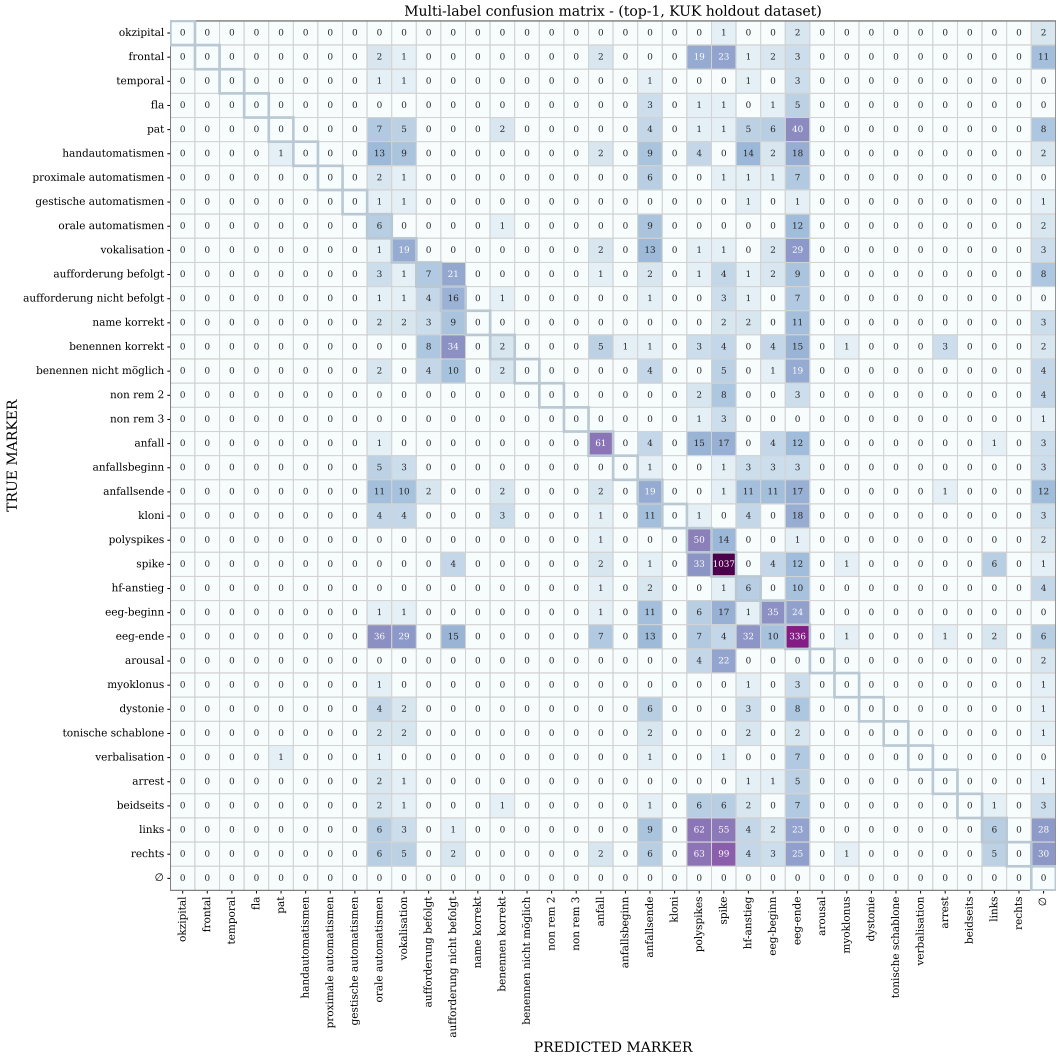


Figure 11: Multi-label marker confusion matrix evaluated without finetuning on the KUK holdout dataset, using top-1 inference. Markers with fewer than 5 occurrences are excluded. True markers are shown on the vertical axis, predicted markers on the horizontal axis. Labels are ordered by semantic proximity in the frozen SBERT embedding space. Cell color intensity reflects the number of predictions.

# H NeuroNTP pretraining using the TUH Seizure Corpus

## H.1 TUSZ preprocessing for the NeuroNTP self-supervised setup

The Temple University Hospital Seizure Corpus (TUSZ) v2.0.3 [86] was incorporated into the NeuroNTP framework to enable self-supervised pretraining on a large public seizure dataset and to support cross-dataset evaluation. TUSZ is part of the TUH EEG corpus[3] and provides seizure-annotated EEG recordings in EDF format with heterogeneous montages and channel configurations.

**EDF conversion and session stitching.** All EDF recordings were converted into the internal HDF-based Epilepsia core-dataset representation used by the NeuroNTP tokenizer and dataset-view pipeline. During conversion, original channels were preserved, channel names and metadata were harmonized, a 60 Hz notch filter and 0.5–100 Hz bandpass filter were applied, signals were resampled to 256 Hz, and amplitudes were converted from $\mu$V to V to match the KUK-pretrained model.

TUSZ recordings are often split into multiple short EDF segments (e.g., `_tNNN`). Since NeuroNTP pretraining requires longer context windows (e.g. $30 \times 10$ s used in our experiments or longer when available), temporally adjacent EDF segments belonging to the same session group were stitched into continuous HDF sessions to reduce sample loss. Sessions below the minimum duration requirement were excluded. The preprocessing step also generated `sessions.csv`, `annotations.csv`, and `skipped.csv` metadata files.

