LibriBrain: Over 50 Hours of Within-Subject MEG to Improve Speech Decoding Methods at Scale

Miran Özdogan[1] **Gilad Landau**[1] **Gereon Elvers**[1] **Dulhan Jayalath**[1]

Pratik Somaiya[1] **Francesco Mantegna**[1,2] **Mark Woolrich**[2] **Oiwi Parker Jones**[1,2]

[1]PNPL🍍, Department of Engineering Science, University of Oxford, UK
[2]Oxford Centre for Human Brain Activity (OHBA), University of Oxford, UK

`{miran, oiwi}@robots.ox.ac.uk`

Abstract

LibriBrain represents the largest single-subject MEG dataset to date for speech decoding, with over 50 hours of recordings—5× larger than the next comparable dataset and 50× larger than most. This unprecedented 'depth' of within-subject data enables exploration of neural representations at a scale previously unavailable with non-invasive methods. LibriBrain comprises high-quality MEG recordings together with detailed annotations from a single participant listening to naturalistic spoken English, covering nearly the full Sherlock Holmes canon. Designed to support advances in neural decoding, LibriBrain comes with a Python library for streamlined integration with deep learning frameworks, standard data splits for reproducibility, and baseline results for three foundational decoding tasks: speech detection, phoneme classification, and word classification. Baseline experiments demonstrate that increasing training data yields substantial improvements in decoding performance, highlighting the value of scaling up deep, within-subject datasets. By releasing this dataset, we aim to empower the research community to advance speech decoding methodologies and accelerate the development of safe, effective clinical brain-computer interfaces.

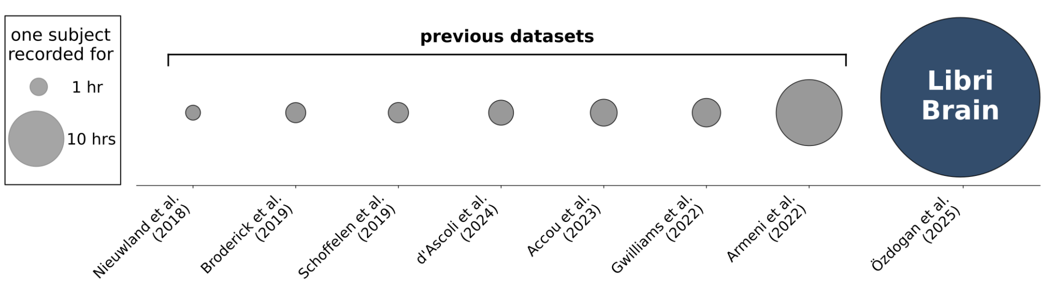


Figure 1: Comparison of within-subject data volume in non-invasive speech datasets.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

1 Introduction

Recent advances in deep learning have demonstrated the critical role of high-quality datasets in enabling robust and generalisable models — a principle that holds true not only in vision and language domains, but increasingly in the study of speech decoding from brain signals. In invasive brain-computer interface (BCI) research, large-scale datasets collected via electrocorticography (ECoG) or microelectrode arrays have made it possible to train models with dozens hours of data per subject, leading to remarkable performance in paralysed individuals (Moses et al., 2016; Metzger et al., 2023; Willett et al., 2023). Most notably, recent systems have achieved less than 5% word error rates (WER) for text decoding in paralysed patients (Card et al., 2024).

However, these achievements come with a fundamental limitation: they rely on data obtained through brain surgery. While neural interfaces based on ECoG or microelectrode arrays are extremely powerful, the need for invasive recording presents a substantial barrier to deployment outside of clinical trials. This has placed renewed emphasis on the search for non-invasive alternatives, particularly those using EEG or MEG, to achieve high-quality speech decoding without requiring surgical intervention.

Although full non-invasive brain-to-text systems have yet to achieve WERs below 100% (Jo et al., 2024; Yang et al., 2024b,c,a), a growing body of work has shown strong results on intermediate tasks such as speech segment identification (Défossez et al., 2023), word-level classification (d'Ascoli et al., 2024), and feature decoding (Jayalath et al., 2024, 2025b). Crucially, these advances have been enabled by the increasing availability of open EEG and MEG datasets, which have lowered the barrier to training powerful models across diverse tasks and conditions.

Early studies focused on pooling data across subjects within a single dataset (Défossez et al., 2023; d'Ascoli et al., 2024), but recent work has begun to explore more ambitious scaling via cross-dataset pre-training (Jayalath et al., 2024, 2025b) and domain adaptation (Ridge and Parker Jones, 2024). These findings establish a basis for future research aimed at leveraging extensive pooled non-invasive datasets to develop general-purpose decoding models. However, an important insight has recently emerged: although training on broader datasets with more subjects tends to improves generalisation, models trained on 'deep' data from a single subject often outperform broader models when matched for total training hours(d'Ascoli et al., 2024).

Most existing non-invasive datasets are 'broad but shallow', offering 1–2 hours per subject. The standout exception is the dataset by Armeni et al. (2022), with over 10 hours per subject, which has underpinned some of the strongest decoding results to date (d'Ascoli et al., 2024). This highlights an important gap in available resources, as non-invasive speech datasets with substantial per-subject depth remain limited yet seem crucial for progress toward reliable brain-to-text decoding.

In this paper, we introduce LibriBrain, a new open-access magnetoencephalography (MEG) dataset recorded from a single healthy volunteer listening to naturalistic, connected speech. With over 50 hours of data, LibriBrain offers the deepest non-invasive speech brain recordings to date, being more than 5× larger than the previous largest within-subject MEG dataset (Armeni et al., 2022), and 50× larger than most datasets in this space (see Figure 1). Inspired by the LibriSpeech dataset for automatic speech recognition (ASR), LibriBrain includes excerpts from seven audiobooks. All source audio is from LibriVox, ensuring public-domain availability and full reproducibility.

Speech decoding tasks can target overt, covert (inner), or heard speech. Of these, heard speech is the most tractable for large-scale non-invasive recording. It not only supports fundamental research in auditory representation and inner speech (e.g. as auditory imagery), but also aids in developing robust, transferable methods for speech BCIs. LibriBrain was designed to fill this methodological gap: enabling reproducible, scalable research on decoding speech from brain activity using naturalistic stimuli.

LibriBrain makes the following contributions:

- **Largest within-subject MEG speech dataset to date**: Over 50 hours of high-quality MEG data from a single participant, exceeding previous datasets by 5–50×.
- **Naturalistic, richly annotated stimuli**: Recordings obtained while listening to full-length audiobooks, including six Sherlock Holmes books and a subset of LibriSpeech excerpts.
- **Designed for ML usability**: Easily accessible via `pip`, with a simple-to-use Python API (`from pnpl.datasets import LibriBrain`) and standard train/val/test splits.

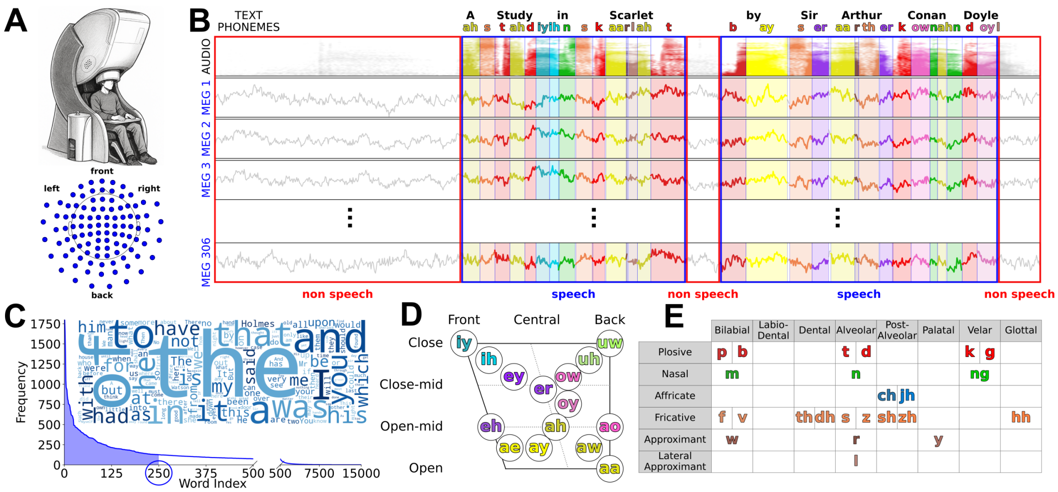


Figure 2: Overview of the LibriBrain dataset. (A) Illustration of the MEG scanner and sensor layout. (B) Example annotation of a sentence, showing phonemes and words aligned with the audio spectrogram and MEG time series. (C) Word frequency distribution and word cloud for the 250 most frequent words. (D) Vowel chart and (E) consonant chart depicting the linguistic properties of the phonemes.

- **Benchmark-ready**: Supports two initial benchmark tasks—speech detection and phoneme classification—and includes baseline models and code.
- **Community engagement**: Supports open ML competitions and a public leaderboard to encourage community-driven progress and reproducibility.
- **Open and reproducible**: All audio, transcripts, and metadata are shared under public-domain licenses, making the entire dataset freely usable and shareable.

2 Related Work

A growing number of EEG and MEG datasets focus on language processing and can be leveraged to scale non-invasive speech decoding models (Table 1). These datasets vary widely across multiple dimensions, including sensor count, linguistic richness, recording duration, and the number of participants.

In terms of sensor coverage, the largest are LibriBrain and the MEG datasets of d'Ascoli et al. (2024), all recorded using 306-channel Elekta/MEGIN systems—matching the highest-density sensor arrays available for non-invasive human neuroimaging.

When measured by total duration, the largest dataset is Nieuwland et al. (2018) with 171 hours of EEG. However, this dataset is very broad and shallow: its 334 participants each contributed on average only 30 minutes of data, limiting per-subject decoding depth. Similar trends appear in the MEG datasets of Schoffelen et al. (2019), which span 81 hours (listening) and 106 hours (reading) but include over 90 subjects, with just 0.8–1.1 hours recorded per individual. In contrast, the dataset of Armeni et al. (2022) is deeper, with approximately 10 hours per subject and 30–34 hours total, and has been shown to yield stronger decoding performance in follow-up studies (d'Ascoli et al., 2024).

LibriBrain extends this direction by offering over 50 hours of MEG from a single subject—approximately 5× deeper than the next closest MEG dataset. While Armeni et al. (2022) focused on one Sherlock Holmes book, LibriBrain spans nearly the entire Conan Doyle canon, with recordings from seven audiobooks. This depth allows for analysis of fine-grained speech representations at scale, and supports strong generalisation to unseen data in downstream decoding tasks (Section 4).

Unlike prior work, LibriBrain is designed with machine learning usability in mind. It is fully open, publicly available on Hugging Face, and comes with a Python API (`pnpl`) for seamless integration with deep learning frameworks such as PyTorch (Paszke et al., 2019). As shown in Table 1, most

Table 1: M/EEG datasets for language tasks.

