GPT2MEG: QUANTIZING MEG FOR AUTOREGRESSIVE GENERATION

Richard Csaky *
Barcelona Computational Foundation, Barcelona, Spain

Mats W.J. van Es
Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, Oxford, UK
Wellcome Centre for Integrative Neuroimaging, Oxford, UK

Oiwi Parker Jones
Wellcome Centre for Integrative Neuroimaging, Oxford, UK
Department of Engineering Science, University of Oxford, Oxford, UK

Mark Woolrich
Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, Oxford, UK
Wellcome Centre for Integrative Neuroimaging, Oxford, UK

ABSTRACT

Large models and language-model training recipes are increasingly repurposed for time series, yet most work emphasizes univariate forecasting and evaluates models primarily via next-step loss. We introduce GPT2MEG, a tokenized GPT-2-style Transformer for multichannel MEG that enables context-informed autoregressive generation via additive embeddings for sensor identity, subject ID, and time-aligned task conditions. We treat channels as a batch dimension. Using simple μ-law companding with uniform quantization, we train with cross-entropy and sample long horizons. To support rigorous evaluation of generative time-series models, we complement next-step metrics with spectral fidelity, HMM-based multivariate dynamics, and task-evoked response alignment. GPT2MEG best matches HMM state statistics and conditioned evoked responses, scales across 15 subjects via subject embeddings, and yields interpretable channel embeddings aligned with sensor geometry. Code available at https://github.com/ricsinaruto/MEG-transfer-decoding.

Track: Research

1 INTRODUCTION

Magnetoencephalography (MEG) and related electrophysiology (EEG/ECoG) provide millisecond-resolved measurements of large-scale brain dynamics, but remain challenging to model due to high dimensionality, strong temporal structure, and substantial variability across subjects and recording setups. Recent self-supervised *foundation models* have begun to learn transferable representations from large collections of unlabelled neural recordings, improving performance and data efficiency on downstream decoding/encoding tasks (Kostas et al., 2021; Wang et al., 2023; Cui et al., 2023; Jiang et al., 2024; Wang et al., 2024; El Ouahidi et al., 2025). In contrast, autoregressive generative models, capable of simulating realistic neural time series, remain comparatively underexplored for MEG, despite their potential for data augmentation, uncertainty-aware decoding, and simulation-based hypothesis testing.

Inspired by tokenized time-series language models (Ansari et al., 2024; Das et al., 2024; Rasul et al., 2023), we ask: *can we repurpose autoregressive language-model training to build a context-informed*

[*] Correspondence to richard.csaky@gmail.com.

generative model for multichannel MEG? We propose GPT2MEG, a GPT-2–style decoder-only Transformer trained by next-token prediction on discretized MEG. A simple tokenization (μ-law companding + uniform bins) turns continuous MEG into symbols, enabling cross-entropy training and autoregressive sampling. As additional context, we incorporate additive embeddings for sensor identity, subject ID, and time-aligned task/stimulus labels, enabling conditional generation of task-evoked activity and multi-subject scaling.

Contributions.

- **Context-informed generative model:** GPT2MEG repurposes a GPT-2 architecture for multichannel MEG via discrete tokenization and additive embeddings for channel, subject, and task context.
- **Rigorous evaluation beyond next-step loss:** we show that one-step accuracy is weakly predictive of long-horizon fidelity and evaluate generations with spectral, HMM-dynamical, and task-evoked metrics.
- **Interpretability and scaling:** we demonstrate multi-subject training via subject embeddings and report interpretable embeddings that reflect sensor geometry and conditioning semantics.

2 Related Work

Self-supervised learning has enabled representation learning from unlabelled EEG/MEG using contrastive, masked, and predictive objectives (Banville et al., 2021; Kostas et al., 2021; Wang et al., 2023), with recent work scaling towards brain foundation models (Cui et al., 2023; Jiang et al., 2024; Wang et al., 2024; El Ouahidi et al., 2025). These approaches primarily target embeddings for decoding/encoding; our focus is complementary: autoregressive *generation* of multichannel MEG with explicit context conditioning and evaluation of long-horizon dynamics.

Tokenization enables language-model training for continuous time series (Ansari et al., 2024; Das et al., 2024; Rasul et al., 2023). However, next-step likelihood alone can miss long-horizon or context-locked structure. We therefore emphasize evaluation of free-running generations using spectral fidelity, multivariate dynamical summaries, and task-evoked alignment, and we analyze learned embeddings for interpretability.

3 Methods

3.1 Problem setup and notation

We consider multichannel sensor-level MEG recordings $\mathbf{X} \in \mathbb{R}^{C \times T}$ with C sensors and T time points. For task datasets we also have a time-aligned condition sequence $\mathbf{y}_{1:T}$ (dynamic exogenous context; e.g. stimulus identity), and for multi-subject training a subject index s (static context). Our goal is to learn a causal conditional generative model

$$p(\mathbf{X} \mid \mathbf{y}, s) = \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}, \mathbf{y}_{\leq t}, s), \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^C$ is the multichannel sample at time t. All deep models are trained by teacher-forced next-token prediction with cross-entropy loss.

3.2 Tokenization and baselines

Following WaveNet (van den Oord et al., 2016), we discretize each channel independently into Q=256 tokens using μ-law companding and uniform quantization. This yields a categorical prediction target and supports sampling without collapsing to mean predictions. Reconstruction and downstream decoding are largely preserved after de-quantization (Appendix A.3).

We benchmark GPT2MEG against a per-channel linear AR(255) baseline and multichannel WaveNet variants (WFC, WFCM). WaveNet architecture details and conditioning equations are provided in Appendix A.1.

3.3 Context-informed GPT2MEG

We adapt GPT-2 to tokenized MEG by treating each channel as a batch dimension (no channel mixing) and adding context embeddings for channel, subject, and task condition. Let $\mathbf{X} \in \mathbb{R}^{C \times T}$ be the tokenized sequence (integer tokens), embedded by a shared token embedding matrix $\mathbf{W}_e \in \mathbb{R}^{Q \times E}$. The initial hidden states are

$$\mathbf{H}^{(0)} = \mathbf{X}\mathbf{W}_e + \mathbf{W}_p + \mathbf{Y}\mathbf{W}_y + \mathbf{O}\mathbf{W}_o + \mathbf{W}_c, \tag{2}$$

where $\mathbf{W}_p$ are positional embeddings, $\mathbf{W}_c \in \mathbb{R}^{C \times E}$ are learned channel embeddings, $\mathbf{W}_o \in \mathbb{R}^{O \times E}$ are subject embeddings, and $\mathbf{W}_y \in \mathbb{R}^{Y \times E}$ are task-condition embeddings (set to zero when no stimulus is present). As in language modeling, the Transformer predicts the next token distribution at each timestep.