**Dataset statistics.** After stitching and filtering, the converted dataset contained 1,586 sessions, while 56 sessions were skipped. The median session duration was approximately 1,567 s ($\sim$26 min), with a mean duration of 3,114 s ($\sim$52 min), a 75th percentile of 3,760 s ($\sim$63 min), and a maximum duration of 86,376 s ($\sim$24 h). The median number of stitched EDF segments per session was 2, with a mean of 4.45 and a maximum of 32.

The median number of seizures per session was 0, the mean was approximately 2.26, and the maximum was 122. Overall, approximately 31.7% of sessions contained at least one seizure. Signal inspection confirmed stable amplitude ranges after conversion, with typical EEG channels densely centered around zero and most values within approximately $[-2 \times 10^{-4}, 2 \times 10^{-4}]$ V.

**Channel heterogeneity and montage diversity.** TUSZ exhibits substantial channel heterogeneity, with 164 distinct channel names and 57 montage configurations; 99.2% of recordings contain at most 40 channels. The most common 17-channel subset is:

$$\{\text{C3, C4, CZ, F3, F4, F7, F8, FP1, FP2, O1, O2, P3, P4, T3, T4, T5, T6}\}.$$

For NeuroNTP pretraining, we defined a fixed 24-channel configuration including 10–20 EEG channels and EKG:

$$\{\text{FP1, FP2, F7, F3, FZ, F4, F8, C3, CZ, C4, P3, PZ, P4, O1, O2,}$$
$$\text{T1, T2, T3, T4, T5, T6, A1, A2, EKG1}\}.$$

Although NeuroNTP can in principle support variable channel configurations, we restricted sampling to frequent configurations covering the majority of recordings to improve training stability.

**Seizure type handling and text encoding.** Unlike KUK, TUSZ does not provide rich temporally aligned clinical text markers. To preserve architectural compatibility, short seizure type codes (e.g., `cnsz`, `fnsz`, `gnsz`, `bckg`) were expanded to their textual definitions and encoded using the same SBERT-based marker embedding pipeline used for KUK. This preserves compatibility with the marker reconstruction auxiliary head, although explicit seizure-type classification may be more appropriate for future TUSZ-specific training.

**TTS/TPS adaptation.** TTS and TPS targets were derived from the seizure onset times in the TUSZ event annotations, using the same definitions as for KUK: TTS is the time until the next seizure onset, and TPS the time since the onset of the previous seizure, computed per patient across all stitched recordings. Time-to-seizure (TTS) and time-past-seizure (TPS) targets were capped at 7,200 s (2 h). This differs from KUK, where longer temporal horizons are meaningful due to substantially longer continuous recordings.

[3] https://isip.piconepress.com/projects/nedc/html/tuh_eeg/

Table 10: **Effect of the pretraining dataset.** NeuroNTP pretrained on KUK (multimodal) or TUSZ (EEG-only), then fine-tuned and evaluated on CHB-MIT and Siena Scalp EEG with biosignal-only input. All values are mean ± standard deviation (in %). Purple denotes best and grey within one std of best (per dataset).

| | CHB-MIT† | | | Siena Scalp | | |
|---|---|---|---|---|---|---|
| **Pretraining** | **BACC** | **AUPRC** | **AUROC** | **BACC** | **AUPRC** | **AUROC** |
| NeuroNTP@KUK | $\mathbf{86.3}^{\pm 0.2}$ | $\mathbf{56.2}^{\pm 0.7}$ | $\mathbf{96.5}^{\pm 0.1}$ | $\mathbf{87.5}^{\pm 2.1}$ | $\mathbf{46.6}^{\pm 7.4}$ | $\mathbf{93.6}^{\pm 1.4}$ |
| NeuroNTP@TUSZ | $85.7^{\pm 0.8}$ | $44.5^{\pm 1.0}$ | $93.4^{\pm 0.6}$ | $79.4^{\pm 0.7}$ | $40.8^{\pm 0.6}$ | $90.7^{\pm 0.2}$ |

† Fine-tuned on all available recordings from subjects 1–19 (cf. Table 9).

**Normalization and validation.** Quantile normalization statistics were precomputed across the converted dataset to accelerate large-scale pretraining. As a preliminary sanity check, logistic regression on a balanced subset yielded reasonable AUROC and AUPRC values, supporting signal integrity after preprocessing.

**Role in self-supervised pretraining.** The converted TUSZ dataset enables pretraining from scratch within the NeuroNTP framework, continued pretraining from a KUK-initialized model, and controlled cross-dataset generalization analysis. Although representational richness is reduced due to the lack of detailed clinical markers, seizure-type annotations provide structured contextual supervision for detection, forecasting, and seizure-type classification tasks.

**Observed performance and generalization.** Across downstream evaluations (Table 10), NeuroNTP pretrained on TUSZ achieves consistently strong AUROC on both CHB-MIT and Siena, indicating that the model captures robust temporal and morphological seizure characteristics even without access to rich clinical context. However, compared to KUK-pretrained models, performance in AUPRC and balanced accuracy is reduced, reflecting the absence of multimodal information and differences in labeling structure. These results suggest that large-scale EEG-only pretraining provides a solid foundation for cross-dataset generalization, while the inclusion of additional modalities and structured clinical markers substantially improves discrimination in highly imbalanced settings. At the same time, the competitive performance of NeuroNTP@TUSZ confirms that the proposed self-supervised framework transfers effectively across datasets and acquisition protocols.