Dataset	Task	Modality	Language	Sensors	Total hrs	# Subj	Hrs/Subj	Stimulus	Public	PNPL Dataloader
LibriBrain (ours)	Listen	MEG	English	**306**	52	1	**52.32**	LibriVox (Sherlock, Books 1–7)	Yes	Yes
Armeni et al. (2022)	Listen	MEG	English	269	30	3	10.0	LibriVox (Sherlock, Book 3)	Yes	Yes (staged for release)
Gwilliams et al. (2022)	Listen	MEG	English	208	49	27	1.0	MASC stories (synthetic voice)	Yes	Yes (staged for release)
Schoffelen et al. (2019)	Listen	MEG	English	275	81	96	0.8	Spoken sentences	Yes	Yes (staged for release)
d'Ascoli et al. (2024)	Listen	MEG	French	**306**	94	58	1.6	Le Petit Prince	No	Not yet
Brennan (2023)	Listen	EEG	English	61	10.1	49	0.2	Alice in Wonderland	Yes	Not yet
d'Ascoli et al. (2024)	Read	MEG	French	**306**	59	46	1.3	Le Petit Prince	No	Not yet
Schoffelen et al. (2019)	Read	MEG	Dutch	275	106	99	1.1	Narrative reading task	Yes	Not yet
Accou et al. (2024)	Read	EEG	English	64	150	85	1.8	Podcasts and audiobooks	Yes	Not yet
Nieuwland et al. (2018)	Read	EEG	English	22	**171**	**334**	0.5	Hemingway short story	Yes	Not yet
Broderick et al. (2018)	Read	EEG	English	128	20	19	1.1	Subset of Nieuwland et al. 2018	Yes	Not yet

recent datasets are public, but the raw MEG from d'Ascoli et al. (2024) remains unavailable at the time of writing. Furthermore, LibriBrain includes ready-to-use benchmarks and reference models for speech detection, phoneme classification, and word decoding—supported by public leaderboards as part of the PNPL competition series—to enable standardised evaluation and accelerate progress in non-invasive speech BCI research.

Despite a growing number of non-invasive datasets and rising interest in foundation models for brain signals, current decoding models rarely generalise across datasets Défossez et al. (2023); d'Ascoli et al. (2024). Historically, decoding has struggled to generalise across subjects, though individual subjects might benefit from group-level models Csaky et al. (2023). However, a number of recent works have successfully pooled data across subjects, though often not between datasets. For example, although d'Ascoli et al. (2024) analyse recordings from over 700 subjects, they report results separately for each dataset—underscoring a lack of cross-dataset generalisability. Recent work by Jayalath et al. (2024, 2025b) and Ridge and Parker Jones (2024) shows that unsupervised pretraining and domain adaptation can help, but high-quality, within-subject data remains essential. LibriBrain takes a complementary approach by scaling data longitudinally within a single participant. It provides the deepest single-subject MEG dataset to date and is designed to reduce friction in neural decoding research, with standardised splits, public loaders, and baseline models to support reproducibility and benchmark-driven progress.

3 The LibriBrain Dataset

3.1 Dataset Overview

The LibriBrain dataset consists of non-invasive magnetoencephalography (MEG) brain recordings, obtained from a single healthy volunteer while listening to over 50 hours of audiobook recordings. Paired event files contain time-locked annotations for linguistic events (e.g. speech, words, and phonemes). The final dimensions of the data are 306 sensor channels $\times$ T time samples, where each sample represents 4 milliseconds. At a high level of abstraction, the dataset is split into standard training, validation, and test sets to facilitate reproducible machine learning applications, with additional hidden sets reserved for competition evaluation (see e.g. Landau et al., 2025). At a lower level, the data are split into the experimental sessions in which they were collected, with whole sessions set aside as holdout data. Non-invasive MEG was collected at the Oxford Centre for Human Brain Activity (OHBA) using a state-of-the-art MEGIN Triux™ Neo system featuring 306 sensor channels that simultaneously probe magnetic fields across the entire head. The recordings were originally sampled at 1 kHz but were downsampled during preprocessing to 250 Hz to preserve oscillations into the high-gamma range (70–125 Hz).

3.2 Data Collection and Structure

MEG data were acquired for audiobook recordings of seven books in the canon of Sherlock Holmes (Table 6 in Supplementary Materials). Recordings were made over 95 sessions, with each session corresponding to a book chapter (or occasionally part of a chapter when chapters were split into multiple parts in the audiobook recordings). Sessions vary in duration, with the average session lasting approximately 34 minutes, with a standard deviation of 15 minutes, corresponding to audiobook recordings with an average of 5421.28 words (standard deviation 2485.19). Each recording session is paired with an event file (in CSV format) containing temporal information about linguistic events such as speech/non-speech, words, and phonemes; this information can be used for various encoding

Table 2: LibriBrain dataset statistics: word, phoneme, and audio duration counts across train, validation, and test splits. To reduce information leakage, the competition holdout data statistics are redacted here (see Landau et al., 2025).

Split	Words	Unique Words	Phonemes	Sessions	Minutes	Hours
Train	459,227	16,753	1,488,392	91	3,094	51.57
Validation	3,427	1,082	11,289	1	22	0.36
Test	3,576	1,145	12,051	1	23	0.38
Total	466,230	16,892	1,511,732	93	3,139	52.32

and decoding tasks, such as speech detection, phoneme classification, and word classification (see Section 4). Full experimental methods for data collection are in section B of the Appendix.

3.3 Data Format, Access, and Supporting Python Library

To lower the entry barrier for machine learning practitioners, we are initially releasing the data in serialised format (HDF5 files). Data are provided in a modular structure (one HDF5 file and TSV file per session), making it easy to load different amounts of data for training. We dedicate one session each for validation and testing (book 1 sessions 11 and 12, respectively). Two additional sessions are being held back for open machine learning competitions (Landau et al., 2025). The validation, test, and competition holdout sessions were acquired on a separate day from all training data, to provide a strong protection against information leakage via 'nonsense correlations' (Harris, 2021).

For this release, we applied minimal preprocessing to maximise accessibility for ML practitioners unfamiliar with neuroimaging pipelines. Briefly, raw MEG recordings were corrected for head movement and signals were filtered (e.g. to remove obvious noise) and then downsampled (see section B.4 for details). Data can be accessed directly using standard HDF5 libraries. However, the recommended way to interact with the data is through our Python library `pnpl`:

```python
#!pip install pnpl
from pnpl.datasets import LibriBrainSpeech, LibriBrainPhoneme
train_data = LibriBrainSpeech(path="./data", partition="train", download=True)
train_data = LibriBrainPhoneme(path="./data", partition="train")
```

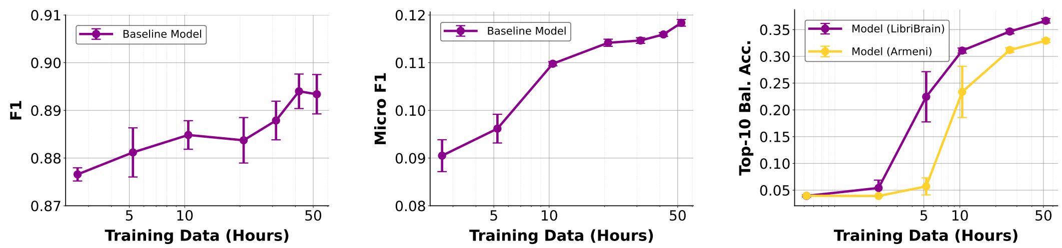
The initial release of the `pnpl` library supports two core tasks aligned with the 2025 PNPL competition: Speech Detection and Phoneme Classification. Nonetheless, the library design is modular, allowing for straightforward extensions to additional tasks derived from the accompanying events file. The source code of the `pnpl` library is available on GitHub.[1] The data is also available for direct download from Hugging Face.[2]

For a future release, we are preparing a BIDS-compliant (Niso et al., 2018) and pseudonymised version of the data in FIF format. This version of the data will take up significantly more memory than the serialised format, but include raw recordings for custom preprocessing. We plan to make it loadable through the `pnpl` library after the 2025 competition (Landau et al., 2025).

4 Decoding Experiments

To validate the LibriBrain dataset and establish strong baselines for future work, we evaluate three neural decoding tasks: speech detection, phoneme classification, and word classification. These span a spectrum of complexity, providing insights into the quality of the dataset and the scalability of neural speech decoding. All experiments are fully reproducible, using the standard train, validation,

[1] https://github.com/neural-processing-lab/frozen-pnpl
[2] https://huggingface.co/datasets/pnpl/LibriBrain



(a) Speech detection (b) Phoneme classification (c) Word classification

Figure 3: Impact of training data volume on model performance across tasks. Subsets were randomly drawn from the training split. Results for speech detection and phoneme classification are averaged over three random seeds, with error bars showing standard error. Word classification results, reproduced with permission from Jayalath et al. (2025a), use five random seeds.

and test splits defined in the `pnpl` data loader. Code to reproduce these analyses is provided on GitHub.[3]

For speech detection and phoneme classification, we employ a lightweight convolutional neural network architecture inspired by SEANet (Tagliasacchi et al., 2020), specifically adapted for MEG data input (Appendix E.1). To ensure meaningful baseline comparisons, we also evaluate performance against a random baseline that assigns labels according to their frequency in the training set. For the word classification task, we replicate the state-of-the-art method introduced by Jayalath et al. (2025a).

4.1 Task 1: Speech Detection

In the speech detection task, the objective is to accurately identify the time points within a segment of MEG recordings when participants were actively listening to speech.

We assess model performance using a number metrics: F1-Score, Balanced Accuracy, Area Under the Receiver Operating Characteristic curve (AUROC), Jaccard Index, and Cross Entropy Loss. We present the results in Table 3.

Additionally, we analyse the relationship between the amount of data used for training and model performance, observing that performance improves approximately logarithmically with increasing amounts of training data (Figure 3a).

Table 3: Speech detection performance. Mean and standard error of metrics over 10 seeds.

Metric	Model	Random Baseline
F1-score	**0.8989*** $\pm$ 0.0007	0.7848 $\pm$ 0.0017
Balanced Accuracy	**0.7082*** $\pm$ 0.0031	0.4990 $\pm$ 0.0028
AUROC	**0.8644*** $\pm$ 0.0017	0.5000 $\pm$ 0.0000
Jaccard Index	**0.6077*** $\pm$ 0.0027	0.3820 $\pm$ 0.0024
Cross Entropy Loss	**0.3802*** $\pm$ 0.0044	0.5449 $\pm$ 0.0000

* $p < 0.01$, one-sided exact permutation test using all 1,024 sign-flip permutations.

4.2 Task 2: Phoneme Classification

The goal of phoneme classification is to predict which of the 39 ARPAbet phonemes (Weide, 1998) corresponds to a given segment of MEG data. Each training sample consists of a short MEG window precisely aligned with a single phoneme from the stimulus audio. This task builds upon the extensive tradition of phoneme-based automatic speech recognition (ASR) (Garofolo et al., 1993), and continues to play a central role in contemporary invasive brain-to-text pipelines (Wilson et al., 2020; Metzger et al., 2023; Willett et al., 2023; Card et al., 2024).

