Neuroprobe: Evaluating Intracranial Brain Responses to Naturalistic Stimuli

Andrii Zahorodnii [1,2]* **Bennett Stankovits**[1]* **Christopher Wang**[1]*

Charikleia Moraitaki[1] **Geeling Chau** [3] **Ila R Fiete** [1,2] **Boris Katz**[1] **Andrei Barbu**[1]

[1]MIT CSAIL, CBMM [2]MIT McGovern Institute [3]Caltech

Abstract

Understanding the relationship between the various tasks the brain performs can shed light on its functional organization. We introduce a benchmark, Neuroprobe, which targets a wide range of multimodal tasks. Neuroprobe borrows several ideas from modern natural language processing: using large scale naturalsitic datasets, probing at scale across tasks as a means to understand black box systems, and evaluating on large benchmarks that test many different skills. For artificial networks, probe analysis attempts to decode attributes from different layers. It is one of the main vehicles used to shed light on the relationship and dependencies between tasks and the algorithms that networks learn. While prior neuroscience benchmarks tend to focus on a single or a very small number of tasks, Neuroprobe uses a fixed set of subjects with a large amount of data across many annotated tasks, which will allow us to create an integrated picture. Furthermore, the results obtained from Neuroprobe evaluations can yield time-orderings between different tasks and recover the functional relationships between tasks that reveal properties of the algorithms the brain uses. The main remaining bottleneck to achieving these type of results is that decoding performance for many tasks is very poor. We demonstrate a few tasks both with simple linear decoders and neural foundation models, then introduce a large number of additional attributes that should, in principle, be decodable but are not. Neuroprobe gives us an opportunity to build higher accuracy decoders, better neural foundation models that are tested across many tasks, and to bring neuroscience closer to the methodology that has worked so well in natural language understanding, and to ultimately discover the functional organization of the brain across many tasks. We make our code publicly available [2] and will maintain a leaderboard [3] to track model progress upon publication.

1 Introduction

The human brain constantly engages in a variety of processing tasks simultaneously: parsing speech, interpreting visual scenes, and performing social reasoning (Schurz et al., 2014). However, a cohesive picture of how these computations are organized across time and regions in the brain remains poorly understood. While modern neuroscience offers glimpses into individual functions, a central challenge is that typical experiments isolate one or two tasks at a time, often using simplified stimuli and contrived lab settings (Nastase et al., 2020). A solution suggests itself from the field of machine learning interpretability, which has developed methods to reverse engineer neural network black

*Equal contribution.

[2]`https://github.com/azaho/neuroprobe`

[3]`https://neuroprobe.dev`

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

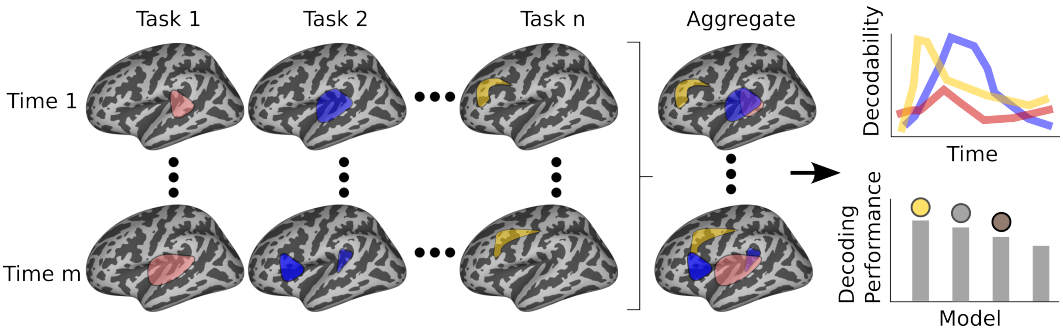


Figure 1: **Overview of Neuroprobe's goals.** Neuroprobe consists of machine learning classification tasks derived from intracranial recordings aligned with annotated stimuli. By running a decoding analysis for each task, we can localize various aspects of multimodal language processing in the brain. Moreover, we can segment the neural recordings by time, repeat the decoding analyses across time bins, and discover a time evolution of each task. Previously, neuroscience experiments have been small, and focused on one task at a time. The results of our analyses can be combined to give a comprehensive picture of language processing in the brain. From this, two things can be achieved. First, we can derive neuroscience insights such as the relative timings for processing of certain tasks. Second, the tasks themselves can be used as a benchmark of neural decoding models.

boxes via probing experiments, e.g. Tenney et al. (2019); Alain & Bengio (2016). These methods are powerful, but there is an obstacle in applying them to study the brain: decoding the contents of brain activity remains a challenging task (Paninski & Cunningham, 2018). While intracranial data offers high temporal and spatial resolution, the raw signals are noisy and high-dimensional. To these ends, we introduce *Neuroprobe*, a benchmark that is designed both to be a setting in which neuroscience probing experiment may be run *and* as a measure of progress to spur improvement of neural decoding models.

Neuroprobe contains 19 decoding tasks that span vision and language, all on the same subjects and the same neural recordings collected while subjects watched movies. Having many different tasks on the same dataset allows one to derive constraints on the relationships between tasks, such as: What is the temporal order between tasks across many subjects? Which tasks share neural real estate? How does latency in one task influence latency in another task? These constraints can then narrow the space of algorithms to regularize models of brain function. Unfortunately, as mentioned, decoding today for many tasks is nowhere near accurate enough to systematically derive these kinds of constraints. So, we develop a public leaderboard for hosting submissions to the Neuroprobe benchmark. As submissions to the leaderboard increase, decoding accuracy will increase, in turn raising our confidence in the spatial and temporal distribution of different tasks uncovered by the probing experiments.

Meanwhile, on the modeling front, more and more foundation models are being developed for neural recordings. There has been an explosion of neural foundation models as of late, including: Neuroformer (Antoniades et al., 2024), BrainBERT (Wang et al., 2023), PopT (Chau et al., 2024), STNDT (Le & Shlizerman, 2022), NDT2 (Ye et al., 2023), MBrain (Cai et al., 2023), Brant (Zhang et al., 2023), MtM (Zhang et al., 2024b), and POYO (Azabou et al., 2023). Most of these models are not tested on standardized decoding tasks. There are few cross-task decoding datasets at present for testing new neural foundation models. This runs contrary to one of the main selling points of foundation models for neuroscience, which is that they will improve decoding accuracy to enable neuroscientists to run more experiments on a variety of tasks with less data. In addition, in a sense the space of tasks determines the space of models considered, since only models that can show an advantage are selected for and published. It is a long term problem for the community that larger batteries of decoding tasks are not a common evaluation practice. This is already reflected in our findings that state of the foundation models for neural recordings don't make a massive different in decoding performance for some tasks, and can even hurt it in a few cases (see section 4).

We have designed Neuroprobe to be usable by members of the ML community even if they have no particular knowledge of neuroscience. Anyone can easily run models and contribute new ideas.

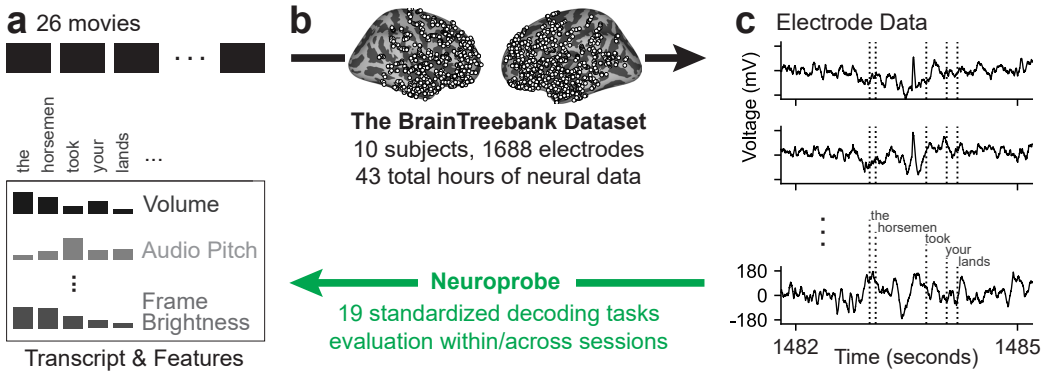


Figure 2: **From raw data to decoding tasks.** As part of the BrainTreebank dataset, 26 movies (a) are watched by 10 patients with stereoelectroencephalography electrodes implanted in various brain regions (b), and the local field potential from the implanted electrodes is recorded (c). Neuroprobe turns this dataset into an evaluation benchmark by segmenting the aligned data into various audio, language, and vision decoding tasks, such as, loudness and pitch of the audio, average pixel brightness, etc.

While Neuroprobe provides the analysis tools to interpret better decoding results. Lowering the barriers to entry ensures that we have a healthier community and attracts many more researchers to these problems.

Neuroprobe, see Figure 2, is derived from the Brain Treebank (Wang et al., 2024), which consists of intracranial neural recordings aligned with the corresponding movie stimuli. The dataset contains annotations from which we derive 19 decoding tasks, see Supplementary Table 1. We select the BrainTreebank because it is at the scale at which modern NLP begins to operate and models being to be understood (43 hours of recordings): comparable to datasets on low-resource languages.

In addition, we standardize a number of aspects of the benchmark. We select test/train splits in different conditions: all the way from training and testing on the same subject and movie, to doing cross-subject cross-movie decoding. We host a centralized website that aggregates results, both as a whole and also by split-type and task, using a JSON schema to validate submissions.

Our contributions are:

1. A new large-scale multitask decoding benchmark: Neuroprobe.
2. Standardized splits and methods to rank neural foundation models and encourage their development in a direction which benefits decoding tasks.
3. Results from a set of baselines and state-of-the-art models on Neuroprobe.
4. An early analysis of the timings and spatial distribution of different task processing pathways in the brain.