# I Architectural choices and efficiency considerations

## I.1 Montage-geometry encoding

To enable spatially consistent modeling across heterogeneous EEG montages, NeuroNTP incorporates explicit electrode geometry information directly into the tokenization process. Standardized montage definitions (e.g., `standard_1020` provided by MNE-Python) supply three-dimensional Cartesian coordinates $(x, y, z)$ for each EEG electrode position. Rather than treating channels as fixed symbolic identities, the proposed approach embeds their physical spatial configuration into the latent representation space.

For each EEG channel, the corresponding Cartesian coordinate vector is projected into the model embedding space through a learned linear transformation and added to the channel token representation during multimodal tokenization. This produces a continuous geometry-aware encoding that preserves relative spatial relationships between electrodes while remaining independent of any specific montage layout. Consequently, electrodes occupying nearby anatomical regions obtain similar spatial representations, improving robustness across acquisition protocols and channel configurations.

To further support montage adaptation and transfer across datasets with differing electrode layouts, NeuroNTP employs distance-based initialization in coordinate space. When adapting pretrained channel-specific components of the multimodal tokenizer to a new montage, parameters are initialized from the nearest spatial neighbor in the source montage according to Euclidean electrode distance. This initialization strategy preserves spatial continuity and enables more stable transfer between partially overlapping electrode configurations.

**Relation to prior work.** Using electrode coordinates to handle variable and unseen electrode layouts is an established direction, in particular for intracranial EEG, where electrode positions differ between subjects; seegnificant [17], PopT [18], and BaRISTA [19] all support adaptation to previously unseen electrodes. We therefore do not claim geometry-based adaptation to unseen electrodes as a unique contribution. Our design differs mainly in its mechanism: electrode positions enter as a continuous linear projection of 3D coordinates added to each channel token, and channel-specific tokenizer parameters for a new electrode are initialized from its nearest spatial neighbor, which allows transfer with a frozen backbone (Table 3).

### I.2 Choice of next-token prediction (NTP) backbone

The NeuroNTP framework is modular with respect to the causal sequence model used for next-token prediction (NTP). While the main results are obtained using an xLSTM backbone, the implementation supports interchangeable NTP components, including standard causal Transformers and linear-attention Transformers, and can be extended to other architectures. State-space models such as Mamba [36, 37] are a further natural alternative; like xLSTM, they allow constant-memory recurrent inference. A matched comparison with state-space backbones within NeuroNTP is left for future work.

In preliminary experiments, we observed that both Transformer-based and linear-attention models can be effectively pretrained within the NeuroNTP framework. Representation quality and downstream predictive performance were not fundamentally limited by the choice of backbone, indicating that the proposed multimodal tokenization and training objective are largely architecture-agnostic.

### I.3 Parameterization

The xLSTM NTP backbone used in our experiments contains approximately 37.3M trainable parameters under the hyperparameter configuration reported in Appendix L. The full NeuroNTP parameter count depends on the multimodal tokenizer (MMT), since modality-specific encoders and channel-adaptation modules are instantiated according to the available modalities and channel configurations of each dataset. Consequently, the total number of parameters can vary across downstream datasets and may increase when extending the model to previously unseen montages or additional modalities such as EMG, EDA, or accelerometry.

For example, the CHB-MIT configuration contains approximately 39.4M trainable parameters, including the xLSTM backbone, dataset-specific tokenizer components, auxiliary heads, and downstream classification head. This design keeps the causal sequence model fixed while allowing the input representation to adapt to heterogeneous clinical acquisition settings.

**Scaling trends.** Tables 11 and 12 show how seizure detection after pretraining changes with backbone size and amount of pretraining data. Within the tested range, AUROC and AUPRC tend to improve with both, with diminishing returns beyond 75% of the data. These are single-run trends, not scaling laws. Pretraining uses event-prioritized sampling: per epoch, 30% of the event and 3% of the non-event samples are drawn (Table 24). Since the number of epochs is fixed, larger data subsets also receive more optimization steps, so the effects of data and compute are not separated.

Table 11: Effect of backbone size on seizure detection performance at a matched pretraining budget. Results are reported in %.

| Backbone | BACC | AUPRC | AUROC |
|---|---|---|---|
| 11M | 53.1 | 44.2 | 74.3 |
| 18.6M | 50.5 | 46.7 | 82.5 |
| 37M | 65.1 | 48.4 | 84.6 |

Table 12: Effect of pretraining data scale on seizure detection performance. Architecture, training pipeline, epochs, and hyperparameters are fixed. Patient-level subsets are nested and stratified by recording duration. Evaluation uses identical linear probing on held-out patients. Performance metrics are reported in %.

| Data Scale | NTP Loss | AUPRC | AUROC |
|---|---|---|---|
| 100% | 0.0356 | 41.65 | 79.34 |
| 75% | 0.0735 | 41.09 | 78.28 |
| 50% | 0.0796 | 38.63 | 77.17 |
| 25% | 0.1368 | 31.70 | 71.83 |

Table 13: Comparison of downstream seizure classification performance between Transformer and xLSTM backbones under matched training conditions.

| Backbone | BACC | AUPRC | AUROC |
|---|---|---|---|
| Transformer* | 52.4 | 41.4 | 79.8 |
| xLSTM | **54.8** | **47.7** | **80.6** |

* Down-projected to approximately match the xLSTM parameter count.