[3] https://github.com/neural-processing-lab/libribrain-experiments

We evaluate model performance using F1-Score, Balanced Accuracy, AUROC, Jaccard Index, and Cross Entropy Loss, with complete results presented in Table 4. For F1 score and AUROC, we report both micro- and macro-averaged values. The results demonstrate that the data enables statistically significant improvements in decoding performance over the random baseline. The absence of improvement in macro F1-score can be attributed to the phoneme frequency distribution, which roughly follows a power-law pattern; consequently, our model does not predict the less frequent phonemes, leading to penalties under this metric. In Landau et al. (2025), we employ a class-weighted loss function, resulting in statistically significant improvements in macro F1-score over the random baseline. For a more comprehensive analysis, refer to Appendix E.

Additionally, we investigate scaling behaviour, finding that decoding accuracy improves approximately logarithmically with additional training data (Figure 3b), consistent with scaling laws in other ML domains (Kaplan et al., 2020; Zhai et al., 2021).

Finally, we investigate the effects of averaging multiple instances of the same phoneme before classification. This procedure simulates brain-computer interface scenarios, where repeated user inputs are leveraged to enhance decoding reliability. We observe that accuracy increases consistently with the number of averaged tokens, achieving a balanced accuracy exceeding 60% with 100 repetitions (Figure 8).

Table 4: Phoneme classification performance. Mean and standard error of metrics over 10 seeds.

Metric	Model	Random Baseline
Micro F1-score	**0.1168*** $\pm$ 0.0003	0.0442 $\pm$ 0.0003
Macro F1-score	0.0253 $\pm$ 0.0005	**0.0258** $\pm$ 0.0003
Balanced Accuracy	**0.0399*** $\pm$ 0.0003	0.0259 $\pm$ 0.0003
Micro AUROC	**0.6360*** $\pm$ 0.0010	0.5000 $\pm$ 0.0000
Macro AUROC	**0.6527*** $\pm$ 0.0013	0.5000 $\pm$ 0.0000
Cross Entropy Loss	**3.2332*** $\pm$ 0.0026	3.3509 $\pm$ 0.0000

* $p < 0.01$, one-sided exact permutation test using all 1,024 sign-flip permutations.

4.3 Task 3: Word Classification

In word classification, we segment a window of MEG data aligned to the onset of a word in the audio. Much like phoneme classification, the task is to classify the word that the neural response is associated with rather than the phoneme. We follow recent work in limiting classification to the set of 250 highest frequency words (d'Ascoli et al., 2024; Jayalath et al., 2025a).

Word classification is a task with precedent in MEG. In a recent preprint, d'Ascoli et al. (2024) developed a method in which they encode a set of consecutive MEG windows, aligned to word onsets, with a transformer and simultaneously predict target word embeddings for these windows. The target word embeddings are extracted from the middle layer of a T5 large language model (Raffel et al., 2020) and the word classifier is optimised for these targets using a variant of the SigLIP contrastive loss (Zhai et al., 2023).

Jayalath et al. (2025a) extend this work further to full sentence reconstruction using LLM-based rescoring and predictive in-filling of out-of-vocabulary words. They find that LibriBrain is the best performing dataset among the existing large speech datasets. Table 5 shows their word classification results in which they compare it to the next largest within-subject MEG dataset (Armeni et al., 2022). Figure 3 shows that LibriBrain scales better than the Armeni et al. (2022) dataset at like for like volumes of training data and continues to scale beyond it. This shows that word classification models trained on LibriBrain perform even better than the previous best-in-class dataset. We refer the reader to Appendix F and especially Jayalath et al. (2025a) for details.

Table 5: Word classification performance. Uncertainty is standard error. Results are reproduced with permission from Jayalath et al. (2025a). They use a vocabulary consisting of the top 250 most frequent words in each dataset. This is enough to cover 67.9% of the story text in LibriBrain. Balanced accuracy represents the macro average over the individual accuracies of each of the words in the vocabulary. The random chance baseline is calculated as $10/250 = 0.04$.

Metric	**Random**	Armeni et al. (2022)	**LibriBrain (ours)**
Top-10 Balanced Accuracy	0.0400 ± 0.0000	$0.3261^{*} \pm 0.0019$	$\mathbf{0.3621}^{*\dagger} \pm 0.0031$

* $p < 0.01$, two-sided permutation test using all sign-flip permutations and † two-sample t-test against Armeni.

5 Discussion

The release of LibriBrain represents a significant milestone in non-invasive speech decoding research, providing, to date, the deepest within-subject MEG dataset — 5× the next biggest (Armeni et al., 2022) and 25–50× most datasets. Unlike previous datasets that emphasise breadth across subjects, LibriBrain's depth within a single subject enables exploration of decoding limits when individual variability is controlled. This approach aligns with observations that deep, single-subject data often yields better performance than broader but shallower multi-subject datasets when total data hours are matched (see e.g. d'Ascoli et al., 2024). The scaling properties observed in our initial experiments align with patterns seen across machine learning domains: performance improves approximately logarithmically with increasing data volume. The effectiveness of averaging multiple instances of the same phoneme is particularly encouraging, as it suggests practical pathways toward useful non-invasive speech BCIs even with current technology.

Our decision to standardise data splits and establish benchmark tasks addresses a critical gap in the field. The absence of widely accepted standards has made it difficult to compare methods across studies and track progress systematically. By providing not only the data but also baseline implementations, benchmarks, as well as a Python library for easy access, we hope to lower the barrier to entry and encourage broader participation from the machine learning community. We anticipate that LibriBrain will serve as a foundational resource, facilitating the continued expansion of non-invasive neural decoding research. Its modular design allows straightforward addition of new subjects, modalities, or tasks in future releases. Its integration with the `pnpl` library provides a unified interface for accessing multiple datasets, facilitating cross-dataset analysis and transfer learning approaches.

Our baseline experiments validate that deep neural networks achieve statistically significant decoding performance on this dataset. Additionally, we observe a clear relationship between the amount of training data and decoding accuracy, characterised by an approximately logarithmic scaling pattern similar to those seen in other machine learning domains (Kaplan et al., 2020; Zhai et al., 2021). The particular effectiveness of averaging multiple phoneme instances suggests practical pathways toward useful non-invasive speech BCIs.

5.1 Broader Impact

Beyond its immediate scientific contributions, LibriBrain has potential for broader impact across several domains:

Clinical applications. While current non-invasive BCIs lag behind invasive approaches in performance, the scale and quality of LibriBrain could accelerate progress toward clinically viable non-invasive alternatives. These would come with the significant advantages of accessibility and risk reduction.

Methodological transfer. Techniques developed for LibriBrain may transfer to other neuroscience domains and recording modalities. Successful methods might transfer to clinical EEG (Jayalath et al., 2024)—which is more widely available than MEG—potentially broadening impact to resource-limited settings.

Community building. By establishing standardised benchmarks, public leaderboards, and providing open infrastructure, LibriBrain aims to foster a more cohesive research community around non-invasive neural decoding that may help consolidate methodological advances that are currently scattered across different research groups, accelerating overall progress in the field.

5.2 Limitations

Despite its significant scale and careful design, LibriBrain has several important limitations that should be considered when interpreting results or designing follow-up studies:

Single-subject design. The deep, within-subject approach limits generalisability across subjects. It was chosen due to previous findings that decoding improves fastest when scaling data within individuals. Future extensions will incorporate multi-subject data.

Listening paradigm focus. The dataset consists of MEG recordings during naturalistic listening to audiobooks, rather than imagined or overt speech. This design choice avoids muscle artefacts that would contaminate MEG signals and it enables collection of extensive high-quality data. Evidence suggests substantial overlap in neural circuits between speech perception and production (Hickok and Poeppel, 2007; Pulvermüller et al., 2006), and recent work suggests that models trained on listening data can, to some extent, transfer to covert speech tasks when decoding fMRI (Tang et al., 2023).

Focused language content. The dataset includes over 50 hours of recordings predominantly from a single author (Arthur Conan Doyle) and genre (detective fiction), read by a single narrator. While this approach may limit the diversity of vocabulary, syntactic structures, and prosodic patterns compared to multi-author or multi-narrator datasets, it establishes a controlled foundation for understanding speech decoding fundamentals. This is meant to be analogous to a single-speaker ASR dataset, to minimise phonetic variance. Future releases will strategically introduce greater linguistic diversity to support broader language understanding and semantic decoding capabilities.

Preprocessing approach. We applied minimal preprocessing (head motion correction, filtering, and downsampling to 250 Hz) and released data in a serialised format to maximise accessibility for machine learning researchers who may be unfamiliar with neuroimaging data, our aim being to make brain data as accessible as computer vision data in standard datasets like CIFAR-10 (Krizhevsky, 2009). Some researchers may prefer access to raw recordings; to address this, we plan to release the raw BIDS-formatted data after the completion of the PNPL competition (Landau et al., 2025).

Absence of Brain-to-Text Decoding. Some readers may question our focus on supervised prediction tasks rather than sequence-to-sequence brain-to-text (B2T) decoding. However, with the exception of Tang et al. (2023) and Jayalath et al. (2025a), current non-invasive B2T approaches have yet to achieve word error rates below 100%, with even the most promising results showing only marginal improvements when evaluated using character error rate or semantic similarity metrics (Jo et al., 2024; Yang et al., 2024b,c,a). To best progress the field, we believe it is essential to establish robust foundations before tackling the more complex objective of B2T. That said, LibriBrain is fully compatible with B2T translation tasks. Moreover, phoneme prediction is a critical component in successful invasive decoding pipelines (e.g. Moses et al., 2021; Willett et al., 2023).

6 Conclusion

In this paper we introduced LibriBrain, a landmark dataset containing over 50 hours of within-subject MEG recordings during naturalistic speech comprehension. This represents the deepest non-invasive speech brain dataset to date, exceeding previous datasets by 5–50×. By providing extensive high-quality data from a single subject, LibriBrain enables the exploration of neural decoding limits when individual variability is controlled—a crucial complement to existing multi-subject datasets.

7 Acknowledgments

We are grateful to our anonymous reviewers for their thoughtful comments and suggestions, and to the NeurIPS Datasets & Benchmarks organisers for all of their efforts in making the track possible. We acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558), Hartree Centre resources, and the NVIDIA Corporation for donating additional GPUs. PNPL is supported by the MRC (MR/X00757X/1), Royal Society (RG\R1\241267), NSF (2314493), NFRF (NFRFT-2022-00241), and SSHRC (895-2023-1022). The doctoral research of Miran Özdogan is further supported by the EPSRC (EP/W524311/1).

References

Accou, B., Bollens, L., Gillis, M., Verheijen, W., Van hamme, H., and Francart, T. (2024). SparrKULee: A speech-evoked auditory response repository of the KU Leuven, containing EEG of 85 participants. Data, 9:94.

Armeni, K., Güçlü, U., van Gerven, M., and Schoffelen, J.-M. (2022). A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Scientific Data, 9:278.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9/10):341–345.

Brennan, J. R. (2023). EEG Datasets for Naturalistic Listening to "Alice in Wonderland" (version 2). Data set.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Current Biology, 28:803–809.

Card, N. S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F. R., Kunz, E. M., and Brandman, D. M. (2024). An accurate and rapidly calibrating speech neuroprosthesis. New England Journal of Medicine, 391(7):609–618.

Csaky, R., van Es, M. W. J., Parker Jones, O., and Woolrich, M. (2023). Group-level brain decoding with deep learning. Human Brain Mapping, 44:6105–6119.

d'Ascoli, S., Bel, C., Rapin, J., Banville, H., Benchetrit, Y., Pallier, C., and King, J.-R. (2024). Decoding individual words from non-invasive brain recordings across 723 participants. arXiv preprint.