In the long run we hope that Neuroprobe will both lead the way to an understanding of the general architecture of the computations that the brain performs as well as bring the ML and neuroscience communities into closer alignment by translating interesting neuroscience questions into questions that are easily digested and then improved on by the ML community.

2 Related work

While there are many publicly available neural recordings that neural decoding models have been developed on, neuroscience still suffers from a dearth of standardized, easy-to-run machine learning benchmarks. This lack of defined decoding tasks, standardized train/test splits, and metrics make it difficult to compare models.

Neural recording datasets The most recently developed models for neural data have relied on several widely accessible datasets. For non-invasive EEG decoding, datasets from Zheng & Lu (2015); Grootswagers et al. (2022); Bhattasali et al. (2020); Tangermann et al. (2012); Obeid & Picone (2016); Broderick et al. (2018); Brennan & Hale (2019) have been used in the construction of

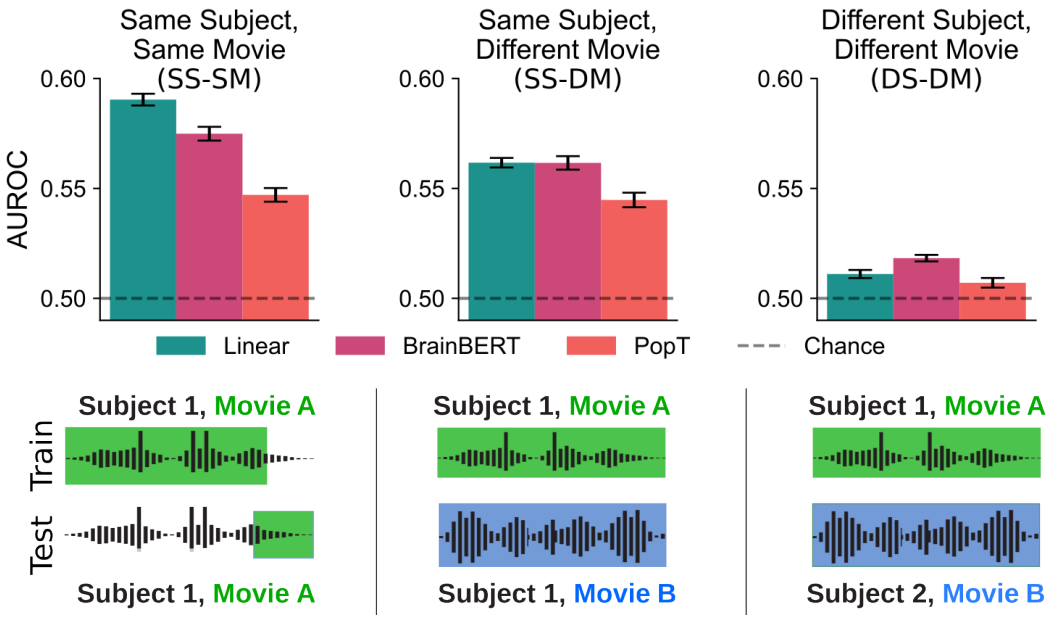


Figure 3: **Neuroprobe splits**. We perform analyses on three different types of splits. In *same subject/same movie* (SS-SM) we train on data from one subject and one movie segment, and evaluate on the same subject, but another segment of the same movie. Performance is measured via cross-validation. In *same subject/different movie* (SS-DM), we train on data from one subject and from one movie. Then, we evaluate on another movie. In *different subject/different movie* (DS-DM), we train on data from one subject and one movie and evaluate on data from an entirely different subject and movie. This is the most challenging split.

models such as those proposed by Jiang et al. (2024); Yang et al. (2023); Yuan et al. (2024); Défossez et al. (2023). For fMRI decoding, (Wehbe et al., 2014; LeBel et al., 2023; Nastase et al., 2021; Li et al., 2022; Allen et al., 2022) have led to models such as those proposed by Scotti et al. (2024); Ozcelik & VanRullen (2023). For MEG decoding, Jan-Mathijs et al. (2019); Hebart et al. (2023) have lead to models such as those proposed by Défossez et al. (2023); Benchetrit et al.. For neural spike decoding Perich et al. (2025); Churchland et al. (2024); Manley et al. (2024); IBL (2024) have lead to models such as those proposed by Azabou et al. (2023); Zhang et al. (2024a). For broadband intracranial neural activity, datasets from (Peterson et al., 2022; Wang et al., 2024; Nejedly et al., 2020) have fueled the development of models proposed by (Peterson et al., 2021; Wang et al., 2023; Chau et al., 2024) However, these datasets do not provide rigorous splits or testing guidelines, so each model is difficult to compare to others.

Existing neural data benchmarks There are a few benchmarks involving neural data. Some of the earliest involve EEG BCI decoding (Tangermann et al., 2012), but are limited in data quality and scale by today's standards. The NaturalScenesDataset (Allen et al., 2022) is close to being a benchmark in that they have splits, but it primarily benchmarks fMRI data, and focuses on visual processing. The clinical-grade Temple University Hospital EEG dataset (Obeid & Picone, 2016) can also be used as a benchmark, but it only contains EEG and has the labels are limited to seizure detection. Benchmarks for neural spikes are proposed by Pei et al. (2021); Karpowicz et al. (2024); Willett et al. (2023); Lueckmann et al. (2025), but these only contain spiking information rather than broadband signals from ECoG or sEEG that capture more neural activity (Parvizi & Kastner, 2018). A benchmark like Neuroprobe for high fidelity intracranial signals with corresponding challenging naturalistic language stimuli is still needed to allow the field to progress forward in building better neural decoding models.

3 Approach

Brain Treebank Neuroprobe is an evaluation-only benchmark environment that uses the raw data from the BrainTreebank (Wang et al., 2024), a publicly available dataset released under a CC

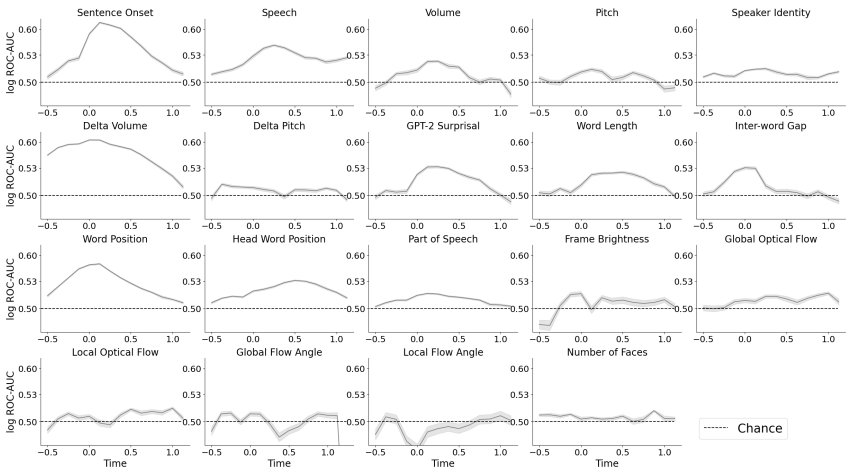


Figure 4: **Neuroprobe enables tracking of information processing in the brain across tasks.** A linear model is fit for a sliding 125ms window of activity. Here, we show the performance of the most decodable 100 electrodes per each task. Error bars show standard error across electrodes. Performance is plotted on a log scale to show trends for tasks that have lower decodability. The x-axis shows time, where $t = 0$ corresponds with word onset. By plotting decoding performance across time, the time course of information availability for each task becomes visible. Audio-linguistic tasks, such as Speech vs. Non-speech, are most decodable closest to word onset.

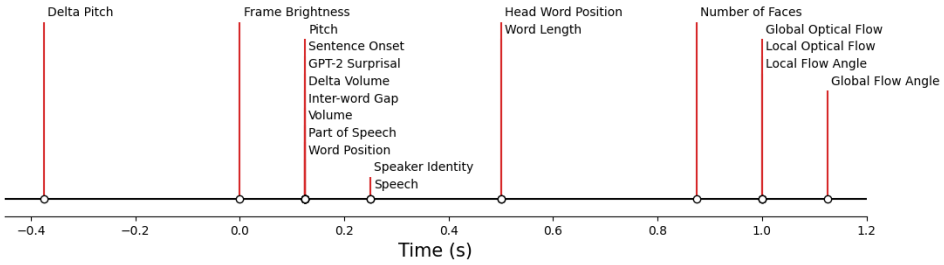


Figure 5: **Time-ranking of decodability** A simple method of finding relationships between tasks it look at when each task is decodable. Consistencies in this order across subjects are an indication of a dependency between tasks. A shortcut to that, is to further restrict ourselves to when each task achieves maximum decodability. Note that we use a window of 125 ms which gives fairly coarse temporal localizations, which is why many tasks overlap. We can already observe some patterns from these results, even with poor decoding accuracy. Notably, *Word head position*, a semantic feature that pertains to the position of the dependency parse head, is decoded later than other language features. A caveat should be offered that these timings are dependent on the type of decoding analysis being performed. As different decoding methods are developed which solidify our ability to decode each task, it is certain that these ordering will change.

BY 4.0 license. The Brain Treebank is a large-scale dataset of intracranial electrophysiological recordings (stereoelectroencephalography; sEEG) collected while 10 human subjects (5 male, 5 female, ages 4–19; Supplementary Table 3) watched 26 total Hollywood movies (Supplementary Table 4). Electrode placements for each subject and their speech-selective responses are shown in Supplementary Figure 10. Spanning 43 hours of neural activity, the dataset aligns recorded brain signals with transcribed and manually corrected speech, word onsets, and universal dependency parses across the 223,068 words in 38,572 sentences. This dataset enables the systematic evaluation of computational models on multimodal neural decoding tasks.