Table 14: Autoregressive inference latency (ms) and GPU VRAM consumption (MB) as a function of effective history length for xLSTM, causal Transformer, and linear-attention Transformer architectures. Measurements were obtained on an 80 GB GPU.

| | xLSTM | | Transformer | | Linear TF | |
|---|---|---|---|---|---|---|
| **History** | **latency** [ms] | **VRAM** [MB] | **latency** [ms] | **VRAM** [MB] | **latency** [ms] | **VRAM** [MB] |
| 100 | 7.5 | 3.5 | 1.5 | 2.5 | 4.0 | 23.4 |
| 1,000 | 7.5 | 3.5 | 6.3 | 141.7 | 10.9 | 175.1 |
| 5,000 | 7.7 | 3.5 | 68.2 | 3161.8 | 47.6 | 840.9 |
| 10,000 | 7.8 | 3.5 | 255.6 | 12473 | 105.5 | 1678 |
| 20,000 | 7.8 | 3.5 | 1028 | 49548 | 220.0 | 3343 |
| $\geq$25,000 | 7.8 | 3.5 | OOM† | — | 275.6 | 4178+ |

† Out-of-memory failure under the evaluated hardware configuration.

### I.4 Controlled comparison: xLSTM vs. Transformer

To isolate architectural effects from data scale, we performed a controlled comparison between xLSTM and a causal Transformer under matched conditions: identical multimodal inputs, fixed context length (30 segments, $\sim$1200 tokens), identical pretraining data, and comparable NTP backbone size ($\sim$38M parameters).

During pretraining, the Transformer exhibits faster initial convergence in next-token cosine similarity, whereas xLSTM reaches a slightly higher final plateau. Downstream seizure classification performance remains comparable between the two architectures, with xLSTM demonstrating a consistent but modest advantage across BACC, AUPRC, and AUROC metrics (Table 13).

These results indicate that both architectures are suitable for the NTP component, with only minor differences in final predictive performance.

### I.5 Efficiency in continuous monitoring

While predictive performance is comparable, the architectures differ substantially in their computational characteristics during inference.

xLSTM supports *stateful inference*, enabling constant memory usage and compute per timestep. In contrast, Transformer-based models require maintaining and updating a growing key-value cache, leading to increasing memory and compute cost with sequence length. Memory-efficient attention kernels such as FlashAttention [12] reduce the memory needed for the attention computation itself, but not the size of this cache.

We simulate a continuous monitoring setting with incremental updates (40 tokens per step), comparing autoregressive update latency and GPU VRAM consumption across architectures as the effective history length increases (Table 14).

xLSTM maintains near-constant latency ($\sim$7.5–8 ms) and memory ($\sim$3.5 MB per update), independent of history length. Transformer-based models scale poorly with increasing context and become infeasible beyond $\sim$20k tokens.

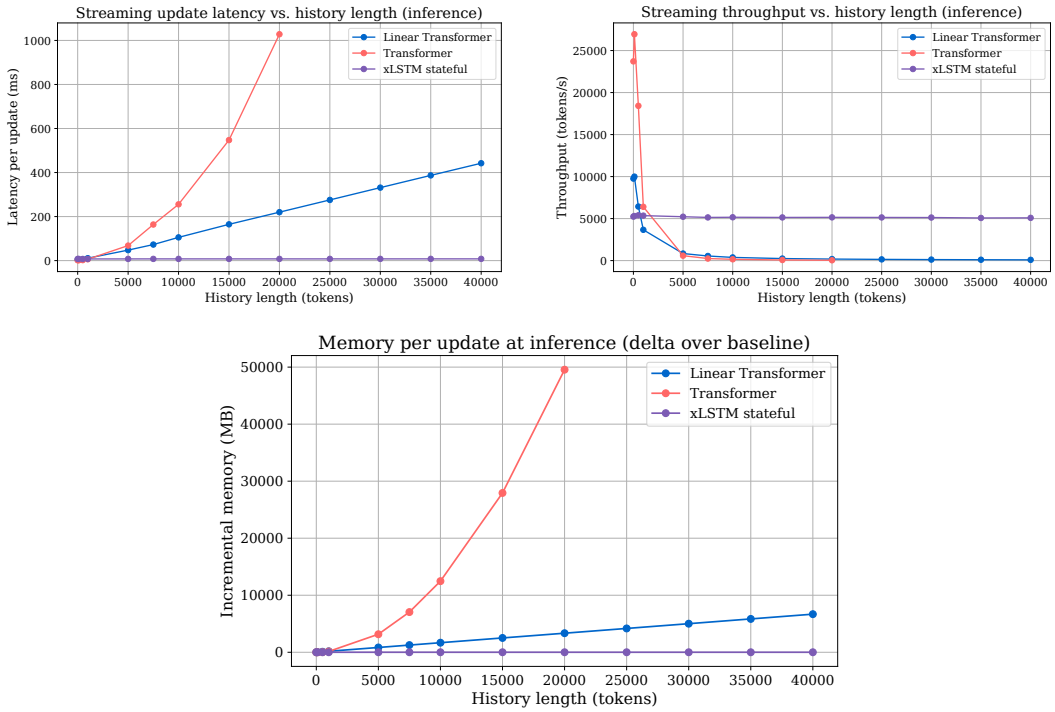


Figure 12: Streaming inference characteristics of the NTP backbone. **Top-left:** latency per update, **top-right:** throughput, **bottom:** memory usage as a function of history length. xLSTM maintains constant latency and memory, while transformer variants scale with context length.