Ding, N. and Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. Frontiers in Human Neuroscience, 8:311.

Doyle, A. C. (1887). A Study in Scarlet. Ward, Lock & Co.

Doyle, A. C. (1890). The Sign of the Four. Spencer Blackett.

Doyle, A. C. (1892). The Adventures of Sherlock Holmes. George Newnes.

Doyle, A. C. (1893). The Memoirs of Sherlock Holmes. George Newnes.

Doyle, A. C. (1902). The Hound of the Baskervilles. George Newnes.

Doyle, A. C. (1905). The Return of Sherlock Holmes. George Newnes.

Doyle, A. C. (1915). The Valley of Fear. George H. Doran Company.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., and King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence, 5(10):1097–1107.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). TIMIT acoustic-phonetic continuous speech corpus. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC93S1.

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. Nature Neuroscience, 15(4):511–517.

Gracco, V. L. and Lofqvist, A. (1994). Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements. Journal of Neuroscience, 14(11):6585–6597.

Gwilliams, L., King, J.-R., Marantz, A., and Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. Nature Communications, 13(1):6606.

Harris, K. D. (2021). Nonsense correlations in neuroscience. bioRxiv.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5):393–402.

Iverson, P. and Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. The Journal of the Acoustical Society of America, 97(1):553–562.

Jayalath, D., Landau, G., and Parker Jones, O. (2025a). Unlocking non-invasive brain-to-text. arXiv preprint arXiv:2505.13446.

Jayalath, D., Landau, G., Shillingford, B., Woolrich, M., and Parker Jones, O. (2024). The Brain's Bitter Lesson: Scaling speech decoding with self-supervised learning. NeurIPS, NeuroAI Workshop.

Jayalath, D., Landau, G., Shillingford, B., Woolrich, M., and Parker Jones, O. (2025b). The Brain's Bitter Lesson: Scaling speech decoding with self-supervised learning. ICML.

Jo, H., Yang, Y., Han, J., Duan, Y., Xiong, H., and Lee, W. H. (2024). Are EEG-to-text models working? arXiv preprint.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. CoRR, abs/2001.08361.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto.

Landau, G., Özdogan, M., Elvers, G., Mantegna, F., Somaiya, P., Jayalath, D., Kurth, L., Kwon, T., Shillingford, B., Farquhar, G., Jiang, M., Jerbi, K., Abdelhedi, H., Mantilla Ramos, Y., Gulcehre, C., Woolrich, M., Voets, N., and Parker Jones, O. (2025). The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. NeurIPS, Competition Track. https://arxiv.org/abs/2506.10165.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. Cerebral Cortex, 15(10):1621–1631.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR.

Luo, H. and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron, 54(6):1001–1010.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174):1006–1010.

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., and Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620:1037–1046.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. The Journal of the Acoustical Society of America, 27(2):338–352.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. The Journal of the Acoustical Society of America, 85(5):2114–2134.

Moses, D. A., Mesgarani, N., Leonard, M. K., and Chang, E. F. (2016). Neural speech recognition: Continuous phoneme decoding using spatiotemporal representations of human cortical activity. Journal of Neural Engineering, 13(5):056004.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., and Chang, E. F. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, 385(3):217–227.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., Donaldson, D. I., Kohút, Z., Rueschemeyer, S.-A., and Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. eLife.

Niso, G., Gorgolewski, K. J., Bock, E., Brooks, T. L., Flandin, G., Gramfort, A., Henson, R. N., Jas, M., Litvak, V., Moreau, J. T., Oostenveld, R., Schoffelen, J.-M., Tadel, F., Wexler, J., and Baillet, S. (2018). MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. Scientific Data, 5:180110.

Obleser, J., Eisner, F., and Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. The Journal of Neuroscience, 28(32):8116–8123.

Ochshorn, R. M. and Hawkins, M. (2015). Gentle: A robust yet lenient forced aligner built on Kaldi.

Parker Jones, O. and Schnupp, J. W. H. (2021). How does the brain represent speech? In Pardo, J. S., Nygaard, L. C., Remez, R. E., and Pisoni, D. B., editors, The Handbook of Speech Perception, Second Edition, chapter 3. Wiley-Blackwell.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. NeurIPS, pages 8024–8035.

Peelle, J. E. and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. Frontiers in Psychology, 3:320.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. Journal of Neuroscience Methods, 162(1-2):8–13.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., Roberts, T., and Rowley, H. (2000). Auditory cortex accesses phonological categories: An MEG mismatch study. Journal of Cognitive Neuroscience, 12(6):1038–1055.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US. IEEE Signal Processing Society.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. PNAS, 103(20):7865–7870.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

Ridge, J. and Parker Jones, O. (2024). Resolving domain shift for representations of speech in non-invasive brain recordings. arXiv preprint.

Schnupp, J., Nelken, I., and King, A. J. (2012). Auditory Neuroscience: Making Sense of Sound. MIT Press, Cambridge, MA.

Schoffelen, J.-M., Oostenveld, R., Lam, N. H. L., Uddén, J., Hultén, A., and Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. Scientific Data, 6:1–13.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science, 270(5234):303–304.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. JMLR, 15(1):1929–1958.

Stevens, K. N. and Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. The Journal of the Acoustical Society of America, 64(5):1358–1368.

Tagliasacchi, M., Li, Y., Misiunas, K., and Roblek, D. (2020). SEANet: A multi-modal speech enhancement network. arXiv preprint.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience, 26(5):858–866.

Taulu, S. and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in meg measurements. Physics in Medicine & Biology, 51(7):1759.

Weide, R. L. (1998). The Carnegie Mellon pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., and Henderson, J. M. (2023). A high-performance speech neuroprosthesis. Nature, 620:1031–1036.

Wilson, G. H., Stavisky, S. D., Willett, F. R., Avansino, D. T., Kelemen, J. N., Hochberg, L. R., Henderson, J. M., Druckmann, S., and Shenoy, K. V. (2020). Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. Journal of Neural Engineering, 17(6):066007.

Yang, Y., Duan, Y., Jo, H., Zhang, Q., Xu, R., Parker Jones, O., Hu, X., Lin, C.-T., and Xiong, H. (2024a). NeuGPT: Unified multi-modal neural gpt. arXiv preprint.

Yang, Y., Duan, Y., Zhang, Q., Xu, R., and Xiong, H. (2024b). NeuSpeech: Decode neural signal as speech. arXiv preprint.

Yang, Y., Jo, H., Duan, Y., Zhang, Q., Zhou, J., Lee, W. H., Xu, R., and Xiong, H. (2024c). MAD: Multi-alignment MEG-to-text decoding. arXiv preprint.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2021). Scaling vision transformers. CoRR, abs/2106.04560.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11975–11986.

Appendices / Supplemental Materials

A Further Discussion

A.1 Ethical Considerations

The development and release of the LibriBrain dataset raise several important ethical considerations that we have carefully addressed throughout the research process:

Informed consent and participant privacy. The participant provided informed consent for data collection and approved the sharing of pseudonymised data for research purposes, in accordance with the University of Oxford's ethical oversight. To protect participant privacy, we have implemented procedures to replace or remove data that directly identifies an individual participant for all released data.

Open science and reproducibility. We have deliberately chosen public-domain materials (LibriVox audiobooks, Project Gutenberg texts) and open-source tools to ensure that the entire dataset can be freely redistributed without licensing constraints. This commitment to open science extends to our analysis code, which is publicly available and documented to facilitate reproducibility.

Risks of misinterpretation. Brain decoding technologies often generate public excitement that can exceed their actual capabilities, as evidenced by recent popular press articles. We have been careful to avoiding claims that might be used to exaggerate the current state of non-invasive speech decoding. For transparency, we report multiple performance metrics and statistical comparisons to chance-level baselines.

Dual-use considerations. The primary goal of this research is to develop assistive technologies for communication in specific clinical populations for whom the technology would be a life changing. However, brain decoding techniques might have the potential to be misused for privacy invasion if applied without consent. The current state of non-invasive methods remains far from enabling such applications. Nevertheless, we emphasise that ethical guidelines for neural recording should always include informed consent and participant autonomy. Our aim in in releasing data and methods to help standardise the field is that there will be opportunities for researchers to come together around a strong set of ethical guardrails for any future technologies.

Long-term data stewardship. We are committed to maintaining the dataset's availability and integrity over time. The BIDS version will similarly be hosted on a standard public platform (e.g. OSF[4]). In addition to the code being hosted on GitHub[5], we have distributed the dataset through Hugging Face[6], ensuring redundancy and continuous accessibility.

B Data Collection Methods

MEG recordings were acquired from a single, right-handed (male) volunteer with normal hearing and vision. The participant reported no history of neurological, developmental, or language-related disorders and was a native English speaker. Prior to participation, informed consent was obtained, authorising the use of pseudonymised data for research purposes. The study was approved by the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (R90053/RE002).

B.1 Stimulus Materials

Audiobooks for the MEG experiments were sourced from LibriVox (https://librivox.org/). As in Armeni et al. (2022), we used the same recording of Sir Arthur Conan Doyle's *The Adventures of Sherlock Holmes* (Doyle, 1892), plus six other audiobook recordings in the cannon of Sherlock

[4] OSF is a free, open platform that is popular for hosting BIDS-formatted datasets (see https://osf.io/).
[5] https://github.com/neural-processing-lab/frozen-pnpl
[6] https://huggingface.co/datasets/pnpl/LibriBrain

Holmes (Doyle, 1887, 1890, 1893, 1902, 1905, 1915). All were read by David Clarke, an adult male with a General British accent (see Table 6). Machine readable text for the audiobooks was recovered on Project Gutenberg (https://www.gutenberg.org/). As all audio and text used in this project are in the public domain, there are no limits on them being openly shared for science.

To prepare the stimuli for the MEG experiments, all audio files were converted to the uncompressed WAV format and resampled to 44.1 kHz with SoX. Our aim was to segment the audio into natural sentences or phrases, separated by pauses, and then match each segment to the appropriate text. This allowed us to send precise triggers to the MEG system during stimulus presentation, enabling continuous verification of the temporal alignment between MEG recordings, audio, and linguistic annotations throughout the experiment. Audio segmentation was achieved in phases: first automatically, using voice activity detection (VAD), then manually. The manual phase addressed several issues: refining segment boundaries where VAD had truncated quiet sounds (such as voiceless plosives in utterance initial position), resolving discrepancies between the written text and narrator's speech, and normalising textual representations (e.g. converting numerical expressions like "1066" to match their spoken form, whether "ten sixty-six" or "one thousand and sixty-six" or something else).

Concretely, VAD was performed using custom scripts in Praat (Boersma, 2001), identifying speech based on an intensity threshold of 59 dB and ensuring that only segments which surpassed this level and that exceeded a minimum duration of 600 ms were classified as speech. The resulting TextGrid annotations were then populated with corresponding text and subjected to meticulous manual correction. This labour-intensive process—requiring over 200 hours of expert human effort—significantly exceeds the typical standards for audiobook stimulus preparation. But it was worth it to us to reduce label noise and produce the highest-quality data possible.