3.4 Generation and evaluation

We evaluate models in both forecasting and free-running generation settings. Tokenized models generate by sampling tokens autoregressively (top-p sampling (Holtzman et al., 2020)); AR models generate by recursively filtering noise. Since one-step prediction accuracy can be weakly informative for long-horizon behaviour, we emphasize generation-based metrics:

- **Spectral fidelity:** power spectral density (PSD) comparisons.
- **Multivariate dynamics:** summary statistics of a 12-state time-domain embedding HMM fit to generated vs. real multichannel time series (Vidaurre et al., 2018b). See Appendix A.2 for implementation details
- **Task-evoked structure:** trial-averaged evoked responses under task conditioning.
- **Downstream utility:** decoding performance when training classifiers on simulated trials.

4 Results

Experimental setup. We use the 15-subject continuous MEG dataset of Cichy et al. (2016) (118 visual images, 30 trials/image). We bandpass filter 1–50 Hz, apply a notch filter for line noise, perform ICA artifact rejection (64 components with manual component rejection), and downsample to 100 Hz. We split each continuous recording (subject) into non-overlapping blocks corresponding to trials, holding out 4 trials/condition for validation and 4 for test. Each sensor is standardized (0 mean, unit variance), clipped $(-10, 10)$, rescaled to $(-1, 1)$, and quantized into Q=256 bins with μ-law companding. This discretization incurs negligible loss for evoked and decoding analyses.

Unless otherwise stated, main-text generative analyses are for a representative subject (single-subject training). We additionally evaluate multi-subject scaling with GPT2MEG-group (15 subjects; Appendix A.6). Hyperparameters and model variants are in Appendix A.4.

Next-step accuracy does not predict long-horizon quality. Across models, next-token forecasting accuracy is only moderately above a repeat-last-value baseline and provides limited separation between architectures (Appendix A.5). This motivates the question of how to evaluate generative time-series foundation models beyond next-step likelihood.

GPT2MEG better matches multivariate dynamics. A good generative model should be expected to be able to recursively generate data that looks like the real data. Here, we first assess the models' ability to do this using the power spectra. For deep learning models we used top-p sampling with $p = 80\%$ (unless otherwise noted in the figure caption) to recursively generate data. We generated 3600 seconds with all models. For models that have task-conditioning (all except AR(255)) we use the task label timeseries from the training set. The models in this section were trained on a single sample subject, containing about 1.5 hours of data downsampled to 100 Hz across 306 channels.

Generated token sequences are first de-tokenised and then the power spectral density (PSD) is computed on the continuous data. Figure 2 compares the PSD of the generated data across our models. Qualitatively, it is clear that AR(255) reproduces PSDs that match best with those computed directly

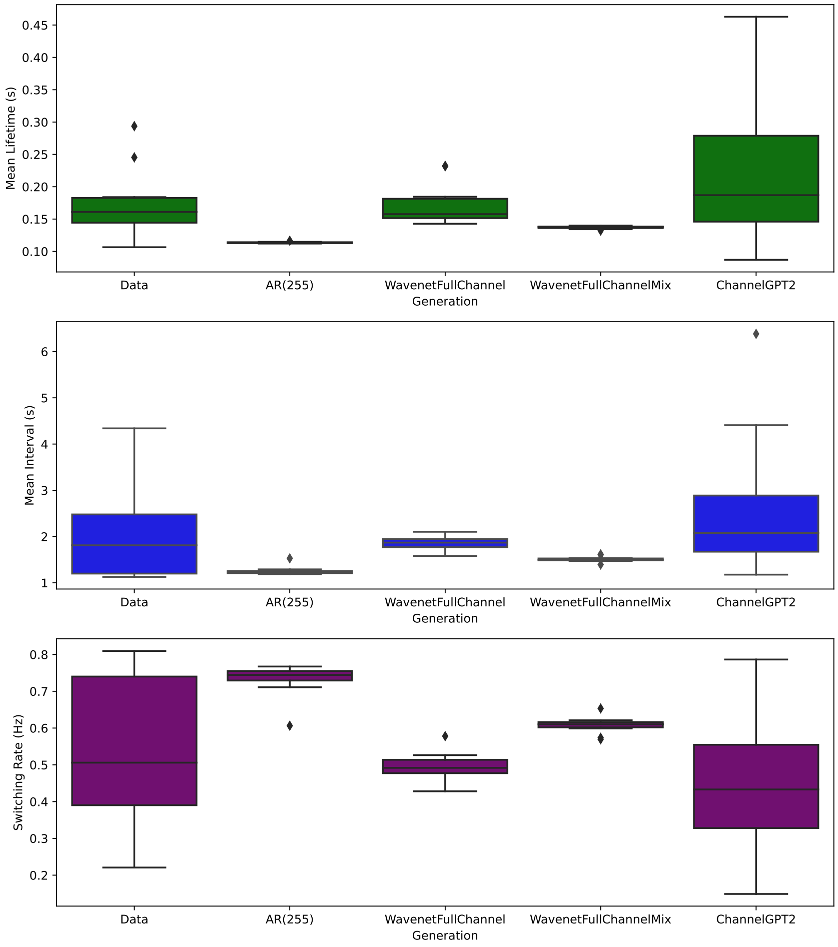


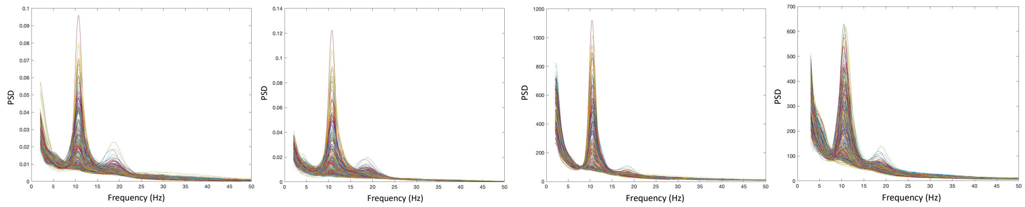
Figure 1: **Distributions of dynamics summary statistics across the 12 states from an HMM** inferred on real MEG multi-channel data (left) and on data generated from different models. `ChannelGPT` refers to `GPT2MEG`.

on the MEG data, while `WavenetFullChannel` and `GPT2MEG` are not far behind. All models capture the characteristic $1/f$ shape, and peaks at 10 and 19 Hz, likely related to alpha and beta band activity.

We evaluate long-horizon generations by fitting a 12-state HMM separately to real and generated multichannel time series and comparing distributions of state-dynamics summary statistics (Vidaurre et al., 2018a). Figure 1 shows that `GPT2MEG` best matches the real-data distributions across switching rate, mean state interval, and other dynamical summaries, whereas AR(255) and `WFCM` generate overly homogeneous state statistics.