### I.6 Implications

These results suggest that while multiple architectures are viable for self-supervised pretraining, the choice of backbone becomes critical in deployment scenarios involving long-horizon continuous monitoring. In such settings, the constant-memory, stateful inference capability of xLSTM provides a clear practical advantage over Transformer-based alternatives.

## J Ablation studies and context length sensitivity

### J.1 Ablations studies

We provide indicative ablation analyses to assess the contribution of input modalities and auxiliary training objectives.

**Input modality ablation.** We evaluate the impact of different input configurations on downstream performance. Results are summarized in Table 15.

Table 15: Impact of input modalities on downstream seizure detection performance on the KUK held-out set.

| Input modality | BACC (%) | AUPRC (%) | AUROC (%) |
|---|---|---|---|
| Non-EEG only | 57.5 | 39.6 | 78.6 |
| EEG only | 68.8 | 45.0 | 82.3 |
| EEG + non-EEG | 71.5 | 52.7 | 84.2 |
| Multimodal (+ text) | **85.3** | **79.3** | **88.4** |

Combining biosignals improves performance over single modalities, while the inclusion of clinical text yields a substantial additional gain, highlighting the value of multimodal contextual information.

**Auxiliary task ablation.** To assess the contribution of auxiliary objectives, we conduct controlled ablations using a reduced model (11M parameters) and limited pretraining (50 epochs). In each experiment, a single auxiliary head is removed while all others are retained.

Table 16: Ablation study of auxiliary pretraining objectives and their impact on downstream seizure detection performance (KUK held-out set, reduced model, limited pretraining).

| Configuration | BACC (%) | AUPRC (%) | AUROC (%) |
|---|---|---|---|
| Full | **53.1** | **44.2** | **74.3** |
| w/o token-type | 52.7 | 44.1 | 73.1 |
| w/o markers | 52.4 | 40.9 | 70.0 |
| w/o TPS | 51.9 | 37.5 | 68.7 |
| w/o TTS | 50.0 | 33.0 | 66.7 |
| w/o seg-type | 51.0 | 42.4 | 71.8 |

The full auxiliary objective consistently yields the best performance. Among individual components, removing time-to-seizure (TTS) has the largest impact, followed by time-past-seizure (TPS) and marker reconstruction. The token-type objective has a comparatively minor effect, while the segment-type head remains necessary for direct classification supervision.

We further observe interactions between auxiliary tasks (e.g., reduced TTS performance when TPS is removed), suggesting that the joint objective provides complementary supervision signals.

Overall, these results indicate that the combined auxiliary framework contributes to improved representation quality.

**Montage geometry ablation.** We compare NeuroNTP with a variant pretrained without the EEG electrode-location encoding; all other settings are identical. Both models are evaluated by linear probing on the KUK held-out set with frozen tokenizer and backbone and biosignal-only input (Table 17). Encoding electrode geometry improves AUPRC by 10.8 and AUROC by 7.5 percentage points.

Table 17: Impact of EEG electrode location encoding on seizure detection performance. Both models are pretrained identically except for the location encoding and evaluated by linear probing (KUK dataset; biosignal-only input; frozen tokenizer and backbone). Results are reported as mean ± standard deviation (in %) across five runs. $\Delta$ denotes the absolute improvement in percentage points when incorporating EEG geometry information.

| Configuration | BACC | AUPRC | AUROC |
|---|---|---|---|
| Without EEG geometry | $50.7^{\pm 0.1}$ | $33.3^{\pm 0.4}$ | $71.8^{\pm 0.4}$ |
| With EEG geometry | $\mathbf{52.9}^{\pm 0.1}$ | $\mathbf{44.1}^{\pm 0.1}$ | $\mathbf{79.3}^{\pm 0.2}$ |
| $\Delta$ (pp) | +2.2 | +10.8 | +7.5 |

**Restricted biosignal channel configurations.** To emulate restricted channel conditions, we fine-tune the multimodally pretrained NeuroNTP model using a reduced electrode set (M1, M2, T7, T8), corresponding to a behind-the-ear configuration, together with ECG and $SpO_2$ signals. Evaluation is performed on the four held-out test folds (4 × 15 patients). In addition, we consider a fully restricted variant using only the BTE electrodes, excluding ECG and $SpO_2$, to isolate the contribution of multimodal physiological signals (Table 18).