B.2 Experimental Design & Procedure

Each session began with visually presented, self-paced instructions. After reading, the participant pressed a button to begin the experiment. A fixation cross was then displayed, followed by auditory presentation of an audiobook segment (a full chapter or partial chapter). At the end of each segment, the participant answered a comprehension question (e.g. "Where is the body of the murder victim found?") via a two-alternative forced choice (2AFC) using a handheld response device (e.g. "A: in the bedroom", "B: in the garden"). Short breaks were allowed between the presentation of one chapter and another.
Instructions and fixation cues were projected on a translucent whiteboard using a DLP LED projector (ProPixx, VPixx Technologies Inc., Saint-Bruno, Canada). The stimulus computer synchronises with the MEG hardware through a parallel port to deliver triggers marking stimulus onset with high temporal precision (at a millisecond timescale). Stimulus delivery was managed via the PsychoPy toolbox (Peirce, 2007). Auditory stimuli were delivered binaurally through tube earphones (Aero Technologies) at approximately 70 dB SPL, with minor adjustments to bass and treble based on participant preference. Responses were recorded using a MEG-compatible ResponsePixx Dual Handheld system (VPixx Technologies Inc., Saint-Bruno, Canada).
This protocol was repeated across multiple sessions. Each session lasted approximately three hours and typically covered from 3 to 5 audiobook chapters. Sessions were spaced at least one day apart, with no more than two months between sessions, depending on participant and experimenter availability.

B.3 Data Acquisition

Before data acquisition, the participant's head shape was digitised using a Polhemus Fastrak 3D digitiser (Polhemus, Vermont, USA). This process included identifying fiducial landmarks (nasion, and left and right pre-auricular points), as well as collecting approximately 250 additional scalp, forehead, and nose surface points. Five Head Position Indicator (HPI) coils were attached to the mastoid bones and forehead to continuously monitor head position via electromagnetic induction during scanning.
MEG data were recorded using a MEGIN Triux™ Neo system (York Instruments Ltd., Heslington, UK), comprising 102 magnetometers and 204 orthogonal planar gradiometers. The scanner was housed within a magnetically shielded room to minimise environmental noise. Prior to entering the room, the participant was screened for metallic objects or other sources of electromagnetic

interference. During scanning, the participant was seated and positioned so that their head was in close contact with the dewar. For each recording session, the participant's head was positioned as close as possible to a standard reference location, allowing for only minimal displacement (within a few millimeters), in order to reduce variability across sessions. Instructions were given to minimise head, body, and limb movements throughout the session. Recordings were sampled at 1000 Hz and band-pass filtered between 0.01 and 330 Hz. Eye-related artifacts were tracked using bipolar electrooculogram (EOG) electrodes—one pair placed horizontally at the outer canthi, and another vertically above and below the left eye. Cardiac activity was monitored via bipolar electrocardiogram (ECG) electrodes located on the clavicle and hip. Articulatory movements were continuously monitored with electromyography (EMG), using one electrode below the cheekbone to monitor jaw movement and another between the lower lip and chin for lip movement. Prior studies suggest that EMG signals measured from these two locations provide sufficient detail to distinguish basic phonemic contrasts (Gracco and Lofqvist, 1994).

B.4 Minimal Preprocessing Pipeline

The MEG data were minimally preprocessed to remove head movement, filter, and downsample to 250 Hz. First, head position information was extracted from the HPI measurements. Any bad channels were identified, removed, and restored using interpolation from nearby channels. External noise (e.g. environmental noise, stationary noise) was removed from MEG recordings offline using a Maxwell Filter software (tsss- filters; Taulu and Simola 2006). We used a temporally non-extended spatial Signal Source Separation (SSS) algorithm to suppress external sources of magnetic interference. Mains power at 50 Hz and at the 100 Hz harmonic were removed using notch filters. Bandpass filtering between 0.1 and 125 Hz was then applied with zero-phase two-pass Butterworth filters, to remove slow drifts and aliasing artefacts related to downsampling. Given the Nyquist theorem, we used 125 Hz exactly (sometimes lower values used) to keep, at least in theory, high gamma signal (despite 1/f power). The data were then downsampled from 1 kHz to 250 Hz, resulting in recordings of 306 channels $\times$ T where T was measured in 4 millisecond samples.

To ensure quality control, we visualised the events and then the PSD (power spectral density) on raw data, after notch filter, after bandpass filter, and after downsampling (see Figure 4 for an example). The choice of hyperparameters was decided based on pilot data, where we explored decoding with and without the SSS step of Maxwell Filtering and different bandpass filter range. So, the values used here reflect the best choices in pilot analyses.

B.5 Annotations/Event Files

Annotations for each session include onset times and durations (in seconds) for event types such as silence, word (e.g. *A*, *Study*, *in*, *Scarlet*), and phoneme (e.g. *ah*, *s*, *t*, *ah*, *d*, *iy*, *ih*, *n*, *s*, *k*, *aa*, *r*, *l*, *ah*, *t*) (see Table 7) The annotations include other information, notably the position of the phonemes within a word. Position is indicated using the conventions from Kaldi (Povey et al., 2011) (*B* = beginning, *I* = inside, and *E* = end; the symbol *S* = singleton, which is both the beginning and end of a word). For each MEG session (FIF file), there is a corresponding event file (TSV file). The event files can be used to specify labels for supervised tasks. The `pnpl` library comes with methods to load speech/non-speech and phonemes. As a community project, we aim to add support for other standard decoding tasks. Pull requests to the github repo are welcome.

The event files were created from the manually corrected (e.g. text normalised) transcripts what were used in the design of the MEG experiment. Text and audio were then force-aligned using Gentle (Ochshorn and Hawkins, 2015). As the text was already aligned to short utterances, as described above (section B.1) using VAD and over 200 hours of quality control and manual correction, the job of the forced-aligner was greatly simplified. Rather than pass entire chapters to the forced-aligner, we passed it the short utterance–text pairs. As a sanity check for the quality of the forced-aligner annotations, we examined the duration of linguistic segments of varying lengths—such as consonants, vowels, and words categorised by character count as short ($\leq$ 3 characters), medium (4–6 characters), or long ($\geq$ 7 characters). As expected, longer linguistic units were associated with correspondingly longer temporal annotations, providing internal validation for the temporal accuracy of the alignment (see Table 8).

When segmentation using the standard configuration options failed, we explored alternative alignment strategies to improve accuracy. Specifically, enabling the "include disfluencies" option allowed the forced aligner to incorporate non-lexical vocalisations and disfluencies (e.g., "um", "uh", false starts) that are typically excluded from transcripts. This adjustment helped the model better capture natural variations in spontaneous speech. Additionally, enabling the "conservative" alignment option prompted the algorithm to assign word boundaries only when it had high confidence in the timing correspondence between the transcript and the audio, thereby reducing misalignments in noisy or ambiguous segments. To further address issues caused by overlapping or coarticulated speech, we ran the alignment process multiple times while systematically removing one word at a time from the transcript. This iterative method helped disentangle problematic word pairs.

Despite these precautions, the forced aligner occasionally produced sub-optimal outputs, which we identified and corrected through a labour-intensive manual inspection and adjustment process. For instance, some word occurrences were incorrectly labelled as out-of-vocabulary (OOV)—not because they were absent from the vocabulary (e.g. foreign language words, proper names, etc.), but because the forced-alignment algorithm failed to detect them for various reasons (e.g. foreign accent, excessive coarticulation, noisy pronunciation etc.). Many of these misclassified OOV instances were subsequently corrected by visually inspecting the speech spectrogram and manually annotating words and phonemes onsets and offsets using Praat (Boersma, 2001). Following this intervention, the proportion of OOV annotations was reduced to approximately 3.5% per book chapter.

Finally, we expect brain responses to lag behind auditory stimuli due to neural conduction delays along the ascending auditory pathway, from the cochlea through the brainstem and thalamus to the auditory cortex and beyond (e.g. Schnupp et al., 2012; Parker Jones and Schnupp, 2021). However, we did not apply a fixed temporal offset (or "fudge factor") when generating the event files. One reason is that conduction delays may vary across brain regions and processing stages. Rather than imposing a single correction, we chose to preserve the original speech timings in the event annotations. This allows users to explicitly model latencies in their analyses or to fit models that can learn temporal offsets directly from the data.

B.6 Standard Data Splits

Through many of the success stories of deep learning, it has been important to have standard data splits (e.g. MNIST, CIFAR-10, ImageNet). This provides like for like comparisons when developing methods. Poor data splits can also be problematic, either to contaminate the training set or leak information. A potential source of leakage in neuroimaging studies is the use of data from the same set of scanning sessions for training and testing. Certainly, taking test data from between samples of training data can be problematic. Taking the last data from the day as test has been used in many cases, but may still leak information that would not be available in the use case scenario in which train and test took place on different days. Therefore, we opted for the more conservative setup, scheduling a special scan to collect all of the holdout data on a separate day from the rest of the training data in the dataset. This means that there are two good reasons to use the designated holdout data in future: (1) consistency in the community, and (2) strongest control against data leakage.

The holdout data, which were acquired on a separate day from any train data, are Sherlock1 chapters 11–14. Specifically, we use Sherlock1 chapter 11 as the validation dataset (duration ≈ 21 minutes), to have the same distribution as the test data. The standard test data are Sherlock1 chapter 12 (duration ≈ 23 minutes).

We are holding out the data from Sherlock1 chapters 13 and 14 as evaluation sets for future machine learnign competitions (duration $\approx 27 + 14$ minutes). This still leaves over 50 hours in this release for training alone; it leaves over 52 hours for training, validation, and test (see Table 2).

C Additional Dataset Details

Further details on the audio books used in LibriBrain are summarised in Table 6. Table 7 provides an example of the event annotations provided in the dataset. Table 8 presents a validation of forced alignment annotations, showing the relationship between linguistic unit length and corresponding temporal duration. Additionally, Figure 4 illustrates the minimal preprocessing steps applied to a representative MEG recording session.

Table 6: Audiobooks available in LibriBrain

Book	Name	Notes	Sessions	Hours
1	A Study in Scarlet (1887)	Novel	14*	04:37:34
2	The Sign of the Four (1890)	Novel	12	04:27:31
3	The Adventures of Sherlock Holmes (1892)	Short Stories	12	10:56:13
4	The Memoirs of Sherlock Holmes (1894)	Short Stories	12	08:53:17
5	The Hound of the Baskervilles (1901–1902)	Novel	15	06:10:32
6	Return of Sherlock Holmes (1905)	Short Stories	14	11:51:17
7	The Valley of Fear (1914–1915)	Novel	14	06:06:17
				53:02:41

*Two sessions held out for competitions.

Table 7: An example of an '*events.tsv' file. Each row logs one event relative to the MEG recording session. Time values for both onset and duration are in seconds. For this example, we only show the events corresponding to one sentence extracted from the first audio book.