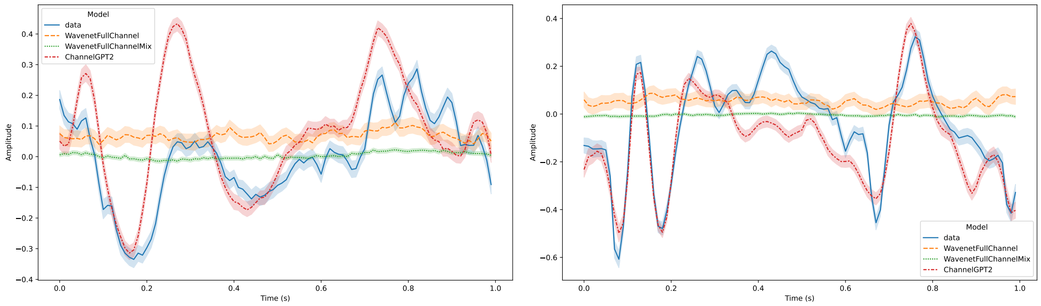
Context conditioning reproduces task-evoked responses. A central goal of context-informed time-series models is to incorporate exogenous inputs. Using time-aligned stimulus labels as conditioning, we epoch generated time series and compute trial-averaged evoked responses. Figure 3 shows that `GPT2MEG` reproduces the timing and amplitude of evoked responses in representative frontal and visual sensors, while WaveNet variants fail to capture task-locked structure.

Interpretability, scaling, and failure modes. Training a single `GPT2MEG-group` model across 15 subjects using subject embeddings preserves task-evoked structure, but tends to smooth subject-specific responses (Appendix A.6). Channel embeddings learned by `GPT2MEG-group` reflect sensor geometry (pairwise distance correlation 0.45; Appendix A.12). Shuffling or collapsing condition labels degrades evoked-response alignment (Appendix Figure 7), indicating the model uses



(a) Data (b) AR(255) (c) WFC (d) GPT2MEG

Figure 2: **PSD comparison** between real MEG (a) and long-horizon generations from different models (b–d) for a representative subject. Each line is a sensor/channel.



(a) Frontal channel (MEG0111) (b) Visual channel (MEG2332)

Figure 3: **Evoked responses for real MEG and task-conditioned generations.** Stimulus onset is at 0 s and offset at 0.5 s. Shading indicates 95% confidence interval of the trial mean. ChannelGPT refers to GPT2MEG.

semantic label information rather than timing alone. Despite these successes, the channel-independent architecture underestimates cross-channel covariance (Appendix A.10), motivating future work on explicit multivariate mixing. Downstream decoding and transfer-learning experiments are reported in Appendix A.7–A.8. GPT2MEG is able to generate evoked responses that are classifiable across 118 classes (image conditions), and moderately improve decoding of real trials.

5 Discussion

GPT2MEG demonstrates that simple tokenization and additive context embeddings can repurpose language-model architectures for high-rate multichannel neural time series, enabling conditional generation and multi-subject scaling. Our results reinforce a key message: next-step loss alone can miss long-horizon and context-locked structure, motivating evaluation protocols that probe dynamics, spectra, and conditional responses. Key directions include explicit cross-channel modeling, richer context (e.g. conditioning on image/text embeddings instead of categorical labels), as well as scaling to multiple large datasets.

Acknowledgements

This research was supported by the NIHR Oxford Health Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. RC is supported by a Wellcome Centre Integrative Neuroimaging Studentship. MVE's research is supported by the Wellcome Trust (215573/Z/19/Z). OPJ is supported by the UK MRC (MR/X00757X/1). MWW's research is supported by the Wellcome Trust (106183/Z/14/Z, 215573/Z/19/Z), the New Therapeutics in Alzheimer's Diseases (NTAD) study supported by UK MRC and the Dementia Platform UK (RG94383/RG89702) and the EU-project euSNN (MSCA-ITN H2020-860563). The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z).

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021. doi: 10.1088/1741-2552/abca18. URL https://doi.org/10.1088/1741-2552/abca18.

Charles W. Brokish and Michele Lewis. *A-Law and mu-Law Companding Implementations Using the TMS320C54x*. Texas Instruments, dec 1997. URL https://www.ti.com/lit/an/spra163a/spra163a.pdf.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, 2016. doi: 10.1038/srep27755. URL https://doi.org/10.1038/srep27755.

Richard Csaky, Mats W.J. van Es, Oiwi Parker Jones, and Mark Woolrich. Interpretable many-class decoding for meg. *NeuroImage*, 282:120396, 2023. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2023.120396. URL https://www.sciencedirect.com/science/article/pii/S1053811923005475.

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. *arXiv preprint arXiv:2311.03764*, 2023. URL https://arxiv.org/abs/2311.03764.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.10688.

Ines El Ouahidi, Razvan-Gabriel Mihai, and Pascal Frossard. Reve: Randomized eeg-to-text encoder for enhanced generalization. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=J0GIpEsV8X.

Chetan Gohil, Rukuang Huang, Evan Roberts, Mats WJ van Es, Andrew J Quinn, Diego Vidaurre, and Mark W Woolrich. osl-dynamics: A toolbox for modelling fast dynamic brain activity. *bioRxiv*, pp. 2023–08, 2023. doi: 10.1101/2023.08.07.549346. URL https://doi.org/10.1101/2023.08.07.549346.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Weibang Jiang, Liming Zhao, and Bao-Liang Lu. Labram: Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, pp. 253, 2021. doi: 10.3389/fnhum.2021.653659. URL https://doi.org/10.3389/fnhum.2021.653659.

Kashif Rasul, Arjun Ashok, Alex Williams, Florian Adam, Nikhil Hassen, Ehsan Khorasani, Patrick Gendron, Tao Jiang, Leonard Berrada, Xiang Li, Jacob Menick, Jonas W. Müller, Valentina Zantedeschi, Yi Wang, Syama Sundar Rangapuram, Sebastian Pineda Arango, Abheesht Gupta, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting, 2023. URL https://arxiv.org/abs/2310.08278.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pp. 125, 2016. URL `https://www.isca-archive.org/ssw_2016/vandenoord16_ssw.html`.

Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Andrew J Quinn, Fidel Alfaro-Almagro, Stephen M Smith, and Mark W Woolrich. Discovering dynamic brain networks from big data in rest and task. *NeuroImage*, 180:646–656, 2018a. doi: 10.1016/j.neuroimage.2017.06.077. URL `https://doi.org/10.1016/j.neuroimage.2017.06.077`.

Diego Vidaurre, Laurence T Hunt, Andrew J Quinn, Benjamin AE Hunt, Matthew J Brookes, Anna C Nobre, and Mark W Woolrich. Spontaneous cortical activity transiently organises into frequency specific phase-coupling networks. *Nature Communications*, 9(1):1–13, 2018b. doi: 10.1038/s41467-018-05316-z. URL `https://doi.org/10.1038/s41467-018-05316-z`.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023. URL `https://arxiv.org/abs/2302.14367`.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *Advances in Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=lvS2b8CjG5`.