Table 18: **Comparison of seizure detection performance between full and reduced biosignal channel configurations.** Results are reported using the KUK dataset 4-fold patient-wise cross-validation on recordings from 60 unseen patients. All values are shown as mean ± standard deviation (in %).
Purple denotes best and grey within one standard deviation of the best result.

| Model | BACC | AUPRC | AUROC |
|---|---|---|---|
| **NeuroNTP** | | | |
| Behind-the-Ear (BTE) channels | $62.6^{\pm 4.5}$ | $48.6^{\pm 4.9}$ | $82.8^{\pm 1.5}$ |
| Behind-the-Ear (BTE) + ECG + $SpO_2$ | $69.0^{\pm 4.5}$ | $50.6^{\pm 3.4}$ | $83.9^{\pm 1.4}$ |
| All biosignal channels | $\mathbf{71.5}^{\pm 2.1}$ | $\mathbf{52.7}^{\pm 5.5}$ | $\mathbf{84.2}^{\pm 0.7}$ |

Even under severe channel reduction, the emulated wearable setup using only four BTE electrodes achieves comparable performance (62.6% BACC, 82.8% AUROC). Augmenting the BHE configuration with ECG and $SpO_2$ further improves performance (69.0% BACC, 83.9% AUROC), indicating that multimodal physiological signals can partially compensate for reduced spatial EEG coverage and improve robustness.

### J.2 Context length sensitivity

To evaluate the influence of temporal context on downstream seizure detection performance, we analyzed the effect of progressively increasing the number of preceding 10-second segments provided to the model during inference. Figure 13 summarizes balanced accuracy (BACC), area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) as a function of context length, evaluated on a subset of the held-out KUK cohort using a biosignal-only input configuration with the pretrained multimodal model.

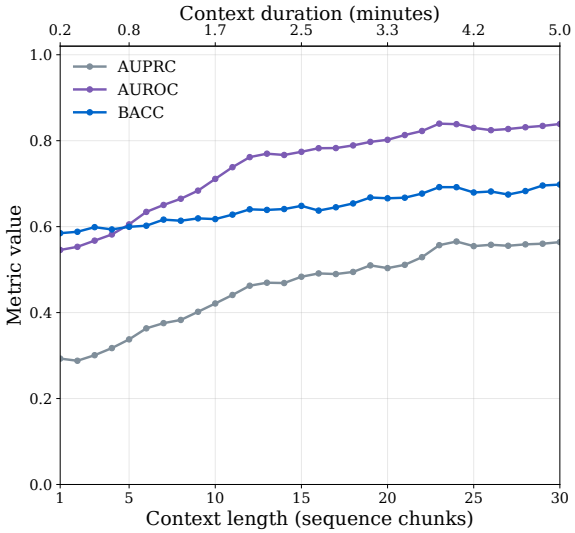


Figure 13: Downstream seizure detection performance as a function of temporal context length. The lower x-axis shows the number of preceding 10-second segments provided to the model, while the upper x-axis indicates the corresponding temporal duration in minutes. Performance improves consistently with increasing context length before approaching saturation for longer sequences.

The results show a consistent improvement in performance as the temporal context increases from short windows toward longer historical sequences. The largest gains are observed during the initial increase from approximately 10–50 seconds of context, indicating that short-term temporal evolution contains highly informative seizure-related dynamics. Performance improvements continue more gradually for longer contexts, with metrics approaching saturation around 150–200 seconds.

These findings support the architectural motivation behind NeuroNTP and its ability to model extended temporal dependencies across multimodal biosignal streams. In particular, the results suggest that seizure-related information is distributed over substantially longer temporal horizons than those typically captured by isolated short-window classification approaches. At the same time, the observed diminishing returns for very long contexts indicate that a moderate sequence length already captures most clinically relevant temporal structure for the evaluated tasks.

## K Compute environment and pretraining resources

We report the computational resources used for NeuroNTP pretraining to support reproducibility and contextualize training cost. All main pretraining runs were performed on NVIDIA A100 80GB GPUs using PyTorch with CUDA-enabled mixed-precision training. Fine-tuning experiments were additionally carried out on NVIDIA A100 40GB GPUs and on 20GB Multi-Instance GPU (MIG) partitions of A100 devices, and a subset of preliminary test runs used an NVIDIA Titan V (12GB). The reported runtimes denote completed training time only and exclude preprocessing, data conversion, exploratory debugging, and failed runs.

Table 19: Pretraining compute summary.

| **Dataset** | **Runs / stages** | **Total runtime** |
|---|---|---|
| KUK | cumulative | 22d 7h 52m (≈535.9 GPU-hours) |
| TUSZ | 2 | 2d 13h 58m (≈62.0 GPU-hours) |

Table 20: Software and hardware environment used for pretraining.

| **Component** | **Version / configuration** |
|---|---|
| Operating system | Linux 5.14.0, glibc 2.34 |
| Python | 3.11.14 |
| PyTorch | 2.4.1+cu121 |
| CUDA runtime | 12.1 |
| NVIDIA driver | 590.48.01 |
| Driver-supported CUDA | 13.1 |
| cuDNN | 9.1.0 |
| Training GPUs | NVIDIA A100 80GB PCIe |
| Available GPUs | 6 × NVIDIA A100 80GB PCIe |
| Memory per GPU | 79.3 GB visible to PyTorch |
| Lightning | 2.5.6 |
| PyTorch Lightning | 2.4.0 |
| NumPy / SciPy / Pandas | 2.3.5 / 1.16.3 / 2.3.3 |
| MNE / h5py | 1.11.0 / 3.15.1 |
| Transformers | Git revision `70179949` |