#	onset	duration	type	segment	position
1	28.772	1.188	silence		
2	30.084	0.1	word	A	
3	30.084	0.1	phoneme	ah	S
4	30.184	0.37	word	Study	
5	30.184	0.08	phoneme	s	B
6	30.264	0.08	phoneme	t	I
7	30.344	0.08	phoneme	ah	I
8	30.424	0.06	phoneme	d	I
9	30.484	0.07	phoneme	iy	E
10	30.556	0.14	word	in	
11	30.556	0.05	phoneme	ih	B
12	30.604	0.09	phoneme	n	E
13	30.696	0.59	word	Scarlet	
14	30.696	0.09	phoneme	s	B
15	30.784	0.07	phoneme	k	I
16	30.856	0.11	phoneme	aa	I
17	30.964	0.01	phoneme	r	I
18	30.976	0.05	phoneme	l	I
19	31.024	0.07	phoneme	ah	I
20	31.096	0.19	phoneme	t	E

Table 8: Forced-aligner validation based on annotated linguistic units duration. Longer linguistic units were associated with longer temporal annotations.

Linguistic unit	Mean	Standard Deviation
Consonants	0.076 s	0.039 s
Vowels	0.082 s	0.042 s
Short Words ($\leq$ 3 characters)	0.158 s	0.083 s
Medium Words (4–6 characters)	0.308 s	0.129 s
Long Words ($\geq$ 7 characters)	0.515 s	0.148 s

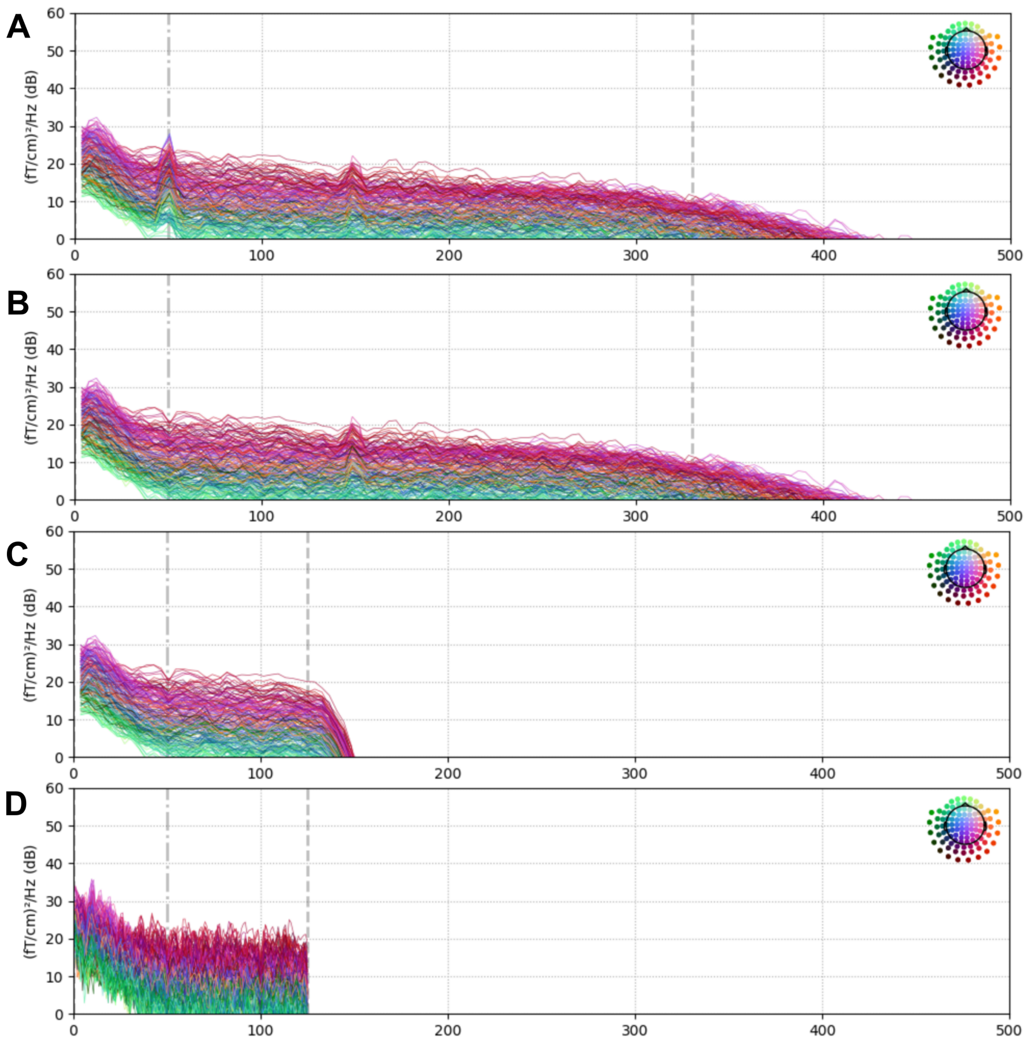


Figure 4: Minimal pre-processing example for one representative recording session. Power spectral density estimated for 204 gradiometers (colour-coded according to their spatial position, see topography in the upper right) is represented in each panel. (A) Raw data, no preprocessing. (B) Notch filter is applied, removing 50 Hz power line noise. (C) Bandpass filter 0.1-125 Hz is applied, attenuating power outside this frequency range. (D) Downsampling to 250 Hz is applied.

C.1 Phonemes colour-coding scheme

Phonemes were grouped into perceptually and acoustically meaningful categories relevant to speech perception (see Figure 2). These groupings are consistent with previous literature investigating how the brain organises speech input along perceptual dimensions that are rooted in acoustic space. The rationale for choosing the colour-coding scheme applied to International Phonetic Alphabet (IPA) vowels and consonants charts is described below.

The vowels are grouped into five categories corresponding to broad acoustic similarity patterns. In particular, formant structure, diphthongisation, and vowel tenseness.

- **High front vowels** $[iy, ih]$: These vowels are characterised by low F1 and high F2, distinguishing them acoustically as front and close. This group reflects fine differences in vowel height and tenseness (e.g., tense 'iy' vs. lax 'ih').
- **High back vowels** $[uh, uw]$: These share low F1 and lower F2 values, indicating backness and closeness. The grouping captures rounding and tongue root position differences.
- **Back diphthongs and mid-back vowels** $[oy, ow, ao]$: These vowels include both monophthongal ('ao') and diphthongal ('oy', 'ow') elements, but all share similar backness and rounded quality, leading to comparable acoustic transitions in F2 and F3.
- **Mid-front vowels** $[eh, er, ey]$: This group includes vowels with mid-height and either fronted or rhoticised qualities. The inclusion of 'er' and 'ey' reflects their transitional or complex spectral features.
- **Low and diphthongal vowels** $[ah, aw, aa, ae, ay]$: These vowels tend to have high F1 and variable F2 trajectories. They include open and central vowels as well as diphthongs like 'ay', capturing broad dynamic articulations.

These groupings reflect acoustic proximity and perceptual clustering (Miller, 1989), which are more informative for understanding neural responses to speech perception than articulatory descriptions alone. In the auditory cortex, vowel processing is driven by spectral characteristics (Mesgarani et al., 2014), particularly formant patterns (Obleser et al., 2008). This grouping schema aligns well with multidimensional scaling studies of vowel similarity and perceptual confusability, supporting their use in neural decoding and classification models of speech perception (Iverson and Kuhl, 1995).

The consonants are grouped into five categories according to manner of articulation, rather than place of articulation.

- **Stops** $[p, b, t, d, k, g]$: Characterised by a closure phase and a burst of release energy, stops are defined by temporal silence followed by transient noise.
- **Fricatives and glottal fricatives** $[f, v, th, dh, s, z, sh, zh, hh]$: Sustained high-frequency turbulence is the key acoustic cue. They form a perceptual continuum based on noise spectra and voicing.
- **Nasals** $[m, n, ng]$: These consonants share low-frequency energy (nasal formants) and anti-resonances.
- **Affricates** $[ch, jh]$: These consonants are acoustically a hybrid between stops and fricatives, with a characteristic stop-like closure followed by frication.
- **Approximants and liquids** $[r, w, y, l]$: These consonants have continuous airflow and formant transitions. Grouping them together reflects their similar spectral dynamics and transitional nature.

While place of articulation is more important during speech production, it is less distinct during speech perception, especially in connected speech, where coarticulation often blurs place of articulation cues (Miller and Nicely, 1955). Moreover, place of articulation cues are acoustically more variable and less reliable than manner cues (Stevens and Blumstein, 1978). In contrast, manner of articulation produces salient acoustic patterns that are more consistently distinguishable in the auditory signal and more likely to drive perceptual differentiation and its neural underpinnings (Shannon et al., 1995). This is supported by psycholinguistic research on speech perception, which shows that listeners are more sensitive to differences in manner than in place, particularly in early perceptual stages (Phillips et al., 2000). Moreover, MEG responses to consonants often show stronger discrimination along acoustic dimensions that reflect temporal envelope (as in stops vs. fricatives) and spectral shape (as in nasals vs. fricatives), both of which map more directly onto manner than onto place of articulation (Liebenthal et al., 2005).

D Additional Speech Detection Details

D.1 Model Architecture and Hyperparameters

For speech detection, we reuse the architecture that we developed for phoneme classification, as described in Appendix E.1. The only hyperparameter we change is the learning rate, which we set to 0.0003 for all speech experiments based on a quick manual exploration.

D.2 Additional Results

During speech perception, neural responses are known to entrain to the rhythmic structure of the auditory stimulus, leading to distinct signatures in MEG recordings. One prominent effect is an increase in amplitude over temporal sensors, particularly those aligned with the location of the primary auditory cortex. This amplitude enhancement reflects stronger evoked responses to the structured acoustic input of speech and is consistent with findings that auditory cortex is highly sensitive to the temporal dynamics of speech (Luo and Poeppel, 2007). In addition to amplitude, phase consistency across trials, as measured by inter-trial coherence (ITC), is significantly elevated during speech perception compared to non-speech intervals. This increase indicates that the phase of low-frequency neural oscillations becomes more aligned across trials, likely driven by the quasi-rhythmic syllabic structure of speech (Ding and Simon, 2014; Peelle and Davis, 2012). Furthermore, power in the delta (1–4 Hz) and theta (4–8 Hz) frequency bands increases during speech perception, particularly in bilateral temporal sensors. This frequency range matches the prosodic and syllabic rhythms of natural speech and reflects the alignment of endogenous oscillatory activity with external stimulus dynamics (Giraud and Poeppel, 2012).

To ensure the reliability and consistency of the data, we compared neural responses between speech and non-speech segments across all recording sessions—including each book and chapter. We consistently observed modulations in signal amplitude, phase, and spectral power, indicating robust differences between conditions (see Figure 5 for an example from one representative recording session). To illustrate this, we averaged the neural responses from 1,000 randomly sampled segments, each lasting 800 milliseconds following stimulus onset. For statistical validation, we employed a bootstrapping method with replacement to estimate variability, followed by a cluster-based permutation test to identify spatio-temporal and time–frequency regions showing significant differences between speech and non-speech conditions. These neural signatures align with established findings in the literature. Importantly, while such differences are robust when using trial-averaging methods, the goal of the speech detection task is to identify these modulations at the single-trial level. This presents a substantially greater challenge, as many of the relevant neural features—particularly those related to phase and power—tend to become evident only when data are averaged across multiple trials.