A Appendix

A.1 Multi-channel Wavenet

Here we describe how we adapted the Wavenet architecture (van den Oord et al., 2016) for electrophysiological data. Wavenet models the conditional probability of each time sample given all preceding samples autoregressively:

$$p(\mathbf{X}) = \prod_{t=1}^{T} p(\mathbf{x}_t | \mathbf{x}_1, ..., \mathbf{x}_{t-1}) \tag{3}$$

where $\mathbf{x}_t$ is the sample at time t and T is the total sequence length. Throughout this paper we use tokenisation and quantisation interchangeably. Both have the aim of discretising a continuous quantity into a finite set of distinct bins/levels/tokens.

In the original paper, the audio waveform is tokenised using a quantisation to 8 bits following a μ-law companding transform (Brokish & Lewis, 1997):

$$f(\mathbf{x}_t) = \text{sign}(\mathbf{x}_t) \frac{\ln(1 + \mu |\mathbf{x}_t|)}{\ln(1 + \mu)} \tag{4}$$

where μ controls the number of quantisation levels, set to 255 as in the original Wavenet. $f(.)$ is applied to each value of $\mathbf{x}_t$ independently. This nonlinear transformation improves reconstruction versus uniform quantisation of the raw input, as it skews the distribution such that more levels are allocated to smaller magnitudes. For MEG data, we observe similar benefits when applying this transform prior to quantisation. Note that the input must be scaled to $(-1, 1)$ first, and clipping outliers above some threshold helps ensure a more uniform mapping.

When adapting Wavenet to M/EEG, a key challenge is the multi-channel nature of the data. We devise two versions: `WavenetFullChannel` as univariate, and `WavenetFullChannelMix` as multivariate. In both, each channel is transformed and tokenised independently to form the input to the models.

In WavenetFullChannel, we first apply an embedding layer to the tokenised data, learned separately per channel. To be clear in this univariate approach the same model is applied to each channel. However, a different embedding layer is learned for each channel, meaning that for example the quantised value of 0.42 in channel x will have a different vector representation than in channel y. This helps the model differentiate between channels.

The embedding operation is given below:

$$\forall c \in 1, 2, \ldots, C : \mathbf{X}_e^{(c)} = \mathbf{W}^{(c)}\mathbf{X}^{(c)} \tag{5}$$

$$\mathbf{H}_0 = \text{Concatenate}(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(2)}, \ldots, \mathbf{X}_e^{(C)}) \tag{6}$$

Here, $\mathbf{X}^{(c)} \in \mathbb{R}^{Q \times T}$ is the tokenised one-hot input and $\mathbf{W}^{(c)} \in \mathbb{R}^{E \times Q}$ is the embedding layer of channel c mapping tokens Q to embeddings of size E. Concatenate concatenates along the channel dimension.

$\mathbf{H}_0 \in \mathbb{R}^{C \times E \times T}$ is the resulting input to Wavenet with C as the batch dimension. Thus, the same model is applied independently to each channel in parallel. At output, a distribution is predicted simultaneously for each channel at $T + 1$. The model is optimised to accurately predict all channels.

WavenetFullChannelMix includes an extra linear layer after summing the skip representations to mix information across the channel dimension:

$$\mathbf{S} = \sum_{l=1}^{L} \mathbf{S}^{(l)} \tag{7}$$

$$\mathbf{S} = \mathbf{S}.\text{permute}(1, 2, 0) \tag{8}$$

$$\mathbf{S}_{out} = \mathbf{S}\mathbf{W}_m \tag{9}$$

where $\mathbf{W}_m \in \mathbb{R}^{C \times C}$ is the mixing weight matrix, and $\mathbf{S}^{(l)}$ is the output of the skip connection at layer l. The permutation is needed to apply the projection to the appropriate channel dimension. After this $\mathbf{S}_{out}$ is permuted back to the original dimension order and the rest proceeds identically to WavenetFullChannel.

In the original Wavenet, audio generation can be conditioned on additional inputs through embedding-based global conditioning or time-aligned local conditioning. For some experiments, we augment the model with local features of task stimuli or subject labels, first embedded into continuous vectors:

$$\mathbf{H}_y = \mathbf{Y}\mathbf{W}_y \tag{10}$$

$$\mathbf{H}_o = \mathbf{O}\mathbf{W}_o \tag{11}$$

$$\mathbf{H}_c = \text{Concatenate}(\mathbf{H}_y, \mathbf{H}_o) \tag{12}$$

where $\mathbf{Y} \in \mathbb{R}^{T \times N}$ contains the condition index $n \in (1, \ldots, N)$ at each time point, and $\mathbf{O} \in \mathbb{R}^{T \times S}$ contains the subject index $s \in (1, \ldots, S)$ at each time point $t \in (1, \ldots, T)$. $\mathbf{W}_y \in \mathbb{R}^{N \times E_n}$ and $\mathbf{W}_o \in \mathbb{R}^{S \times E_s}$ are embedding matrices mapping the labels to learned continuous vectors of size E_n and E_s, respectively. The subject index is the same across time points of the recording from the same subject. The condition index is set to the (visual) stimuli presented (e.g., one of the 118 images in Cichy et al. (2016)), for exactly those time points when the stimulus is on. At any other time, the task condition embedding $\mathbf{H}_y$ is set to 0.

$\mathbf{H}_c$ is the conditioning vector fed into Wavenet at each layer:

$$\mathbf{Z}^{(l)} = \tanh\left(\mathbf{W}_f^{(l)} * \mathbf{H}^{(l)} + \mathbf{W}_c^{(l)} * \mathbf{H}_c\right) \odot \sigma\left(\mathbf{W}_g^{(l)} * \mathbf{H}^{(l)} + \mathbf{W}_c^{(l)} * \mathbf{H}_c\right) \tag{13}$$

where $\mathbf{W}_c^{(l)}$ (1x1 convolution) projects $\mathbf{H}_c$ before adding it to the input representation ($\mathbf{H}^{(l)}$). $\mathbf{W}_f^{(l)}$ is the filter convolution weight, $\mathbf{W}_g^{(l)}$ is the gate convolution weight, and $\mathbf{Z}^{(l)}$ is the output representation at layer l. $\odot$ is element-wise multiplication. This formulation conditions the prediction on both past brain activity and stimuli:

$$p(\mathbf{X}|\mathbf{Y}, \mathbf{O}) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_1, ..., \mathbf{x}_{t-1}, \mathbf{y}_1, ..., \mathbf{y}_{t-1}, \mathbf{o}_1, ..., \mathbf{o}_{t-1}) \tag{14}$$

In single-subject models we only use the task labels $\mathbf{Y}$.

A.2 Evaluation details

We fit 12-state time-domain embedding HMMs using osl-dynamics (Gohil et al., 2023) with 15 embeddings and PCA to 80 dimensions (sequence length 2000). Decoding uses the 4-layer linear network of Csaky et al. (2023).