# L Hyperparameters and framework configuration

Given the large number of tunable hyperparameters and the resulting high-dimensional search space, our empirical exploration was necessarily selective and hypothesis-driven. We primarily focused on the components expected to have the strongest impact on representation quality and temporal modeling, namely the MMT signal encoder and the backbone architecture, as well as on the composition of the input modalities. In contrast, the loss formulation and regularization strategies were explored in a more limited manner, aiming at stability and consistency rather than exhaustive optimization. We report the configuration used for pretraining on the KUK dataset and, where applicable, indicate the ranges of values explored during ablation. The tables that follow summarize the most relevant hyperparameters, while omitting secondary or implementation-specific settings for clarity and conciseness.

Our hyperparameter exploration leveraged the modular design of the NeuroNTP framework. We first investigated individual subsystems in isolation, in particular the biosignal encoder and the MMT module, using simplified supervised objectives to identify stable and performant parameter regimes. Once viable ranges were established, we proceeded to end-to-end optimization of the full architecture under the self-supervised objective, using short training runs to efficiently screen configurations before committing to longer pretraining schedules. Validation loss was used as the selection criterion. This staged strategy allowed us to balance empirical rigor with computational feasibility.

In the final configuration, the biosignal encoder employs an STFT with an FFT size of 256 and a hop length of 64 samples. The shared convolutional stem consists of three residual blocks with progressively increasing channel width and temporal downsampling, yielding a compact yet expressive representation that preserves clinically relevant temporal structure while remaining computationally efficient.

For training, we adopt a weighted multi-objective loss with coefficients $\alpha = 0.5, \varepsilon = 0.15, \beta = \gamma = \delta = \zeta = 0.1$, which assign primary importance to the next-token prediction objective while ensuring that the auxiliary tasks contribute as structured regularizers. This design encourages the model to learn a coherent shared representation that supports seizure-related reasoning without allowing any single auxiliary objective to dominate the optimization process.

For the quantile-based regression components, we define the set of quantiles as $\mathcal{Q} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$, with corresponding weights $\boldsymbol{\omega}_\tau = \{0.1, 0.2, 0.4, 0.2, 0.1\}$. This configuration was chosen to place the strongest emphasis on the median and the interquartile range while still accounting for the distribution tails. However, as these quantiles and weights are symmetric around the median and one prediction is shared across quantiles, the loss reduces to $\frac{1}{2}|y_j - \hat{y}_j|$ per dimension, i.e., a scaled L1 loss (Section 4.3). The quantile formulation was retained because it supports asymmetric weightings, which we explored during development.

Table 21: **Architecture and causal model configuration.**

| Hyperparameter | Value | Explored range | Description |
|---|---|---|---|
| `embedding_dim` ($d_{emb}$) | 1024 | 512-1024 | Multimodal embedding dimension. |
| `marker_emb_dim` | 768 | – | Marker embedding dimension. |
| `seq_len` | 30 | – | Context length used in experiments. |
| `chunk_size` | 10 s | – | Size of the chunk used in experiments. |
| `overlap` | 0 s | – | Consecutive chunk overlap. |
| `sampling_rate` | 256 Hz | – | Sampling rate used in pretraining and experiments. |
| `max_seq_len` | 5000 | – | Maximum multimodal sequence length. |
| `ntp_type` | xLSTM | lin.transformer | Causal next-token predictor backbone. |
| `xlstm.num_blocks` | 8 | 8-24 | Number of xLSTM blocks. |
| `xlstm.context_length` ($L$) | 1536 | – | Context length of causal model. |
| `xlstm.slstm.num_heads` ($H$) | 16 | 8-24 | Number of sLSTM heads. |
| `xlstm.slstm.conv1d_kernel_size` ($k$) | 4 | – | Convolutional kernel size in sLSTM. |
| `xlstm.feedforward.proj_factor` | 1.3 | – | Feedforward expansion factor. |

Table 22: **Modalities and tokenization configuration.**

| Hyperparameter | Value | Explored range | Description |
|---|---|---|---|
| `add_eeg` | true | – | Include EEG channels. |
| `add_non_eeg` | true | {false, true} | Include non-EEG biosignals. |
| `add_eeg_location` | true | {false, true}* | Encode EEG electrode position. |
| `add_tps`, `add_tts` | true | {false, true} | Include TPS and TTS tokens. |
| `add_markers`, `add_notes` | true | {false, true}* | Include clinical markers and notes. |
| `add_gender`, `add_age` | true | {false, true}* | Include demographic metadata. |
| `add_hod` | true | {false, true}* | Hour-of-day token. |