D.3 Computational Requirements

Each of our individual speech detection experiments can be replicated in under 12 hours on an NVIDIA H100 GPU. Given that we have conducted a total of 31 experiments this means that all of the speech detection results are reproducible in less than 372 GPU hours. They require no more than 64GiB of memory. We conducted experiments on an internal cluster. The research project required additional compute for preliminary exploration and choosing hyperparameters.

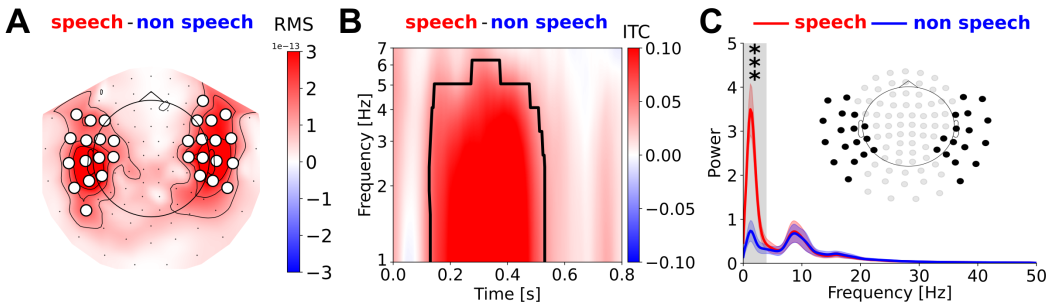


Figure 5: Differences in amplitude, phase, and power between speech and non-speech segments. (A) Topographic map of the root mean square (RMS) difference between speech and non-speech event-related potentials (ERPs), showing that speech segments exhibit higher average amplitude in temporal sensors. Statistically significant sensors are indicated by white dots. (B) Time–frequency representation of the difference in inter-trial coherence (ITC) between speech and non-speech segments. Speech segments demonstrate greater phase consistency across trials in lower frequencies (1–7 Hz; delta–theta range) in temporal sensors. Statistically significant clusters are outlined with a black contour. (C) Power spectral density (PSD) for speech and non-speech segments, revealing higher power in lower frequencies (1–7 Hz; delta–theta range) in temporal sensors during speech. Statistically significant frequency bands are highlighted with a semi-transparent grey rectangle.

E Additional Phoneme Classification Details

E.1 Model Architecture and Hyperparameters

We choose the SEANet architecture as a starting point because it has been successfully applied to MEG data in previous work (Jayalath et al., 2024, 2025b). We removed several layers because they did not increase performance. Furthermore, we add dropout (Srivastava et al., 2014) which we found to increase validation performance. A visualisation of our architecture can be found in Figure 6. Hyperparameters can be found in Table 9 and 10. We chose these based on a manual exploration. We evaluate the test partition using the model checkpoint with the best validation loss. For the experiments with averaged phoneme samples we change the learning rate to 0.00025. An exception are the scaling experiments with 10 averaged input samples where we use learning rate 0.0005.

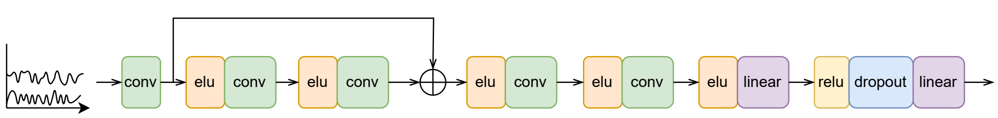


Figure 6: Convolutional Neural Network for Phoneme Classification

Table 9: Phoneme Classification Hyperparameters

Hyperparameter	Value
Learning Rate	1e-4
Batch Size	256
Dropout Rate	0.5
Seeds	[0-9]
Optimiser	Adam (Kingma and Ba, 2017)
Beta 1	0.9
Beta 2	0.999
Epsilon	1e-08

Table 10: Architecture Hyperparameters (Layer-by-Layer)

Layer #	Type	In Channels	Out Channels	Kernel Size	Stride	Padding
1	Conv1D	306	128	7	1	same
2a	ResNet Block – Conv1D	128	128	3	1	same
2b	ResNet Block – Conv1D	128	128	1	1	same
3	ELU Activation	-	-	-	-	-
4	Conv1D	128	128	50	25	none
5	ELU Activation	-	-	-	-	-
6	Conv1D	128	128	7	1	same
7	ELU Activation	-	-	-	-	-
8	Flatten	-	-	-	-	-
9	Linear	512	512	-	-	-
10	ReLU Activation	-	-	-	-	-
11	Dropout	-	-	-	-	-
12	Linear	512	#classes	-	-	-

E.2 Additional Results

Here we provide additional results on phoneme classification. In Figure 7B-C, we show the phoneme specific F1 scores of our model. For this we use the output of the full phoneme model and for each phoneme consider the one-vs-all binary classification problem. For reference, in Figure 7A, we show the frequencies of each phoneme in the English language. In Figure 9, we show the confusion matrix for the phoneme classification model predictions averaging over 10 samples for each phoneme. Furthermore, we consider the task of predicting averaged phoneme samples. For this we take the MEG window of multiple samples corresponding to the same phoneme and take their average both during model training and evaluation. In Figure 8 we show how performance increases with the number of samples averaged.

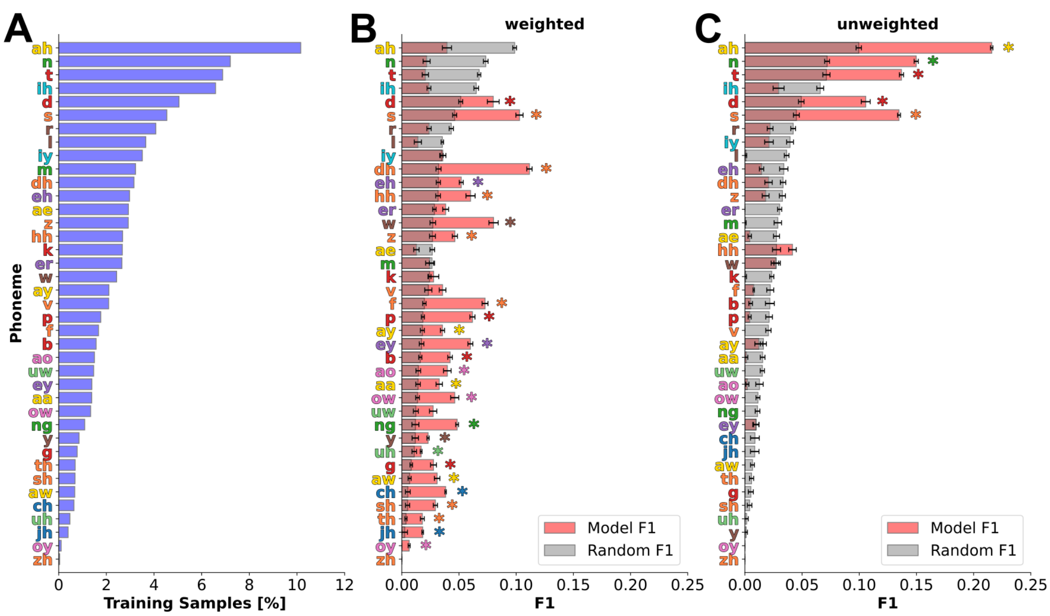


Figure 7: Per-phoneme analysis of our model's performance. (A) Distribution of training samples across phonemes. (B) Per-phoneme F1 scores when using a weighted loss function; significant differences to random baseline indicated ($p < 0.01$, one-sided exact permutation test). (C) Per-phoneme F1 scores as in (B), but using the default unweighted loss function.

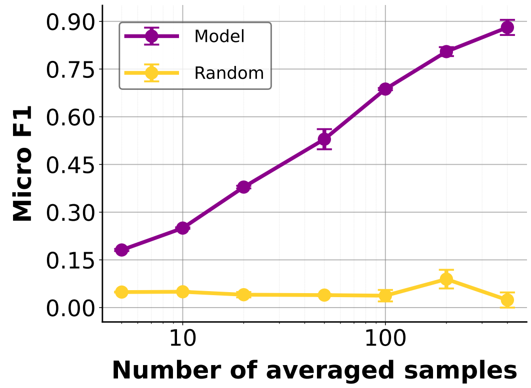


Figure 8: Effect of sample averaging on phoneme classification performance. Performance improves with the number of input windows averaged per phoneme.

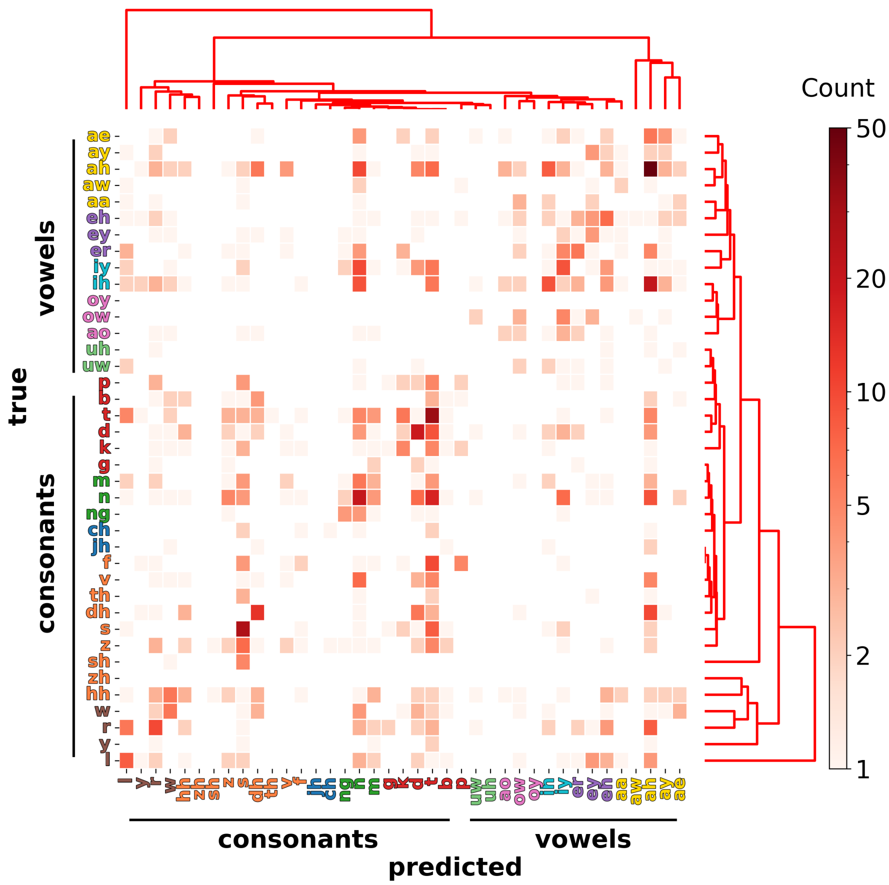


Figure 9: Confusion Matrix for the Phoneme Classification Task averaging over 10 samples for each phoneme. The distribution of predicted versus true phoneme labels is represented for 39 phonemes. Tick labels are colour-coded according to a predefined colour scheme based on the International Phonetic Alphabet (IPA) classification, grouping phonemes by place and manner of articulation: consonants and vowels. The 24 consonants are listed first, followed by 15 vowels. Dendrograms along the top and right margins represent the output of an unsupervised hierarchical clustering algorithm applied to the confusion matrix rows and columns, respectively, highlighting structural similarities in prediction patterns without affecting the phoneme order.