A.3 Tokenization quality

We verified that de-quantized signals preserve evoked responses and decoding performance compared to raw continuous data; reconstruction error is low (less than 5%) for both μ-law and linear quantization.

A.4 Hyperparameters

We match receptive fields across deep models (255 samples). WaveNet variants use two dilation stacks (7 layers each) with early stopping. We set dilation and residual channels to 128, and skip channels to 512. `GPT2MEG` uses 12 layers and 12 attention heads with embedding size 96 for single-subject models and 240 for group models; optimization uses Adam (Kingma & Ba, 2015) with early stopping. Batch size is set to the number of channels, so 1 full example for channel-independent models.

A.5 One-step prediction is weakly informative

First, we assessed the models' forecasting performance, i.e. the prediction accuracy of the label at the next time point. We used two different modified versions of Wavenet (`WavenetFullChannel` and `WavenetFullChannelMix`) alongside `GPT2MEG`. For comparison, we also evaluated the performance of a linear autoregressive (AR) model of order 255. For AR models we simply binned the predicted continuous output to compute accuracy and compare with other models.

The results on a sample subject is shown in Figure 4. Beyond standard accuracy, we also evaluated top-5 accuracy, counting a prediction as correct if the true bin was within the 5 most probable bins. Surprisingly, all models performed only moderately better than a naive baseline of repeating the previous timestep's value. However, as we shall investigate later, this does not necessarily reflect the richness of the structure in data recursively generated by the models.

As expected, the linear AR model had lower MSE but worse accuracy than the nonlinear models. This can be because MSE measures the distance of the prediction to the target, while accuracy is only 1 if the prediction is in the target bin. Thus, it may be that the AR model always predicts values that are slightly closer to the target, but never quite falling in the target bin. While `WavenetFullChannel` appears to be worse, `WavenetFullChannelMix` and `GPT2MEG` have nearly identical performance. Based on these results it is inconclusive whether deep learning models improve over the linear AR model. Further, long-range structure analyses are presented in the next sections to elucidate this.

A.6 Scaling to multiple subjects with subject embeddings

We next looked at whether combining data from multiple subjects improves modelling and generation performance. This is in line with the overall goal of training such foundational forecasting models on large datasets containing multiple subjects. Here we took a first step in exploring this by scaling `GPT2MEG` to the 15 subjects in the Cichy et al. (2016) data, which we refer to as `GPT2MEG-group`.

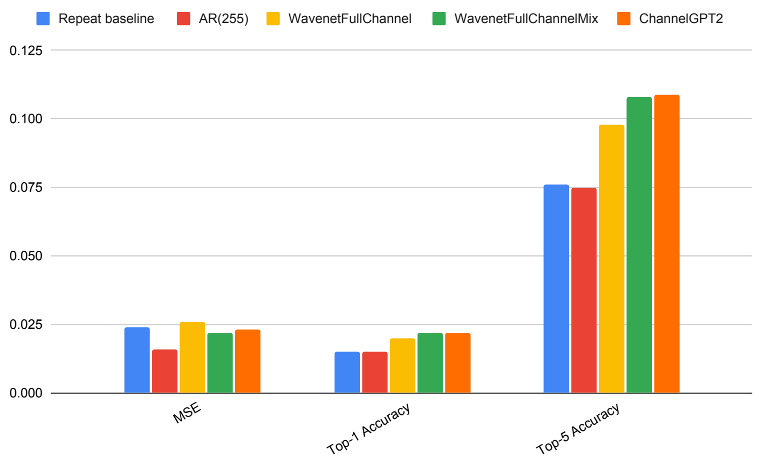


Figure 4: **Next-timestep prediction performance across the different forecasting models.** Accuracy values are on left-out test data and are given in 0-1 units. Chance-level is $1/256$, however predicting the majority class (quantised value) is somewhat higher, since the distribution over bins is not uniform. `ChannelGPT` refers to `GPT2MEG`.

For adapting to multiple subjects and to capture variability over subjects, we used subject embeddings (see Methods). The main reason for only evaluating `GPT2MEG` on group data is the comparatively much poorer performance of Wavenet-based models in evoked timeseries generation.

We were interested in whether the model generated evoked responses improved their similarity with the evoked responses from the real data, when using data from more subjects. To compare with the single-subject training we generated data using the subject embedding of that subject. The comparison of the evoked response of single-subject and group models for one 1 visual channel is shown in Figure 5. We found that generally `GPT2MEG-group` produces evoked responses that are more smoothed than the single-subject model. This is possibly because the model learns to generate data that is closer to the average statistics over subjects, and while it can adapt its generation based on the subject label, this ability is not perfect.

A.7 GPT2MEG-GROUP GENERATES CLASSIFIABLE EVOKED RESPONSES

We have shown that the channel-independent, Transformer-based model (`GPT2MEG`), can generate data with spatial, temporal, and spectral signatures similar to real data. We next investigated whether we can use this as a foundational model in a downstream task. Specifically, we look at the ability of `GPT2MEG` to aid in the decoding of experimental task conditions in visual task dataset from Cichy et al. (2016).

We first investigated whether the task responses generated by the `GPT2MEG` model can be classified with performance comparable to trials of real data. This also further tests how well the model captures spatiotemporal task-related activity and information. Furthermore, if similar performance can be obtained, then `GPT2MEG` could be used to simulate an arbitrarily large number of trials to potentially improve decoding of real data through pretraining on the simulated data. This is a form of transfer learning, where the decoding model, not the forecasting model (e.g. `GPT2MEG`), is transferred.

First, we generated 20 trials for all 118 conditions for one sample subject, using both `GPT2MEG` trained on the sample subject and `GPT2MEG-group` trained on all subjects (with the appropriate subject embedding of the chosen sample subject). We then trained separate linear neural network models on the real data (20 trials/condition) and these generated datasets, with an appropriate 4:1 train and validation set ratio. This achieved validation accuracies of 17.6% (real data), 1.9% (`GPT2MEG`), and 7.2% (`GPT2MEG-group`). In short, while the group model generates more classifiable subject-specific task-responses, it still does not reach the classification accuracy of real data. Nonetheless, this provides further evidence that `GPT2MEG-group` successfully leverages larger datasets to produce more accurate task-related activity.