Decoding tasks We use the movie annotations and the alignment with the corresponding neural data to create a suite of 19 decoding tasks, spanning visual, audio, and language domains. For every task, the neural data is the input and the annotation label is the target output, where we formalize all of the

tasks as binary classification by thresholding the labels. For example, for the GPT2 Surprisal task, the positive label corresponds to surprisal annotations above the 75%th percentile of the distribution within a session, and the negative label to the values below the 25%th percentile. For non-scalar labels (such as speaker identity or part of speech of the word) we pick a main target class (i.e. most frequent speaker, or Verb for the part of speech task), and formulate the task as one-versus-rest classification. See more details in Appendix A.

Splits The Neuroprobe evaluation takes place across three different types of splits. For the *same subject/same movie* (SS-SM) splits, train data and test data come from a single movie-viewing session. Decoding results are cross-validated with an 80-20 train-test split. Importantly, the indices for the cross-validation splits are not drawn from the whole movie uniformly, but rather the train examples are taken from a single contiguous block and the validation examples are taken from a separate block. This is done to prevent models from over-fitting to auto-correlation in the signal.

For the *same subject-different movie* SS-DM split, the train data consists of examples drawn from the longest movie viewed by a given patient, and the test data comes from the second longest movie.

For the *different subject-different movie* DS-DM split, the train data consists of data from a single session (trial 4), viewed by subject 2, chosen because this is the longest trial and the subject with the most electrodes in both hemispheres. Testing then consists of the average performance across selected sessions for all other subjects (see Appendix F). This split in particular presents a demanding test of model generalizability, especially since electrode placements vary widely between patients (see Figure 10).

Experiments In Neuroprobe, experiments can either be performed at the *single-electrode* level or the *population* level, i.e., using all electrodes in a given subject as model input. To give a sense of the types of neuroscience insights that can be derived from Neuroprobe, we perform a collection of single-electrode analyses across the SS-SM splits for all BrainTreebank sessions. In particular, for each task, we fit a linear classifier to do decoding over a fixed window of activity (250 ms). This window slides along a longer period, from 0.5s before word onset to 1.25s after word onset, with a stride of 125ms. This provides a picture of the time-course of decodability in the brain. Electrodes marked as corrupted in the original BrainTreebank dataset are excluded. See Section 4.

Neuroprobe-Lite Benchmark Outside of analyses described above, for the purposes of comparing models, running experiments over all sessions and electrodes is prohibitively expensive. To this end, we subset the data to create Neuroprobe-lite by selecting a smaller portion of subjects and sessions (6 subjects, 2 trials each) for training and evaluation.

Furthermore, the total number of electrodes per subject is capped at 120. The electrodes in Neuroprobe-lite were chosen specifically to cover as much of the brain in each participant as possible. This was done by randomly taking a specified proportion of electrodes from every probe, to ensure that every probe is represented in the Neuroprobe-lite data features. This ensures that the input for each task is standardized matrix which has predictable memory and computational requirements. We maintain a public leaderboard which will display model performance on this benchmark, both on the single-electrode and population level; see Supplemental fig. 12.

Models To show the utility of the Neuroprobe tasks as a benchmark, we evaluate on a few baselines and models. For the purposes of benchmarking, all models are run on Neuroprobe-lite (see above). All inputs are given as a population, i.e., the data from all electrodes is provided as input, concatenated. The models we benchmark span the range of simple classifiers to large, pretrained models. These include three linear regression models, which take as input either the raw voltage time-series inputs, Fourier transform input, or Short-time Fourier transform (STFT) inputs. For pretrained models, we also train a regression on BrainBERT (Wang et al., 2023) inputs, and fine-tune a linear layer on top of a pretrained PopT (Chau et al., 2024), a pretrained transformer for encoding arbitrary sets of electrodes. More details on the models available in Appendix H.

Metric calculations The primary evaluation metric was the Area Under the Receiver Operating Characteristic curve (AUROC), aggregated across electrodes. We adjusted the aggregation strategy to be compatible with each model to obtain the different subjects-different movie DS/DM results shown in Figure 3. Before running our linear regressions, we preprocessed the neural data to represent activity in each cortical region (using averaging per subject/trial pair), as defined from the 34 regions by the Desikan-Killiany atlas. Similarly, we ran BrainBERT, with the same region averaging strategy.

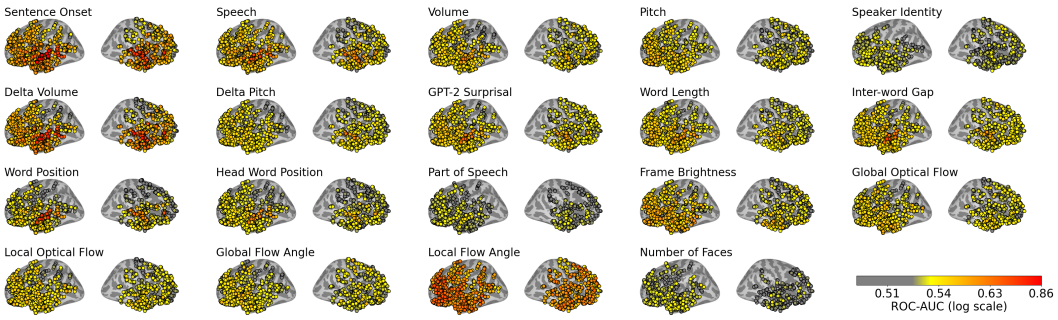


Figure 6: **Distribution of task processing throughout the brain** A linear decoder is trained on the *single-subject/single-movie* split. Color shows ROC-AUC on a logarithmic scale. Performance is computed by averaging over cross-validation folds ($k = 5$) and movies and then taking a max over time bins. Language features like *Sentence Onset* and *GPT-2 Surprisal* are most decodable in the temporal and frontal lobes.

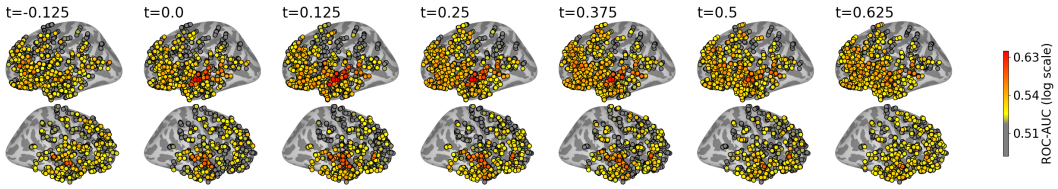


Figure 7: **Time evolution of surprisal decodability throughout the brain.** The decodability of features vary in both time and space. Close to word onset ($t = 0$), surprisal is most decodable in the superior temporal gyrus. Time zero here refers to the onset of a given word. Most words are interior to sentences or to conversations. Since most modern surprisal metrics are contextualized, one can immediately predict surpiral even from the neural activity left over from prior words As time progresses, surprisal becomes more decodable in the frontal areas. Full progressions for all tasks can be seen in Appendix L and in a movie at this url: `https://neuroprobe.dev/neuroprobe_time_course.mp4`.

For the PopulationTransformer we use all electrodes that can be bipolar-rereferenced and are in the set of 'clean' electrodes (see (Chau et al., 2024)) for evaluation. No accomodation for the DS/DM split was necessary for the PopulationTransformer, which is designed to handle subject-transfer.

4 Results

Timing analysis To investigate the time course of linguistic information processing in the brain, we aligned neural data to word onsets and split it into narrow time-bins (width $= 125$ms), and train a separate linear decoder on each bin for multiple tasks. Decodability is computed as the average across cross-validation folds ($k = 5$). For each task, we restrict our attention to the top 100 electrodes with the highest decodability. Decoding performance as a function of time shows the course of processing after the word onset ($t = 0$, Figure 4). Interestingly, the beginning of a new sentence can be decoded with better-than-chance AUROC even before the word onset ($\mu = 0.53, \sigma_M = 0.0015$ at -250ms), hinting at the predictive nature of processing. Moreover, we can find a time-ranking of features by looking at when decodability peaks for reach feature (Figure 5). For example, we note that the high-level semantic feature 'word head position' is decodable only later (decodability peaks at $t = 0.5s$ vs. volume and pitch at $t = 0.125s$).

Spatial analysis By examining the linear decodability of features, a picture emerges of which features modulate activity in which areas of the brain (Figure 6). Using the single electrode analysis, we find that audio-linguistic tasks such as 'sentence onset', 'speech vs. non-speech', 'delta volume' are most decodable in the superior temporal gyrus, especially close to Herschel's and Wernicke's area, with