** Tested in ablation experiments. For boolean parameters, {false, true} indicates that both states were evaluated.*

Table 23: **Signal encoder configuration.**

| Hyperparameter | Value | Explored range | Description |
|---|---|---|---|
| `sigenc_fft` ($N_{\text{FFT}}$) | 256 | – | FFT size of STFT. |
| `sigenc_hoplen` ($h$) | 64 | – | Hop length. |
| `sigenc_stem_layers` | 3 | 2-4 | Convolutional stem depth. |
| `sigenc_stem_init_ch` | 16 | 12-20 | Initial stem width. |
| `sigenc_stem_out` | 32 | 16-32 | Stem output channels. |
| `sigenc_cnn_stride` | 2 | 2-4 | Channel-head stride. |
| `sigenc_bottleneck` | 16 | 8-16 | Channel-head bottleneck dimension. |
| `sigenc_lora_rank` | 4 | – | LoRA rank. |

Table 24: **Loss weights, sampling and regularization parameters.**

| Hyperparameter | Value | Explored range | Description |
|---|---|---|---|
| `event_frac` | 0.3 | 0.1-1.0 | Fraction of event samples per epoch. |
| `nonevent_frac` | 0.03 | 0.01-1.0 | Fraction of non-event samples per epoch. |
| `signals_channel_drop_prob` ($p$) | 0.3 | 0.0-0.6 | Channel dropout probability. |
| `tps_mask_prob` | 0.7 | 0.3-1.0 | Mask probability for TPS token (initial). |
| `markers_mask_prob` | 0.7 | 0.3-1.0 | Mask probability for marker tokens. |

Table 25: **Pretraining loss weights and objective parameters.**

| Hyperparameter | Value | Explored range | Description |
|---|---|---|---|
| `pt_loss_alfa` ($\alpha$) | 0.5 | 0.4-0.6 | Next-token prediction loss weight. |
| `pt_loss_beta` ($\beta$) | 0.1 | 0.05-0.1 | Token-type loss weight. |
| `pt_loss_gamma` ($\gamma$) | 0.05 | 0.05-0.1 | TPS loss weight. |
| `pt_loss_delta` ($\delta$) | 0.1 | 0.05-0.1 | TTS loss weight. |
| `pt_loss_epsilon` ($\epsilon$) | 0.15 | 0.1-0.2 | Segment-type loss weight. |
| `pt_loss_zeta` ($\zeta$) | 0.1 | 0.05-0.1 | Marker reconstruction loss weight. |

Table 26: **Training and optimization configuration.**

| Hyperparameter | Value | Explored range | Description |
|---|---|---|---|
| `mode` | pretrain | – | Training regime. |
| `optimizer` | AdamW | – | The optimizer used for training. |
| `epochs` ($E$) | 100 | 50-150 | Number of training epochs. |
| `batch_size` ($B$) | 8 | 4-12 | Global batch size. |
| `lr` ($\eta$) | $2 \times 10^{-4}$ | $10^{-5}$–$10^{-4}$ | Base learning rate. |
| `weight_decay` ($\lambda$) | $5 \times 10^{-5}$ | – | AdamW weight decay. |
| `lr_scheduler` | cos_warmup | w/o scheduler | Scheduler with cosine annealing and warm-up. |
| `lr_anneal` | cos | – | Cosine annealing policy. |
| `lr_warmup` | 0.05 | – | Warm-up fraction of total steps. |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims stated in the abstract and introduction are directly supported by the experimental results in Sections 4–5, including consistent improvements across multiple datasets (KUK, CHB-MIT, Siena) and analyses in the appendix that clarify scope and limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly discussed in the Discussion section, including dataset privacy constraints, potential bias due to cohort characteristics, limitations in temporal invariance, and challenges related to interpretability and deployment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results requiring proofs; the contribution is empirical and methodological.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experimental protocols, preprocessing steps, model configurations, and evaluation procedures are described in detail in the main paper and appendix, including exact dataset splits, preprocessing pipelines, and training procedures to enable reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to patient privacy regulations, the primary KUK dataset cannot be publicly released. However, we provide full details of the methodology and validate results on public datasets (TUH/TUSZ, CHB-MIT, Siena), and we release code and pretrained model weights to facilitate reproducibility and further research.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies dataset splits, preprocessing pipelines, model architecture, training procedures, and hyperparameters in the main text and appendix, with additional implementation details provided for all evaluation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All reported results include mean and standard deviation over multiple runs, with variability explicitly stated in tables and described in the text.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information on compute resources, including GPU type, software environment, and total training time for pretraining and evaluation in the appendix, along with cumulative compute estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics, including proper handling of sensitive clinical data, anonymization, and compliance with institutional ethical approval procedures.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification:

Justification: Broader societal impacts are discussed in the Discussion section, including potential benefits for seizure management and quality of life, as well as risks related to privacy, dataset bias, generalization, interpretability, and safe clinical deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not involve releasing high-risk generative models or sensitive datasets; only pretrained weights are planned for release, without raw clinical data.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and prior methods used (e.g., TUH/TUSZ, CHB-MIT, Siena, BIOT preprocessing) are properly cited, and publicly available resources are referenced with their original sources.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide documentation for the NeuroNTP architecture, training pipeline, and evaluation procedures in the paper and appendix, and plan to release pretrained model weights and code to support reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The study does not involve direct interaction with human participants or crowdsourcing; all data are retrospective clinical recordings.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The use of clinical data was approved by the responsible ethics committee (EK 1021/2023), as stated in the Ethics section.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models are not used as part of the core methodology; they are only used as frozen embedding models for textual markers and do not constitute a novel or generative component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.