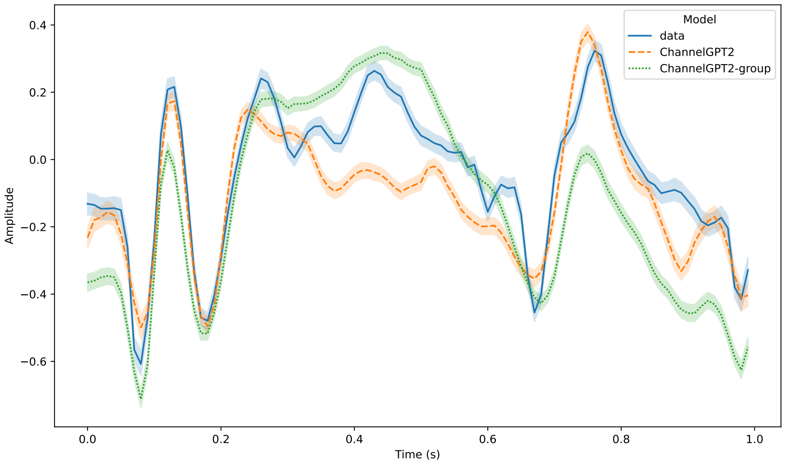


Figure 5: **Comparison of evoked responses in a visual channel (MEG2332) for a single sample subject**; using: real MEG data (blue), data generated from the `GPT2MEG` model trained on data from the single sample subject (orange), and data generated for the single sample subject (using an appropriate subject embedding) from the `GPT2MEG-group` model trained on all subjects (green). The stimulus onset is at 0 s and the stimulus offset is at 500 ms. Shading indicates 95% confidence of the trial mean. `ChannelGPT` refers to `GPT2MEG`.

A.8 Transfer learning

A key advantage of generated data is the ability to generate huge amounts of surrogate data. As in the previous section using 20 trials per condition, we generated additional datasets with 40 and 60 trials per condition using the `GPT2MEG-group` trained model. Training a decoder on these achieved validation accuracies of 7.2% (20 trials per condition), 21.7% (40 trials) and 44.2% (60 trials), exhibiting linear scaling of classification performance with the amount of simulated data.

Critically, we next assessed whether this simulated data can pretrain classifiers for transfer learning. First, we took the neural network decoder pre-trained on the 20-, 40-, and 60-trial generated datasets. We then finetuned the decoder (trained it further) on the real MEG dataset (20 trials per condition), and evaluated it on separate validation trials from the MEG data. As the quantity of generated data used for pretraining increased, accuracy of the finetuned model improved rapidly. Zeroshot (no finetuning) performance on real MEG data was above chance with 2% (20 trials per condition), 3% (40 trials), and 4% (60 trials) accuracy. Final accuracies after finetuning were 19.5% (20 trials), 21.5% (40 trials), and 23% (60 trials). Thus, each additional 20 `GPT2MEG-group` model-generated trials per condition improved final decoding by 2% on the real MEG trials. These results are summarised in Figure 6.

A.9 Ablation experiments

We performed ablation experiments with `GPT2MEG` for a single sample subject to investigate how well it can generate task-related brain activity under varied conditions without further training.

We performed two experiments to determine whether `GPT2MEG` relies solely on timing information or also utilises the semantic content of the condition labels. First, we trained a model (`GPT2MEG-randomlabel`) where the condition labels were shuffled randomly during training, breaking the semantic alignment between labels and evoked responses. Second, we trained a model (`GPT2MEG-1label`) using a single condition label for all trials. This tests whether the model cheats by learning an average evoked response instead of adapting to each task-condition.

As evident in Figure 7, models with either with shuffled or single condition labels typically failed to generate distinct evoked responses for different semantic conditions. This demonstrates that `GPT2MEG` leverages both timing and semantic information in the conditioning labels, rather than simply learning a stereotyped temporal template. Quantitatively, evoked response correlation with

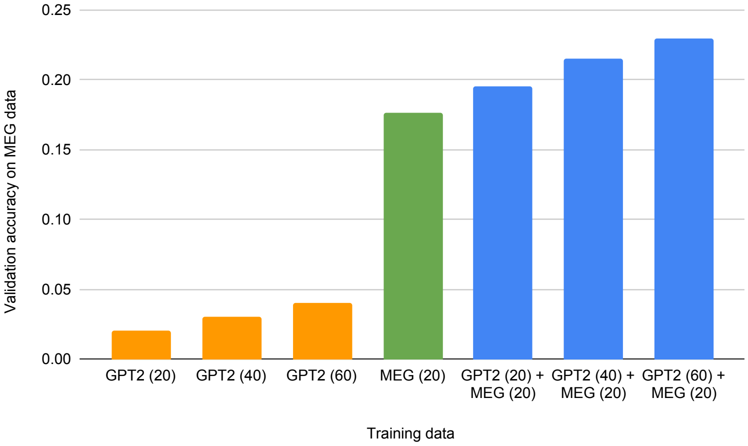
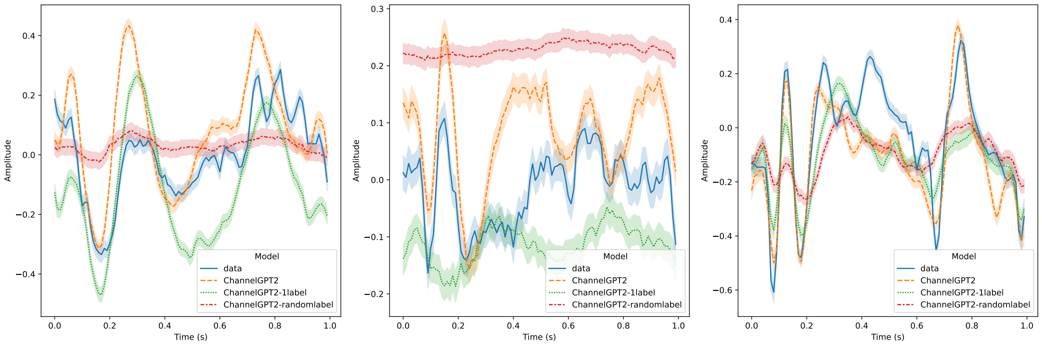


Figure 6: **Summary of the decoding accuracies of visual stimuli when using different amounts of transfer learning**. The horizontal axis represents which data the decoder was trained on. GPT2 (N) refers to the `GPT2MEG-group` generated data, while GPT2 (N) + MEG (20) is the fine-tuned decoder on the MEG data, where N is the number of trials per condition generated by `GPT2MEG-group`. The vertical axis shows the validation accuracy on the validation trials of the MEG data. Orange shows zeroshot performance, while with blue we denote the finetuned models. Chance level is $1/118$.



(a) Ch. MEG0111

(b) Ch. MEG1011

(c) Ch. MEG2332

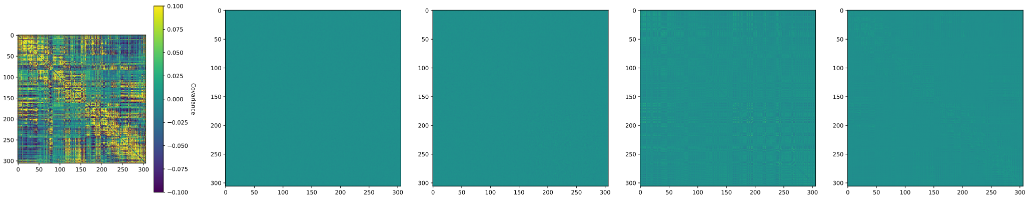
Figure 7: **Evoked responses for `GPT2MEG` models trained with shuffled or single condition labels**, indicating reliance on semantic content. Three representative channels are presented for a single sample subject. MEG0111 is anterior-left, MEG1011 is anterior-central, and MEG2332 is posterior-central. See main text for an explanation of model types. Stimulus onset is at 0 seconds, with stimulus offset at 0.5 seconds. `ChannelGPT` refers to `GPT2MEG`.

real data dropped to 44% and 56% for `GPT2MEG-randomlabel` and `GPT2MEG-1label`, respectively, compared to 74% for the full `GPT2MEG`.