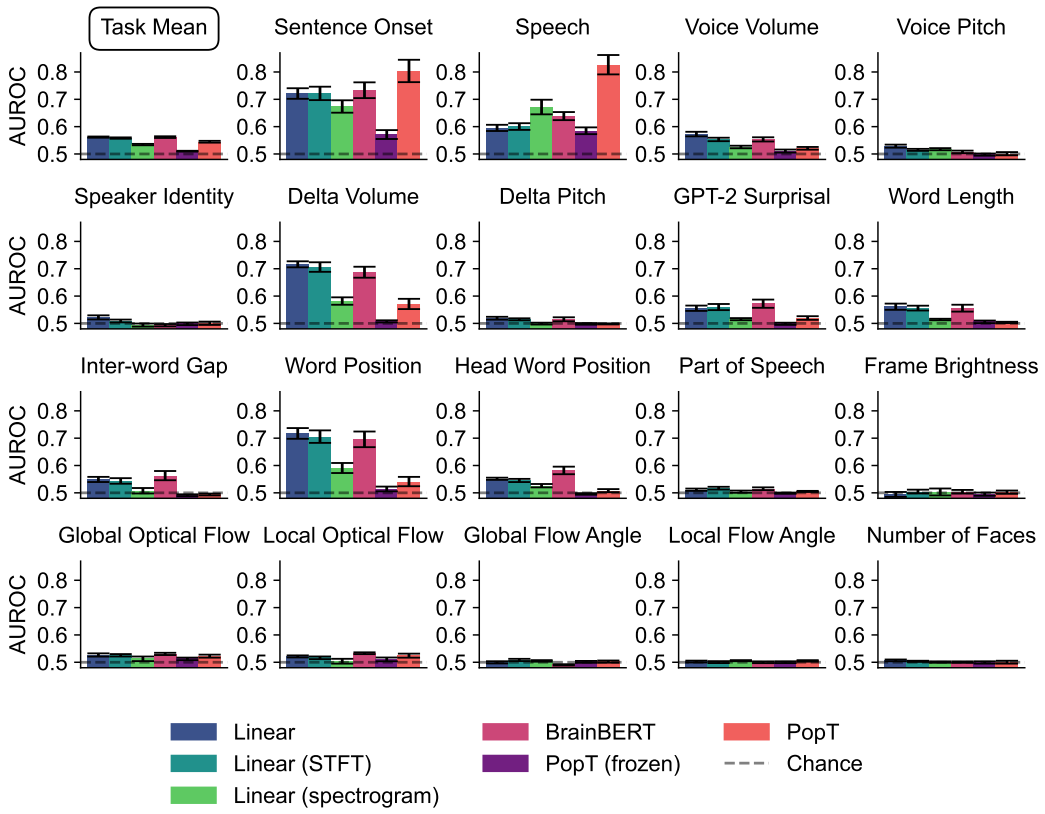


Figure 8: **Performance of baseline models on the 19 tasks of Neuroprobe.** Evaluation is done on the same subject, same trial (SS-ST), using 5-fold cross-validation. Normalized audio volume traces and the distribution of detected faces with corresponding word counts are shown in Supplementary Figures 9 and 11, respectively. The performance of four models is shown: (1) logistic regression either from raw voltage signal of all electrodes to the labels, or (2) from the spectrogram of the signal to the labels, as well as (3) BrainBERT (Wang et al., 2023) and (4) PopulationTransformer (Chau et al., 2024). Neural data was cut to include one second following each word onset. In case of multi-class classification, AUROC was computed using a one-vs-all strategy and averaged together. Performance on different trials for the same subject were averaged together. Error bars denote s.e.m. across all subjects. These results can be seen in tabular form in Appendix I.

average AUROCs of 0.61, 0.55, and 0.62, respectively in the gyrus of the temporal transverse. Here region results are given with respect to the Destrieux atlas; see Appendix M.

Spatio-Temporal analysis We do a deep dive on the surprisal feature and show that after word onset, it is most decodable in the temporal lobe (AUROC $= 0.58$ at $t = 0$ in the transverse temporal), but decodability spreads to the frontal lobe as time progresses (AUROC $= 0.50$ at $t = -0.125$ and AUROC $= 0.52$ at $t = 0.5$); see Figure 7. A movie of this for all tasks can be seen at `https://neuroprobe.dev/neuroprobe_time_course.mp4`.

Comparison of basic decoding methods on Neuroprobe. We compare the performance of two simple baseline models—logistic regression applied to raw voltage signals and logistic regression applied to spectrogram features—across the 19 decoding tasks in Neuroprobe. Performance is evaluated using area under the receiver operating characteristic curve (AUROC), with chance-level performance ($ROC = 0.5$) included for reference. We also compare with BrainBERT and PopT using their publicly released off-the shelf-weights. Because of this there may be some discrepancy due to the fact that both models were trained on 5s intervals, whereas we train on 1s intervals across all models for consistency. In general, linear decoding is very good (see Figure 3), achieving the best overall performance on the SS/SM (0.590 ± 0.003) split, with the second best model being BrainBERT (0.575 ± 0.003). On the SS/DM split, the linear baseline tied BrainBERT (0.562 ± 0.002 vs

0.562 ± 0.003, respectively), outperforming PopulationTransformer (0.545 ± 0.003). But BrainBERT performs the best on the difficult DS/DM split (0.518 ± 0.001) with the next best model being the linear baseline (0.511 ± 0.002).

Finally, for SS-SM, a breakdown by task can be seen in Figure 8. The PopulationTransformer, despite being pretrained, underperforms on many tasks, but achieves the highest performance on the Sentence Onset and Speech vs. Non-speech tasks.

5 Conclusion

Neuroprobe can be used in several ways by different communities: (1) Machine learning practitioners can contribute by improving decoding performance. (2) At the intersection of ML and neuroscience, Neuroprobe can be used to assess how good a given neural foundation model is at improving decoding accuracy. (3) Neuroscientists can use Neuroprobe to uncover relationships between different tasks that the brain executes which puts constraints on the kinds of algorithms our brains are using.

Using Neuroprobe, questions about processing in the brain become machine learning decoding tasks which can be rapidly iterated on. This will drive improvements both in decoding ability and the ability to draw neuroscience conclusions from large scale data. As we have seen in other fields, this can also lead to a virtuous cycle in which neuroscientists are encouraged to share more datasets to the effort.

Despite the weakness of current decoding models, Neuroprobe can still find interesting trends in both the spatial and temporal organization of tasks in the brain. As decoding models improve, the clarity of such findings will improve and their variance will decline. Each decoding task induces a map across the brain of when and where processing specific to that task is performed. By overlaying many of these maps, a functional picture of the brain emerges of which language, vision, and audio features modulate activity in each region. We see this approach as a way of answering the long-standing neuroscience question: What is the underlying circuit basis of language processing in the brain?

Limitations Our decoding results from the baselines we tested are low for a few tasks, such as speaker identity and pitch, and thus drawing any conclusions from their results is fraught. While our data offers unprecedented combination of scale and resolution, it is collected from a clinical population undergoing invasive monitoring, and results should not be overgeneralized. We only have 10 subjects currently. This is because it is difficult to obtain this kind of data, which requires invasive surgery to implant electrodes. However, each subject has many sessions.

Broader impacts Neuroprobe provides a standardized benchmark for evaluating models of human brain activity, with potential applications in neuroscience, machine learning, and clinical technologies such as brain-computer interfaces. By releasing our data, code, and leaderboard, we aim to democratize access to high-quality neural benchmarks and foster cross-disciplinary collaboration.

Future work Our framework is general enough to accommodate future annotations, allowing for investigations of low-level language processing, such as part of speech, or high-level semantic processing such as thematic roles or language model embeddings. We also seek, in near-term future work, to add to the library of tasks and datasets in Neuroprobe. As we continue to build out the benchmark, researchers will be able to study the question of how various tasks interact with each other.

6 Acknowledgements

This work was supported by the Center for Brains, Minds, and Machines, NSF STC award CCF-1231216, the NSF award 2124052, the MIT CSAIL Machine Learning Applications Initiative, the MIT-IBM Watson AI Lab, the CBMM-Siemens Graduate Fellowship, the DARPA Mathematics for the DIscovery of ALgorithms and Architectures (DIAL) program, the DARPA Knowledge Management at Scale and Speed (KMASS) program, the DARPA Machine Common Sense (MCS) program, the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000, and the Air Force Office of Scientific Research (AFOSR) under award number FA9550-21-1-0399. This work also has been supported by ONR award N00014-19-1-2584, by NSF-CISE award IIS-2151077 under the Robust Intelligence program, by the ARO-MURI award W911NF-23-1-0277, by the Simons Foundation SCGB program 1181110, the K. Lisa Yang ICoN Center, the Caltech Chen Institute,

and the Caltech Carver Mead New Adventures Fund. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

References

International brain lab. https://internationalbrainlab.org, 2024. Accessed: 2024-11-23.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*, 2016.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and Multitask Generative Pretraining for Brain Data, March 2024.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A Unified, Scalable Framework for Neural Population Decoding, October 2023.

Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: toward real-time reconstruction of visual perception. october 2023. In *URL https://openreview. net/forum*.

Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. The alice datasets: fMRI & EEG observations of natural language comprehension. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 120–125, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.15/.

Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.

Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. MBrain: A Multi-channel Self-Supervised Learning Framework for Brain Signals, June 2023.

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, October 2024.

Mark Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. Neural population dynamics during reaching. Data set, 2024. URL https://dandiarchive.org/dandiset/000070/draft.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10): 1097–1107, 2023.

Tijl Grootswagers, Iris Zhou, Austin K. Robinson, et al. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9:3, 2022. doi: 10.1038/s41597-021-01102-7. URL https://doi.org/10.1038/s41597-021-01102-7.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Schoffelen Jan-Mathijs, Robert Oostenveld, Lam Nietzsche HL, Uddén Julia, Hultén Annika, and Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, 6(1), 2019.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Brianna M Karpowicz, Joel Ye, Chaofei Fan, Pablo Tostado-Marcos, Fabio Rizzoglio, Clay Washington, Thiago Scodeler, Diogo de Lucena, Samuel R Nason-Tomaszewski, Matthew J Mender, et al. Few-shot algorithms for consistent neural decoding (falcon) benchmark. *Advances in Neural Information Processing Systems*, 37:76578–76615, 2024.

Trung Le and Eli Shlizerman. STNDT: Modeling Neural Population Activity with a Spatiotemporal Transformer, June 2022.

Alexandre LeBel, Laura Wagner, Siddharth Jain, et al. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10:555, 2023. doi: 10.1038/s41597-023-02437-z. URL `https://doi.org/10.1038/s41597-023-02437-z`.

Jixing Li, Shohini Bhattasali, Shaolei Zhang, et al. Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, 9:530, 2022. doi: 10.1038/s41597-022-01625-7. URL `https://doi.org/10.1038/s41597-022-01625-7`.

Jan-Matthis Lueckmann, Alexander Immer, Alex Bo-Yuan Chen, Peter H Li, Mariela D Petkova, Nirmala A Iyer, Luuk Willem Hesselink, Aparna Dev, Gudrun Ihrke, Woohyun Park, et al. Zapbench: A benchmark for whole-brain activity prediction in zebrafish. *arXiv preprint arXiv:2503.02618*, 2025.

Jason Manley, Sihao Lu, Kevin Barber, Jeffrey Demas, Hyewon Kim, David Meyer, Francisca Martínez Traub, and Alipasha Vaziri. Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. *Neuron*, 112 (10):1694–1709.e5, 2024. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2024.02.011. URL `https://www.sciencedirect.com/science/article/pii/S0896627324001211`.

Samuel A. Nastase, Ariel Goldstein, and Uri Hasson. Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 222:117254, 2020. doi: 10.1016/j.neuroimage.2020.117254. URL `https://doi.org/10.1016/j.neuroimage.2020.117254`. Open access under CC license.

Samuel A. Nastase, Yung-Fang Liu, Harrison Hillman, et al. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8:250, 2021. doi: 10.1038/s41597-021-01033-3. URL `https://doi.org/10.1038/s41597-021-01033-3`.

Petr Nejedly, Vaclav Kremen, Vladimir Sladky, Jan Cimbalnik, Petr Klimes, Filip Plesinger, Filip Mivalt, Vojtech Travnicek, Ivo Viscor, Martin Pail, et al. Multicenter intracranial eeg dataset for classification of graphoelements and artifactual signals. *Scientific data*, 7(1):179, 2020.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Liam Paninski and John P. Cunningham. Neural data science: Accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current Opinion in Neurobiology*, 50:232–241, 2018. doi: 10.1016/j.conb.2018.04.007. URL `https://doi.org/10.1016/j.conb.2018.04.007`. Copyright © 2018 Elsevier Ltd. All rights reserved.

Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, 21(4):474–483, 2018. doi: 10.1038/s41593-018-0108-2. URL https://doi.org/10.1038/s41593-018-0108-2.

Felix Pei, Joel Ye, David M. Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. In *Advances in Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2109.04463.

Matthew G. Perich, Lee E. Miller, Mehdi Azabou, and Eva L. Dyer. Long-term recordings of motor and premotor cortical spiking activity during reaching in monkeys. Data set, 2025. URL https://doi.org/10.48324/dandi.000688/0.250122.1735.

Steven M Peterson, Zoe Steine-Hanson, Nathan Davis, Rajesh PN Rao, and Bingni W Brunton. Generalized neural decoders for transfer learning across participants and recording modalities. *Journal of Neural Engineering*, 18(2):026014, 2021.

Steven M Peterson, Satpreet H Singh, Benjamin Dichter, Michael Scheid, Rajesh PN Rao, and Bingni W Brunton. Ajile12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific data*, 9(1):184, 2022.

Matthias Schurz, Joaquim Radua, Markus Aichhorn, Fabio Richlan, and Josef Perner. Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, 42:9–34, 2014.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings, February 2023.

Christopher Wang, Adam Uri Yaari, Aaditya K Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R. Madsen, Scellig Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli, 2024. URL https://arxiv.org/abs/2411.08343.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9(11):e112575, November 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112575. URL http://dx.plos.org/10.1371/journal.pone.0112575.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural Data Transformer 2: Multi-context Pretraining for Neural Spiking Activity, September 2023.

Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. Brainwave: A brain signal foundation model for clinical applications. *arXiv preprint arXiv:2402.10251*, 2024.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation Model for Intracranial Neural Signal. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Yizi Zhang, Yanchen Wang, Donato Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards, Renee Tung, Olivier Winter, Eva Dyer, Liam Paninski, et al. Towards a" universal translator" for neural dynamics at single-cell, single-spike resolution. *Advances in Neural Information Processing Systems*, 37:80495–80521, 2024a.

Yizi Zhang, Yanchen Wang, Donato Jimenez-Beneto, Zixuan Wang, Mehdi Azabou, Blake Richards, Olivier Winter, International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution, July 2024b.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

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: Yes, we outline the types insights that can be derived from our benchmark and then show preliminary neuroscience results that take steps to producing those insights in Section 4.

   Guidelines:

   - The answer NA 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 NA 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: Yes, we discuss this in the conclusion.

   Guidelines:

   - The answer NA 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: [NA]

Justification: We present a benchmark that only pertains to empirical results.

Guidelines:

- The answer NA 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: We release the code on github with a quickstart notebook as well as the scripts that produce all results and figures. The appendix contains specification of trials used in splits (see Appendix F).

Guidelines:

- The answer NA 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: [Yes]

Justification: Same as above. See item 4. We release our code on github and include a quickstart jupyter notebook as well as scripts to obtain our results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Hyperparameters are given in Appendix H and splits are specified in Appendix F

Guidelines:

- The answer NA 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]

For our empirical results, we report standard error across cross-val folds.

Guidelines:

- The answer NA 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]

   We discuss this in Appendix J.

   Guidelines:

   - The answer NA 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]

   We adhere to the code of ethics.

   Guidelines:

   - The answer NA 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]

    We discuss this in the Conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We're using a public dataset only for evaluation purposes.

    Guidelines:

    - The answer NA 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: In the approach section we specify BrainTreebank's license.

    Guidelines:

    - The answer NA 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 are an evaluation-only benchmark. We make the code necessary for our benchmark public.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: We use a previously existing public dataset.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: We use a public dataset that is openly published and available on the internet to construct our benchmark (BrainTreebank, https://braintreebank.dev). As such, we did not require any IRB approvals or equivalent to conduct our research.

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We do not use LLMs as core components of our methods. One of our tasks is "GPT2 Surprisal", tasking the model with decoding the LLM negative log likelihood of the words in the dataset, however this feature was extracted from the sentences following standard protocol.

    Guidelines:

    - The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

TODOs

- TODO: chris, geeling make an appendix with all the hyperparameters for PopT
- TODO: bennet write richer website description in appendix. Basically write up what will be displayed on the page. Put a new screenshot in.
- DONE: chris Put parcellation figure in appendix
- DONE: chris Put time series superposition figure in appendix
- DONE: chris Put time course for all features in appendix
- DONE: chris Make an appendix that has compute details of PopT
- TODO: andrii Make appendix H in tabular form.
- TODO: andrii (only if you have time; low priority) make a figure in the appendix for population level decoding over time.
- TODO: chris / andrii Fix table 2 to have corresponding info to the data.

A Decoding tasks

#	Feature	Description	Benchmark Task
1	frame_brightness *(visual)*	The mean brightness computed as the average HSV value over all pixels	Binary classification: low (percentiles 0%-25%) vs high (75%-100%)
2	global_flow *(visual)*	A camera motion proxy. The maximal average dense optical flow vector magnitude	Same as above
3	local_flow *(visual)*	A large displacement proxy. The maximal optical flow vector magnitude	Same as above
4	global_flow_angle *(visual)*	As 2, averaged over orientation (degrees) and selected by maximal magnitude	2-way classification: Left vs Right (180 degree intervals)
5	local_flow_angle *(visual)*	The orientation (degrees) of the largest local flow vector	Same as above
6	face_num *(visual)*	The maximum number of faces per frame during the word	2-way classification: 0, or ≥ 1
7	volume *(auditory)*	Average root mean squared watts of the audio	Binary classification: low (0%-25%) vs high (75%-100%)
8	pitch *(auditory)*	Average pitch of the audio	Same as above
9	delta_volume *(auditory)*	The difference in average RMS of the 500ms windows pre- and post-word onset	Same as above
10	delta_pitch *(auditory)*	The difference in average pitch of the 500ms windows pre- and post-word onset	Same as above
11	speech *(language)*	Whether any speech is present in the given time interval	Binary classification
12	onset *(language)*	Whether a new sentence starts in the interval, or there is no speech at all	Binary classification
13	gpt2_surprisal *(language)*	Negative-log transformed GPT-2 word probability (given preceding 20s of language context)	Binary classification: low (0%-25%) vs high (75%-100%)
14	word_length *(language)*	Word length (ms)	Same as above
15	word_gap *(language)*	Difference between previous word offset and current word onset (ms)	Same as above
16	word_index *(language)*	The word index in its context sentence	2-way classification: 0 (the first word in the sentence), or other (1)
17	word_head_pos *(language)*	The relative position (left/right) of the word's dependency tree head	Binary classification
18	word_part_speech *(language)*	The word Universal Part-of-Speech (UPOS) tag	2-way classification: verb (0), or other (1)
19	speaker *(multimodal)*	The movie character that speaks the given word.	2-way classification: most frequent speaker (0), or other (1)

Table 1: **Extracted visual, auditory, and language features used to create the evaluations for Neuroprobe.** For all classification tasks, the classes were rebalanced. The difference between local and global flow is that global is the averaged optical flow, with the average being taken over all optical flow vectors on the screen, whereas local is the largest individual optical flow vector on the screen. The table is adapted from Chau et al. (2024).

B Subject and movie information

Subj.	Age (yrs.)	# Electrodes	Movie	Recording time (hrs)	Neuroprobe-Lite
1	19	154	Fantastic Mr. Fox	1.35	
			The Martian	2.43	x
			Thor: Ragnarok	1.77	x
2	12	162	Venom	1.54	x
			Spider-Man: Homecoming	2.05	
			Guardians of the Galaxy	1.90	
			Guardians of the Galaxy 2	2.13	x
			Avengers: Infinity War	2.30	
			Black Panther	1.42	
			Aquaman	2.19	
3	18	134	Cars 2	1.64	x
			Lord of the Rings 1	2.25	x
			Lord of the Rings 2 (extended edition)	3.58	
4	12	188	Shrek 3	1.38	x
			Megamind	1.44	x
			Incredibles	0.85	
5	6	156	Fantastic Mr. Fox	1.35	
6	9	164	Megamind	0.68	
			Toy Story	1.29	
			Coraline	0.84	
7	11	246	Cars 2	1.64	x
			Megamind	1.44	x
8	4.5	162	Sesame Street Episode	0.94	
9	16	106	Ant Man	1.80	
10	12	216	Cars 2	1.33	x
			Spider-Man: Far from Home	1.93	x

Table 2: **Subject statistics** Subjects in the BrainTreebank dataset, and the trials used in the benchmark tasks. Table adapted from Wang et al. (2023). The second column shows the total number of electrodes. The average amount of recording data per subject is 4.3 (hrs).