We further ablated the channel and condition embeddings and analyzed the learned channel-embedding geometry; these results indicate both embeddings are important and that spatial structure emerges in the channel embeddings (Appendix A.12).

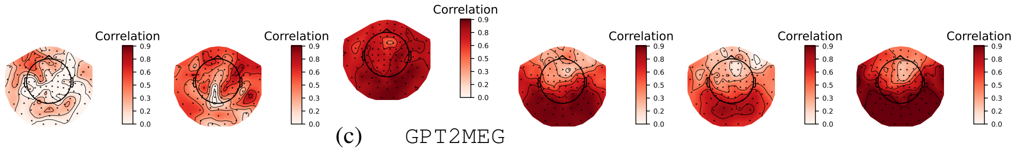
A.10 Generated covariance

As the PSD is a channel-independent measure, we also looked at generated data covariance which captures the interactions between different channels (Figure 8). All models produce data with covariances much closer to 0 than real data. This is perhaps expected for channel-independent models which generate data independently for each channel, but somewhat surprising for `WavenetFullChannelMix`.



(a) Data (b) AR(255) (c) WFC (d) GPT2MEG (e) WFCM

Figure 8: **Covariance of generated data between channels** (vertical and horizontal axes). All plots have the same scaling as (a).



(a) WFC mean (b) WFCM mean (c) GPT2MEG mean (d) WFC var. (e) WFCM var. (f) GPT2MEG var.

Figure 9: (a)-(c) **Correlation between the time-courses of the mean (over individual epochs) evoked responses**; from the real MEG data for a single sample subject and the mean evoked responses from data generated by the different forecasting models trained on the single sample subject. (d)-(e) **Correlation between the time-courses of the variance (over individual epochs) of the mean evoked responses**; from the real MEG data for a single sample subject and the mean evoked responses from data generated by the different forecasting models trained on the single sample subject. For all figures, the correlation values are visualised across sensors. WFC refers to WavenetFullChannel and WFCM refers to WavenetFullChannelMix. Darker reds indicate higher correlation.

A.11 Additional evoked-response analyses

To quantify the similarity between real and model generated evoked activity, we computed the correlation of the mean (across individual epochs) time-courses of the evoked response for each channel separately. Note that we averaged over the different MEG sensors (the magnetometers and gradiometers) found at the same location. The result of this is plotted in Figure 9, allowing insights into the spatial pattern of similarity.

As expected, GPT2MEG generates data with evoked responses that have much higher correlation with evoked responses from real data, and slightly higher correlation in visual areas compared to other channels, matching the known topography of visual evoked responses. In other models the correlation is low, and spatially better in frontal areas, likely because the evoked responses here are noisier providing an easier fit.

Figure 9 also shows the correlation between the variance (over individual epochs) time-courses of the mean evoked response obtained from the actual data and the evoked responses obtained from data generated by each model. This captures a measure of the ability of the models to represent the trial-to-trial variability found in the real data. Again, GPT2MEG generates data that has the highest correlations with the real data, with higher values in channels in the back of the head, appropriately capturing the topography of response variability. Other models have similar spatial distribution, and notably WavenetFullChannel also produces evoked responses with variance partially matching the real data.

Finally, a different way to assess task-related activity is to examine the evoked state time-courses from the HMMs fitted on the real and model generated timeseries. Rather than looking at individual channels, this provides an overall view of which HMM state gets activated when, during individual trials. This is computed by simply epoching the state timecourse, and averaging over all trials. We plot these for the real data and each generated timeseries in Figure 10. As expected, the HMM trained on models other than GPT2MEG shows poor evoked state time-courses. GPT2MEG generated data produces states with similar evoked dynamics and variability as the real data. In the next section we show how this generalizes over multiple subjects.

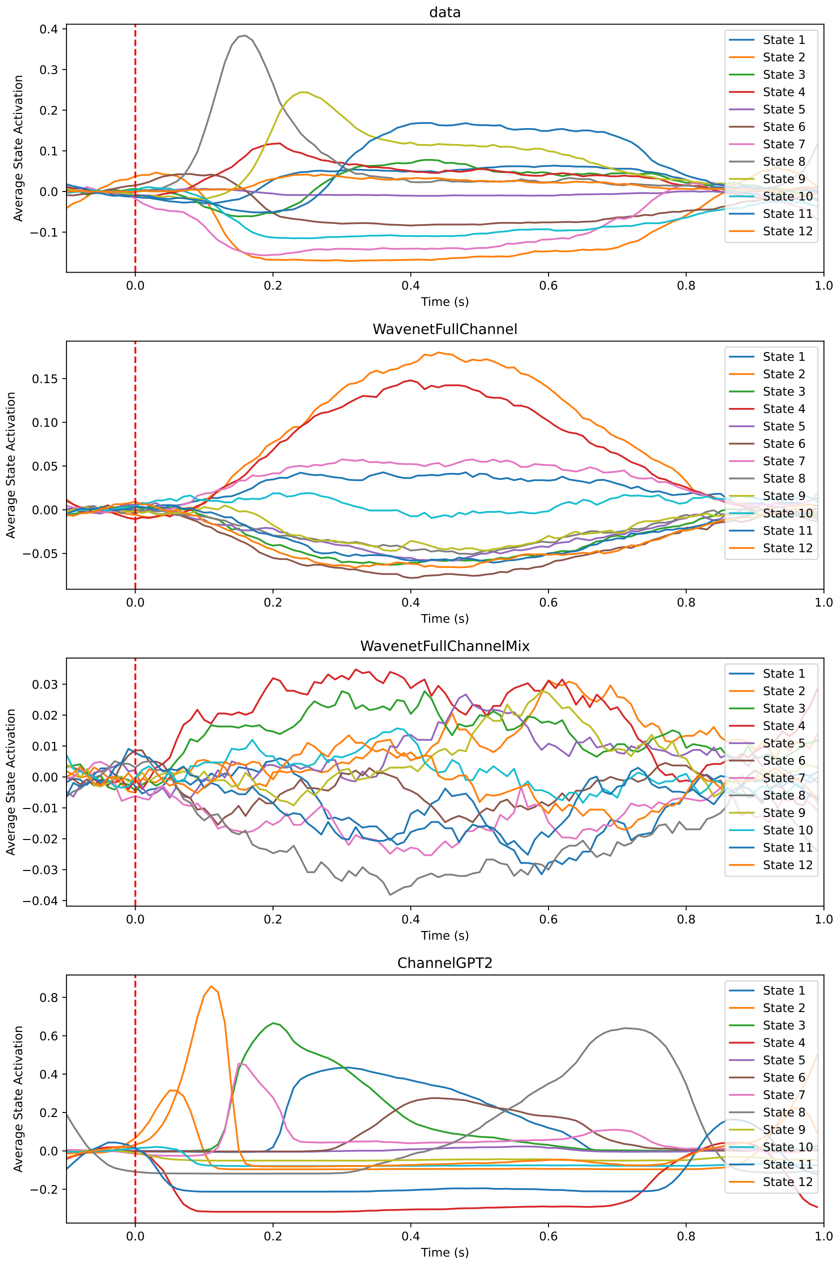
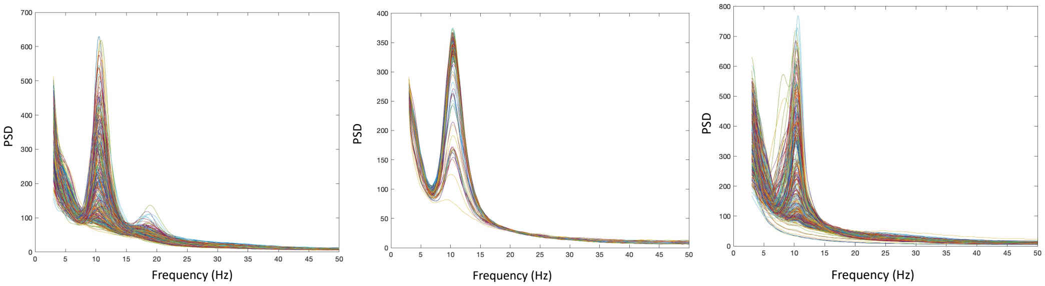