Subj.	Age	Sex	Movies	Time (h)	# Sent.	# Words	# Lemmas	# Elec.	# Probes
1	19	M	7, 18, 19	5.6	4372	27424	4489	154	13
2	12	M	2, 3, 4, 8, 9, 17, 21	13.5	9870	57731	9164	162	47
3	18	F	5, 11, 12	7.5	5281	31596	4547	134	12
4	12	F	10, 13, 15	3.7	4056	23876	4017	188	15
5	6	M	7	1.35	1282	7908	1481	156	12
6	9	F	6, 13, 20	2.8	3789	20089	3349	164	12
7	11	F	5, 13	3.08	3523	19068	2828	246	18
8	4	M	14	0.94	860	3994	537	162	13
9	16	F	1	1.80	1558	9235	1480	106	12
10	12	M	5, 16	3.08	3981	22147	3004	216	17

Table 3: **All subjects language, electrodes and personal statistics.** Columns from left to right are the subject's ID and information (age and gender), the IDs of the movies they watched (corresponding to Supplementary Table 4), the cumulative movie time (hours), number of sentences, number of words (tokens) and number of unique lemmas (canonical word forms), as well as the number of probes the subject had and their corresponding number of electrodes. Table adapted from Wang et al. (2024).

#	Movie	Year	Length	Sent.	Words	Unique words	Nouns	Unique nouns	Verbs	Unique verbs
1	Antman	2015	7027	1558	9869	1944	1358	705	1545	580
2	Aquaman	2018	8601	1054	7233	1544	1069	520	1104	508
3	Avengers: Infinity War	2018	8961	1523	8529	1750	1083	607	1317	495
4	Black Panther	2018	8073	1254	7580	1606	1093	553	1209	508
5	Cars 2	2011	6377	2051	11407	2037	1572	724	1664	577
6	Coraline	2009	6036	997	5433	1232	784	409	805	348
7	Fantastic Mr. Fox	2009	5205	1282	8461	1864	1229	681	1227	484
8	Guardians of the Galaxy 1	2014	7251	1174	8295	1779	1096	603	1250	529
9	Guardians of the Galaxy 2	2017	8146	1290	9405	1824	1224	626	1370	532
10	Incredibles	2003	6926	1521	9430	1954	1226	652	1557	591
11	Lord of the Rings 1	2001	13699	1514	10566	1998	1473	679	1487	598
12	Lord of the Rings 2	2002	14131	1716	11041	2065	1588	743	1619	646
13	Megamind	2010	5735	1472	8891	1726	1172	602	1347	496
14	Sesame Street Ep. 3990	2016	3440	860	4220	787	717	231	706	217
15	Shrek the Third	2007	5568	1063	7226	1590	977	568	1071	422
16	Spiderman: Far From Home	2019	7764	1930	12189	1969	1459	668	1785	560
17	Spiderman: Homecoming	2017	8008	2196	12295	2066	1583	777	1808	572
18	The Martian	2015	9081	1570	11374	2192	1757	812	1677	622
19	Thor: Ragnarok	2017	7831	1583	9683	1789	1195	599	1419	548
20	Toy Story 1	1995	4863	1320	7216	1510	1019	548	1027	395
21	Venom	2018	6727	1379	7937	1513	897	507	1217	433

Table 4: **Language statistics for all movies.** Columns from left to right are the movie's ID, name, year of production, length (seconds), number of sentences, number of words (tokens), number of unique words (types), number of nouns, number of unique nouns, number of verbs and number of unique verbs. Table adapted from Wang et al. (2024).

C Composition of movies by volume

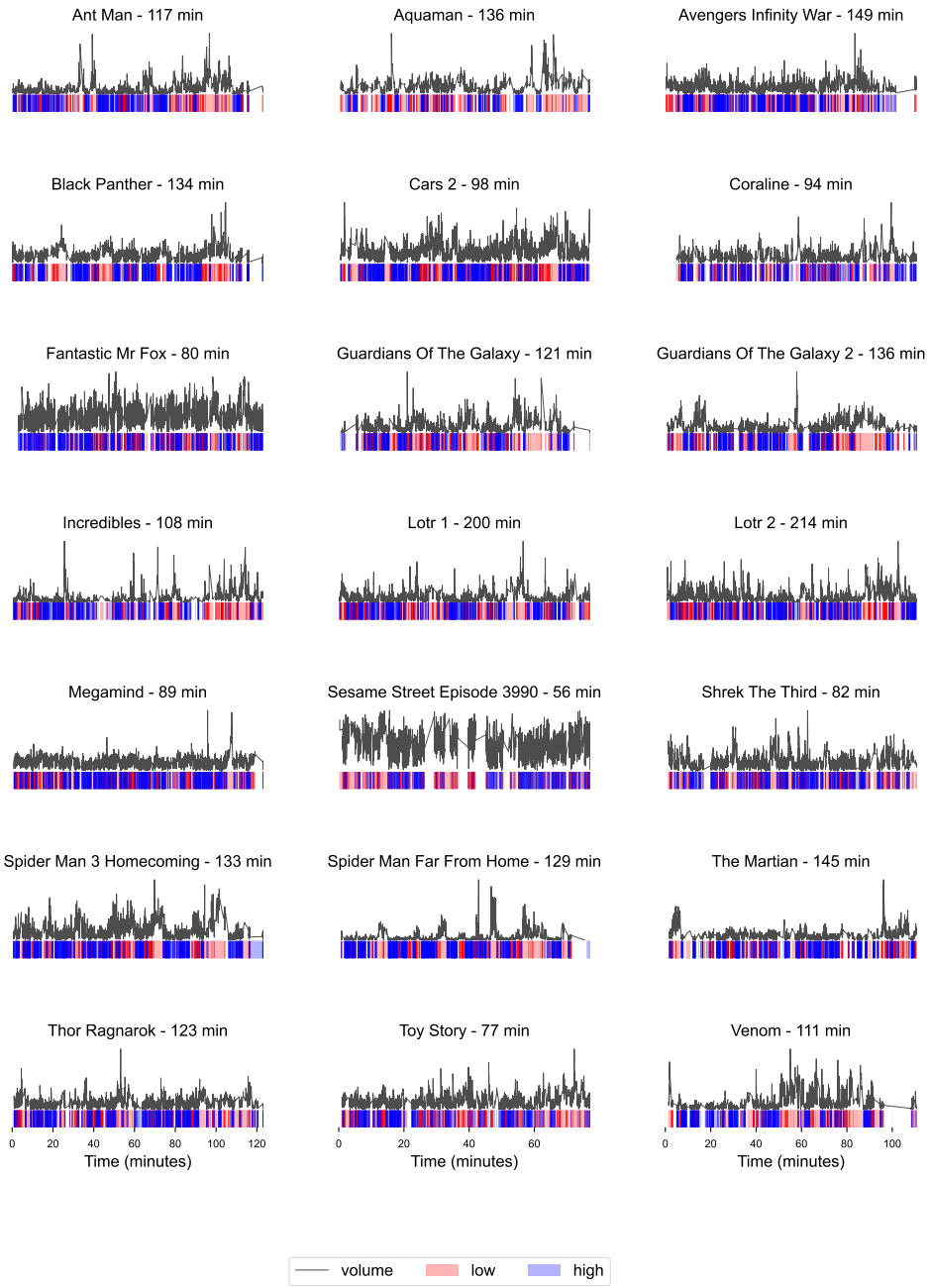


Figure 9: **Volume comparison across movies.** The black line shows the normalized audio volume over time for 18 feature-length films and one TV episode shown to subjects. Below each volume trace, colored bars indicate periods of relatively low (red) and high (blue) volume, defined as the bottom 25% and top 25% of volume values respectively.

D Speech localization

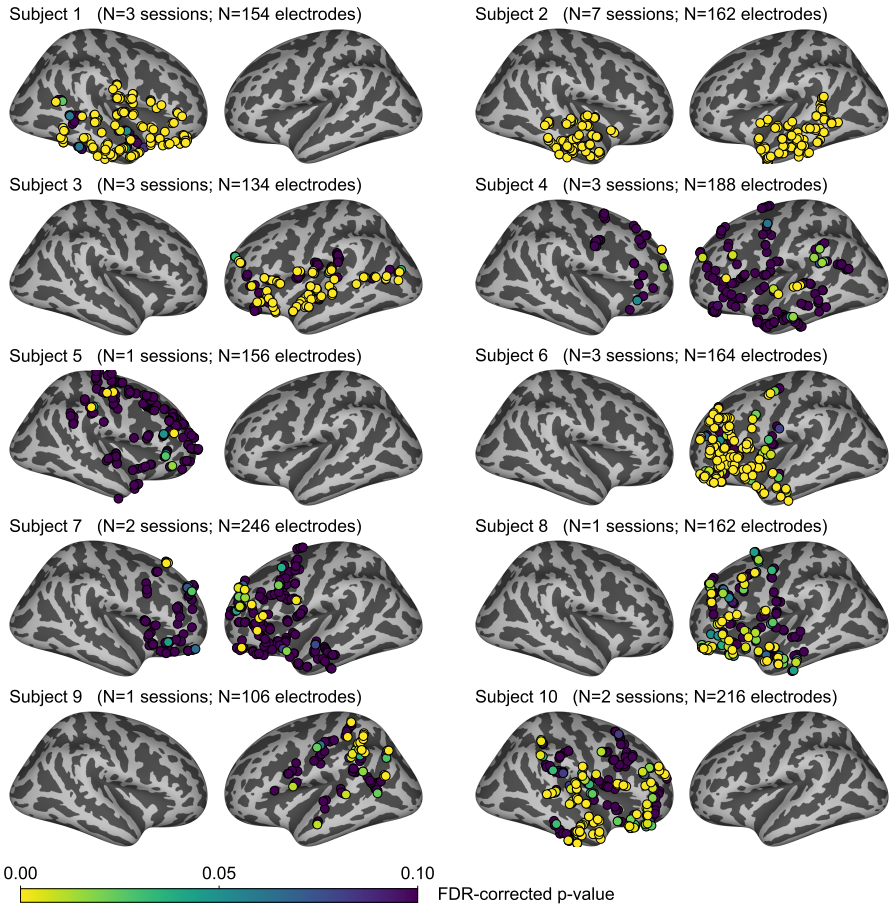


Figure 10: **Electrode locations and speech selectivity across subjects.** Brain reconstructions showing electrode placement and speech-selective responses for all 10 subjects. Each dot represents an electrode, colored by its FDR-corrected p-value from a speech vs. non-speech classification (color scale above, yellow indicating stronger selectivity). Left and right hemispheres are shown separately, with session counts and total electrodes noted. Speech selectivity was assessed by comparing high gamma power (70–300 Hz, dB) during the first 125 ms after word onset to non-speech intervals of equal duration. A two-sample t-test determined significance, with Benjamini-Hochberg correction applied for multiple comparisons.

E Face distribution

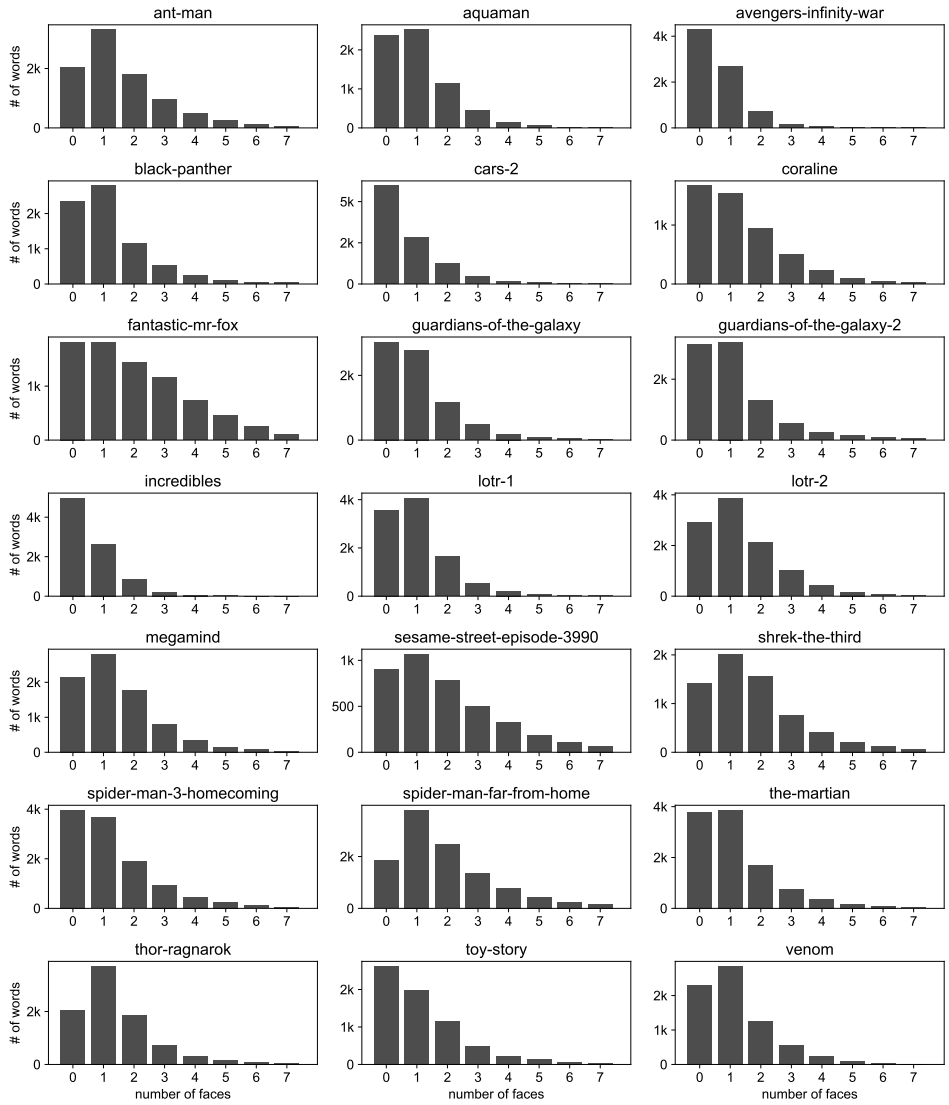


Figure 11: **Distribution of faces detected per frame across different movies.** Histograms show the number of words (y-axis) that occur during frames containing different numbers of faces (x-axis) for 18 feature-length films and one TV episode (Sesame Street)

F Splits

Neuroprobe includes 3 different types of splits.

Same subject/Same trial

Same subject/Different movie This is a slightly more difficult split. It ensures completely that no data-contamination due to auto-correlation has occurred.

Different subject/Different movie This is the most difficult split. It tests the model's ability to generalize between subjects *and* stimuli.

TODO: describe what the splits are and which trials are in each split.

G Neuroprobe-lite

The following subject-trial pairs are included in Neuroprobe Lite:

- Subject 1: Trials 1, 2
- Subject 2: Trials 0, 4
- Subject 3: Trials 0, 1
- Subject 4: Trials 0, 1
- Subject 7: Trials 0, 1
- Subject 10: Trials 0, 1

For every task, the number of datapoints was trimmed at 3500 datapoints (i.e. if a specific movie has more than 3500 annotations for any task, only the first 3500 are taken for the Lite benchmark). When selecting the subject/trial pairs for Neuroprobe Lite, we selected the trials that contained the most tasks which hit the 3500 datapoints limit.

H Models benchmarked

Linear For this evaluation, raw voltage traces sampled at 2048 Hz were taken from the BrainTreebank data, then line noise was removed at 60 ± 5 Hz and the 4 harmonics, and the resulting vectors of sampled features were fed as input to the linear regression. We found almost identical results when removing line noise or passing the data raw to the linear regression.

Linear (STFT) For this baseline evaluation, the features are the STFT of the raw signal with the following parameters (given that the sampling rate is 2048Hz):

- nperseg=256
- noverlap=0
- window=boxcar

After this step, the data turns into an array of arrays where first dimension is the time bin and the second dimension is the STFT result (a complex number); for the downstream regression, all of these features are concatenated together, with the real and imaginary parts of the complex features being split into two features each.

Linear (spectrogram) For this baseline evaluation, first the STFT of the raw voltage signal was taken as in the Linear (STFT) description, and then the absolute value of each complex number was taken to obtain the final real number features for each example.

BrainBERT For this evaluation, the BrainTreebank data was Laplacian rereferenced (as described in the original BrainBERT paper by Wang et al. (2023)), with line noise removed, and then passed into the BrainBERT model as provided by Wang et al. (2023). The output features were concatenated and used as input to the linear regression. For the electrodes which could not be Laplacian rereferenced, non-rereferenced data was inputted into BrainBERT. The BrainBERT model was frozen and only the final linear regression layer was fine tuned, in order to compare the quality of features generated by the foundation model.

For all linear regression, we used the sklearn package, class LinearRegression, with the tolerance parameter set as 0.001. In all cases, the features were first normalized using the sklearn StandardScaler. We found that it helps with convergence and often produces higher regression values for the baselines.

PopulationTransformer Off-the-shelf Population Transformer (PopT) is a SSL pretrained model for encoding arbitrary ensembles of iEEG electrode data for general downstream decoding (Chau et al., 2024). The model consists of a transformer backbone that learns functional and spatial relationships between input channels whose temporal activity is encoded. We use the publicly available weights which were pretrained on data from 10 iEEG subjects, using 5s BrainBERT temporal embeddings from individual channels. For Population Transformer, we followed the implementation and used

the weights from (Chau et al., 2024). The fine-tuning protocol is taken to be directly the same as in the authors' original paper (including linear rate, number of epochs, a factor of 10 between learning rates of the linear output layer vs the transformer blocks, etc), but reduce the number of steps to $steps = 1000$. We finetune Population Transformer in two conditions: either by only finetuning the final linear output layer while keeping the rest of the model weights frozen (the "frozen" condition), or finetuning through the whole model (the default PopT condition).

I Benchmark results

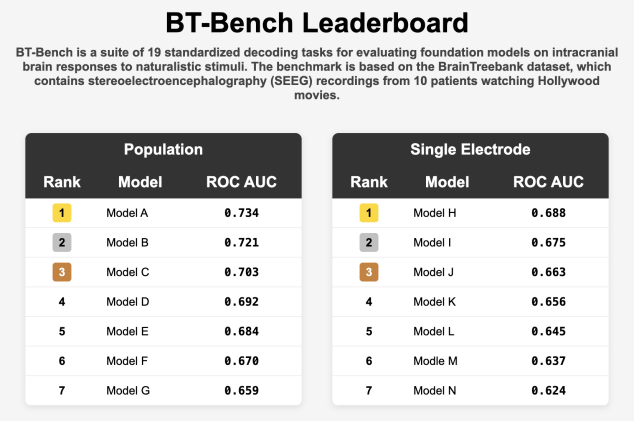
TODO fill in with tabular form of fig. 8.

J Compute requirements

Every Linear regression was run on a CPU-only instance, with 2 virtual CPU cores and 64GB RAM for the population level results and 2 CPU cores with 6GB RAM for the single electrode decoding results. For BrainBERT, the necessary resources also included a GPU with at least 9GB of memory along with 128GB of RAM and 2 CPU cores. For the PopulationTransformer, the fine-tuning was done on 2 GPUs (NVIDIA GeForce GTX TITAN X) with at least 12GB of GPU RAM.

K Leaderboard

TODO: describe leaderboard website



BT-Bench Leaderboard

BT-Bench is a suite of 19 standardized decoding tasks for evaluating foundation models on intracranial brain responses to naturalistic stimuli. The benchmark is based on the BrainTreebank dataset, which contains stereoelectroencephalography (SEEG) recordings from 10 patients watching Hollywood movies.