Figure 10: **Evoked response state timecourses of HMMs inferred on a single sample subject**; using real MEG data (top), and on generated data from each of our task-conditioned models trained on the single sample subject. Note that the HMM states are not matched between models. Image presentation starts at 0 seconds and ends at 0.5 seconds. `ChannelGPT` refers to `GPT2MEG`.

A.12 Additional ablations and channel-embedding analysis

We also investigated the contributions of the channel and condition embeddings, by training two separate ablated models. As shown in Figure 11, removing the channel embeddings resulted in very similar PSD across channels in the generated data, indicating that the model relies heavily on these embeddings to adapt generation per channel. The evoked responses in Figure 12 confirm that without channel embeddings, variability between channels is reduced. Removing the condition embeddings resulted in noisier power spectra of the generated data and no 20 Hz peak.

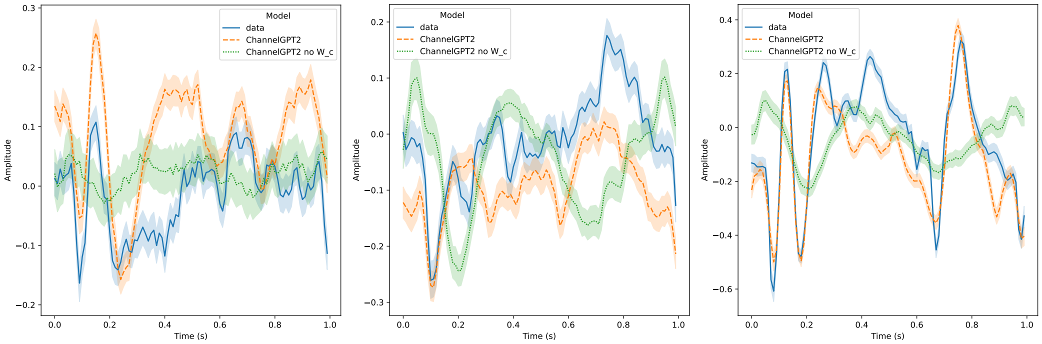


(a) Full GPT2MEG

(b) No channel embedding

(c) No condition embedding

Figure 11: **Comparison of generated power spectra with different ablations**: (a) full GPT2MEG model, (b) GPT2MEG with ablated channel embeddings and (c) GPT2MEG with ablated condition embeddings. Shown for a single sample subject. Both channel and condition embeddings are critical for accurate spectral content.

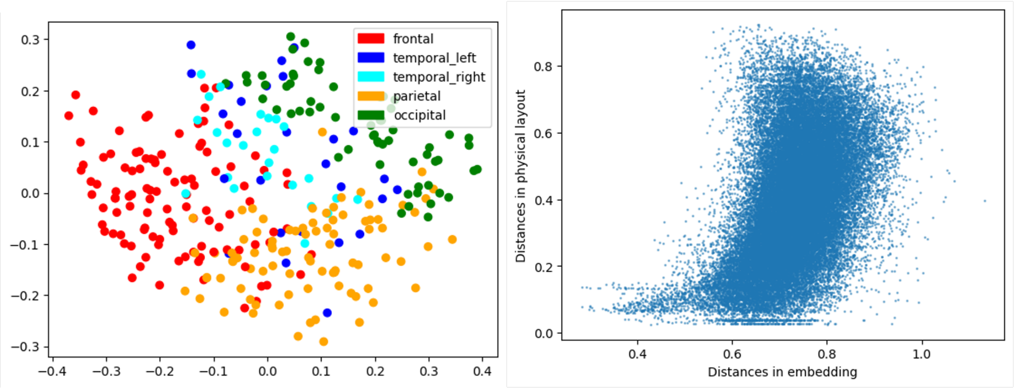


(a) Channel 1 (MEG1011)

(b) Channel 2 (MEG1813)

(c) Channel 3 (MEG2332)

Figure 12: **Comparison of generated evoked responses with ablated channel embeddings** in the GPT2MEG model, shown across 3 representative channels (a)-(c) and for a single sample subject. Without channel embeddings the model fails to adapt evoked responses to different channels. The stimulus onset is at 0 seconds and the offset is at 0.5 seconds. ChannelGPT refers to GPT2MEG.



(a) PCA

(b) Correlation

Figure 13: **Visualisation of channel embeddings.** (a) 2D projection of the channel embeddings from GPT2MEG-group with PCA. Channels are coloured by their location on the scalp grouped into 5 major brain areas. (b) Plotting pairwise Euclidean distances of channels in real, physical space versus embedding space. Sensors that are near to each other in the real sensor montage tend to have more similar embeddings. Each point represents a different pair of channels. Correlation is 0.45.

Finally, we found that the channel embeddings encode spatial relationships, as sensors that are near to each other in the real sensor montage tend to have more similar embeddings. This is shown through a PCA projection of the embedding space in Figure 13. Correlation between pairwise Euclidean distances of channels in physical space and embedding space was 0.45 (Figure 13b).