Population		
Rank	Model	ROC AUC
1	Model A	0.734
2	Model B	0.721
3	Model C	0.703
4	Model D	0.692
5	Model E	0.684
6	Model F	0.670
7	Model G	0.659

Single Electrode		
Rank	Model	ROC AUC
1	Model H	0.688
2	Model I	0.675
3	Model J	0.663
4	Model K	0.656
5	Model L	0.645
6	Modle M	0.637
7	Model N	0.624

Figure 12: **The leaderboard for the task of classifying sentence onset.** The public webpage link will be made available upon publication. TODO: revisit caption

L Time course of task decodability

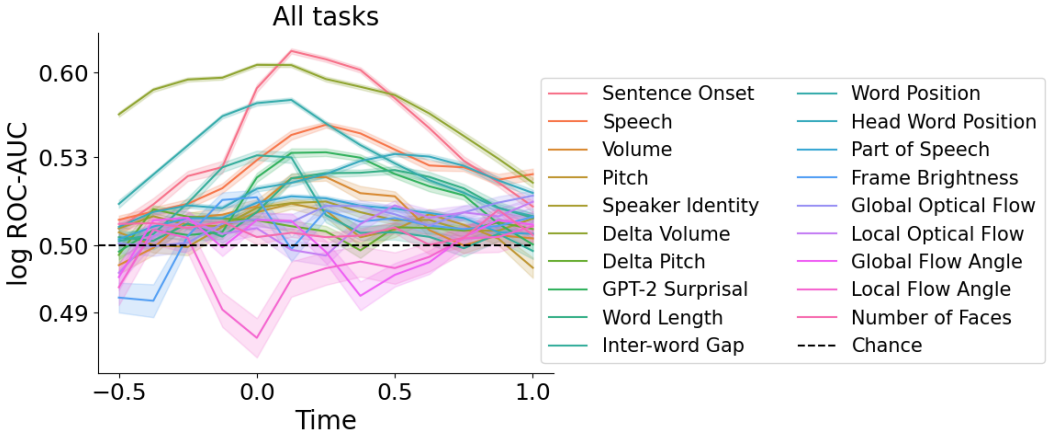


Figure 13: **TODO: revisit caption All the plots from Figure 4 overlaid. Error bars show standard error from variability across all electrodes (from all subjects and all sessions).**

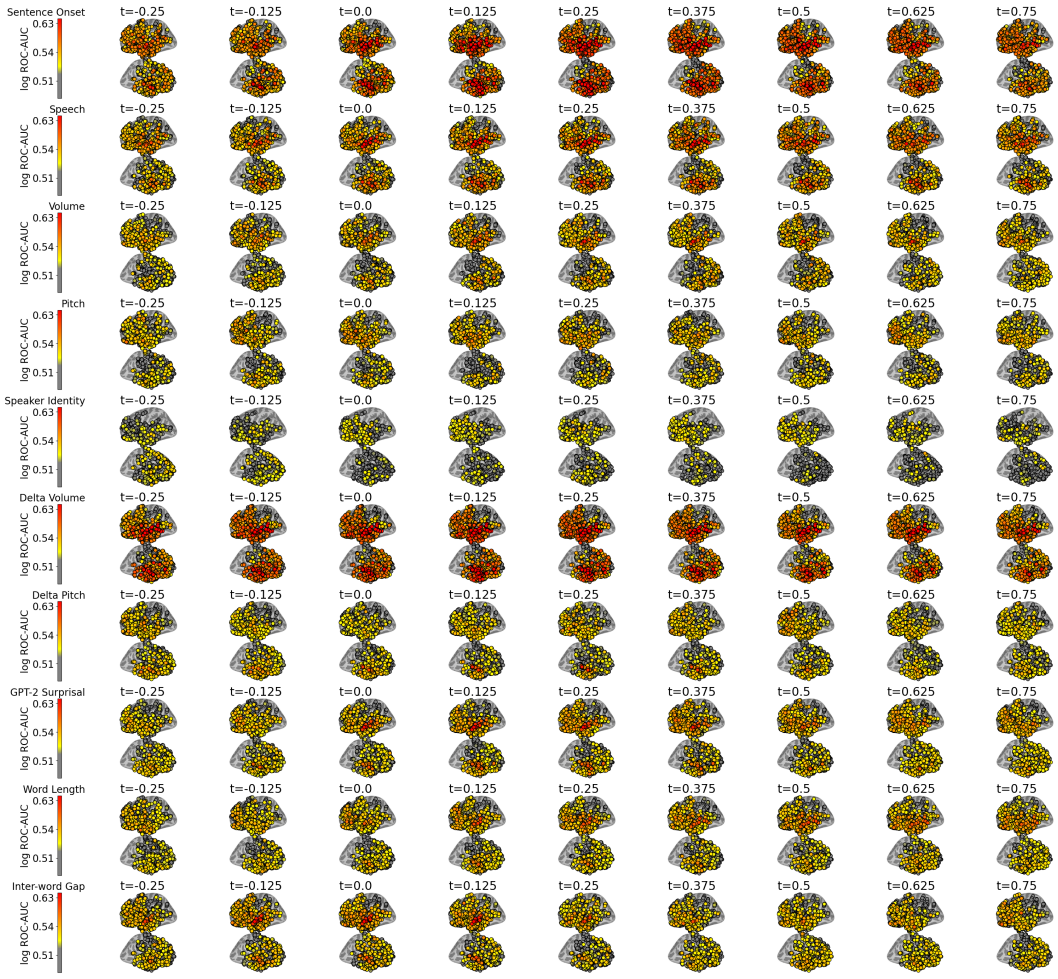


Figure 14: **TODO: revisit caption Same as Figure 7 but for all features. Pt 1**

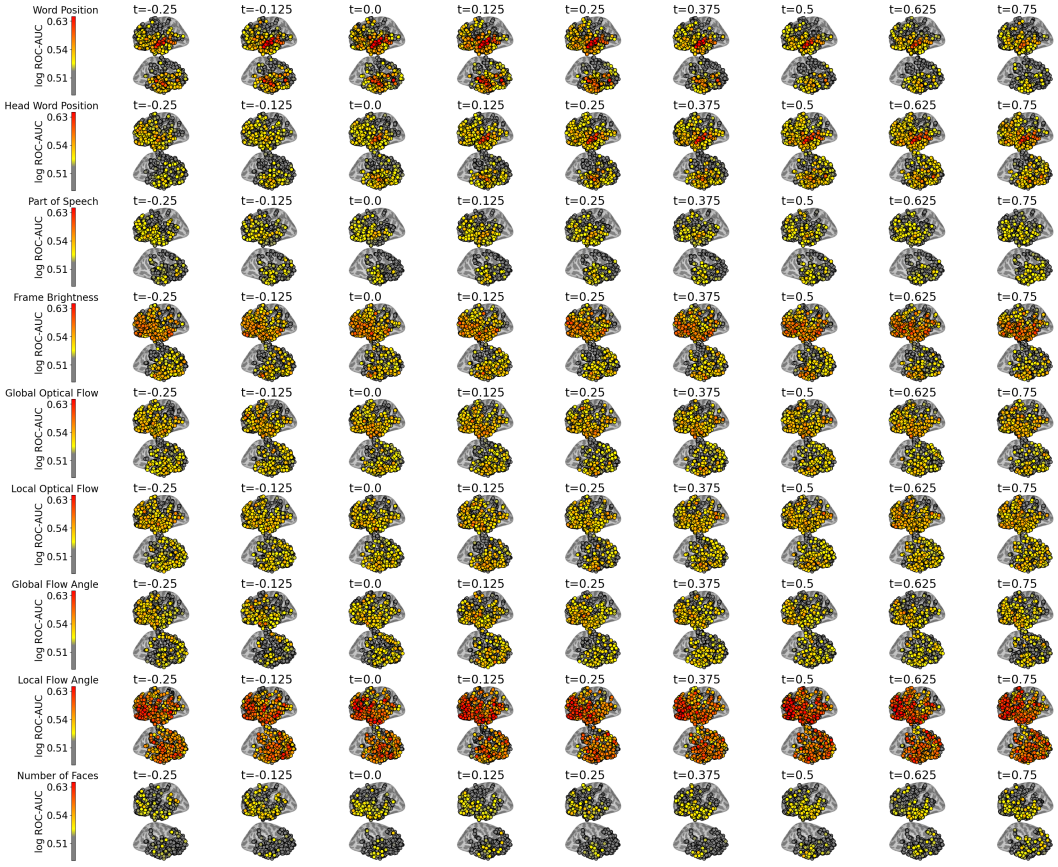


Figure 15: **TODO: revisit caption Same as Figure 7 but for all features. Pt 2**

M Region analysis

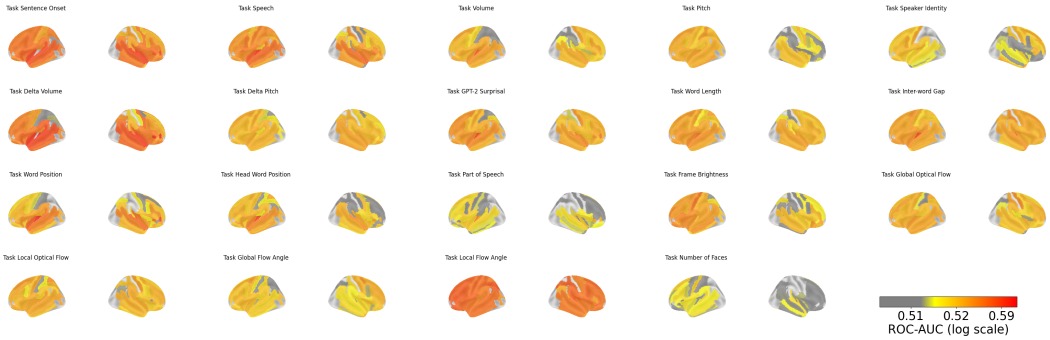


Figure 16: **TODO: revisit caption** top 10-th percentile of electrodes in each region are plotted Make it top k=100?.