E.3 Computational Requirements

Each of our individual phoneme experiments can be replicated in under 12 hours on an NVIDIA H100 GPU. Given that we have conducted a total of 52 experiments this means that all of the phoneme results are reproducible in less than 624 GPU hours. They require no more than 64GiB of memory and less than 100GiB of storage. We conducted experiments on an internal cluster. The research project required additional compute for preliminary exploration and choosing hyperparameters.

F Additional Word Classification Details

As the word classification results are taken from Jayalath et al. (2025a), we restate their model description and hyperparameters here, which themselves are mostly reproduced from d'Ascoli et al. (2024). MEG windows are first encoded by a signal encoder. The input windows, aligned to consecutive word onsets, pass through a spatial attention (Défossez et al., 2023) that projects each input window channel dimension into a latent dimension of size 270 through an attention mechanism with scores derived from the (x, y) positions of the sensors. Then, they are encoded through a set of dilated convolutions using the same encoder as Défossez et al. (2023) which produces a temporal embedding for every time point of the input window. They take the mean over the time dimension to reduce this to a single embedding. If there are N word-aligned input windows, there will be N such embeddings.

Finally, these embeddings pass through a 1024-dimensional bidirectional attention transformer encoder with 16 layers, 16 heads, and rotary positional embeddings. The output embeddings of this transformer are optimised for the target middle layer embeddings of the T5 LLM (where words consisting of multiple tokens are averaged into a single target embedding).

Baseline results were obtained using the dataset released by Armeni et al. (2022), made available under the RU-DI-HD-1.0 license.

Table 11: Word classification hyperparameters.

Hyperparameter	**Value**
Batch size / seq. length	64
Learning rate	1e-5
Optimiser	AdamW (Loshchilov and Hutter, 2019)
Annealing schedule	Cosine (min. 1e-6 after 50 epochs)
Early stopping patience	5 epochs
Early stopping metric	Val. top-10 word class. accuracy
Encoder	Brain model (Défossez et al., 2023)
Transformer depth	16
Transformer heads	16
Transformer dimension	1024
Transformer attn. dropout	0.1
Transformer pos. emb.	Rotary

F.1 Additional Results

The word classification model output consists of predictive probabilities for each word in the vocabulary, indicating the likelihood of each word being the target word in the test set. For illustration, Figure 10 represents four successful examples where the ground truth word was among the model's top ten predictions. An inspection of these top predictions across the test set revealed that classification performance is influenced by both word frequency and word type. To investigate the effect of word frequency, we divided the 250 most frequent words into four equally sized quartiles (0–25%, 25–50%, 50–75%, 75–100%) and evaluated model performance separately for each group. As shown in Figure 11B-C, accuracy drops markedly after the first quartile, indicating that higher-frequency words are more reliably predicted. To assess the influence of word type, we categorised the same set of frequent words by part-of-speech (POS): nouns, verbs, adjectives, adverbs, and function words. Performance was largely consistent across POS categories, with the exception of function words, which were predicted with notably higher accuracy. This finding aligns with the frequency effect, as function words are typically repeated more often in the training set (Figure 11A). Overall, these results underscore the importance of scaling: model performance is significantly better for words that occur frequently in the training data, highlighting the impact of the training set size on word classification accuracy.

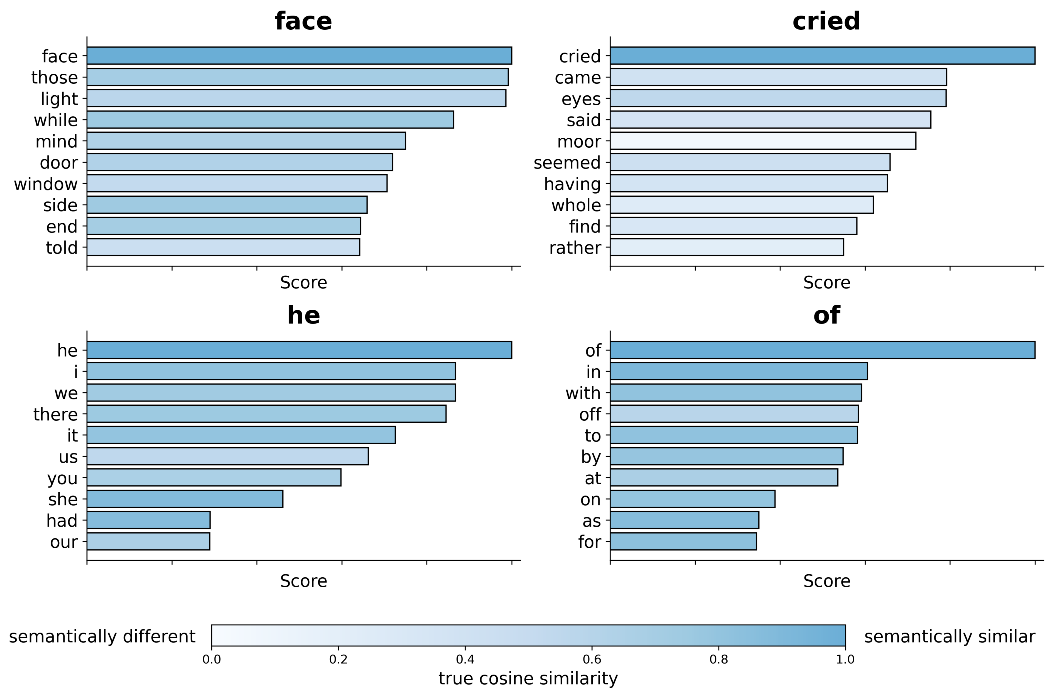


Figure 10: Top-10 model predictions for the word classification task. For illustration, four examples in which the model successfully included the correct ground truth word within its Top-10 predictions are shown. Model outputs were ranked based on softmax-normalised probabilities. To provide an interpretable view of the model's semantic performance, we visualised pairwise semantic similarity between the ground truth and the predicted words. Semantic similarity was estimated using cosine distance between 300-dimensional GloVe word embeddings.

F.2 Computational Requirements

Each of our individual word classification experiments can be replicated in under 12 hours on an NVIDIA V100 GPU. Given that we have conducted a total of 30 experiments this means that all of the word classification results are reproducible in less than 360 GPU hours. They require no more than 64GiB of memory. We conducted experiments on an internal cluster. The research project required additional compute for preliminary exploration and choosing hyperparameters.

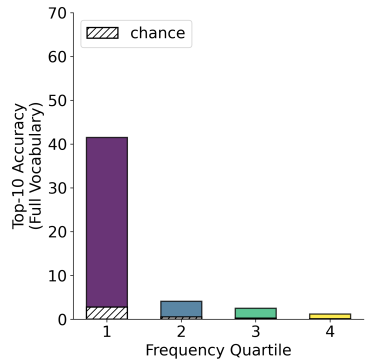


Figure 11: Impact of word frequency in the training set on classification model performance. Words in the test set were divided into four quartiles based on their frequency of occurrence in the training set, ranging from the most frequent to the least frequent. Top-10 accuracy was then computed across the full vocabulary, with each quartile reflecting the model's performance on words belonging to that frequency range. Chance-level performance was estimated by randomly shuffling the Top-10 predictions 1,000 times to create a null distribution for comparison.

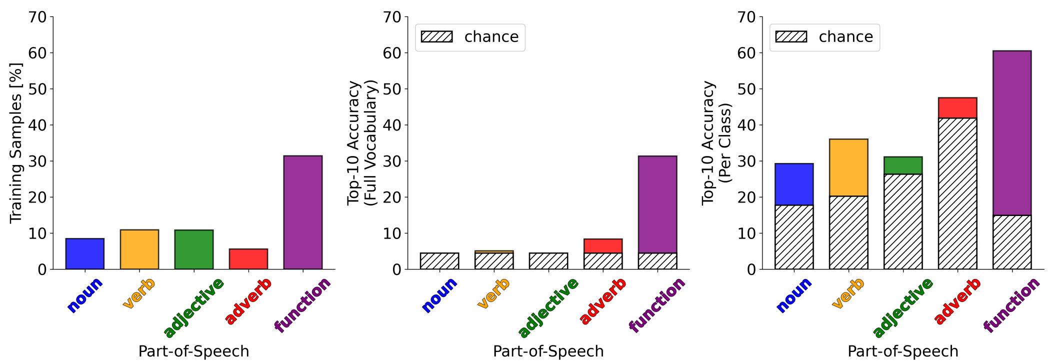


Figure 12: Effect of Part-of-Speech (POS) category on word classification model performance. (A) Distribution of training samples across POS categories, including nouns, verbs, adjectives, adverbs, and function words. (B) Top-10 accuracy computed over the full vocabulary, incorporating all instances of the 250 most frequent words in the test set. (C) Top-10 accuracy computed within each POS category, using the same 250 most frequent words grouped by their POS class. Chance-level performance was estimated by randomly permuting the model's Top-10 predictions 1,000 times, providing a baseline for comparison.

Figure 13: Top Function Words Cloud.

Table 12: Top Function Words

Word	Freq	Word	Freq	Word	Freq	Word	Freq
the	26096	so	1794	about	775	against	275
and	12778	by	1737	before	763	between	270
of	12300	all	1726	only	715	while	243
i	11510	were	1713	am	697	under	229
to	11155	been	1703	here	672	being	216
a	11102	an	1571	did	671	both	211
that	7892	what	1553	any	650	having	200
in	7640	your	1501	how	644	i'll	189
it	7326	are	1477	other	638	whom	173
he	7229	very	1390	than	631	each	173
was	6738	her	1387	their	570	does	169
you	6415	if	1371	where	555	yourself	144
his	5559	when	1335	own	490	because	133
is	4523	out	1319	after	471	during	130
had	4213	then	1254	such	470	nor	123
have	3822	she	1227	these	465	i've	108
with	3677	will	1211	through	431	above	108
my	3659	up	1204	just	428	doing	101
as	3327	who	1167	once	393	i'm	100
for	3278	they	1148	most	393	ourselves	92
at	3271	our	1142	again	386	you'll	80
which	3132	some	1137	why	368	further	73
we	2828	has	1131	himself	365	won't	71
but	2789	do	1093	off	345	itself	71
me	2538	into	1044	until	339	yours	62
not	2536	or	959	too	331	he's	61
this	2508	now	946	its	330	didn't	60
be	2465	down	923	don't	316	themselves	60
him	2303	them	877	myself	306	i'd	55
there	2235	can	874	few	299	you've	43
from	2080	more	863	those	299	we'll	42
on	1968	should	855	same	289	herself	39
no	1891	over	818	it's	285	you're	38

Figure 14: Top Content Words Cloud.

Table 13: Top Content Words.

Word	Freq	Word	Freq	Word	Freq	Word	Freq
said	2226	however	427	looked	276	also	203
upon	1893	go	420	half	274	evening	202
one	1860	saw	402	life	273	help	202
holmes	1846	left	399	mind	270	already	201
man	1533	yet	396	knew	269	death	201
would	1501	away	396	turned	265	sat	200
could	1320	three	389	course	262	across	198
us	1088	get	385	seemed	260	hardly	198
well	1041	thought	383	even	258	table	198
mr	975	matter	378	though	252	paper	198
see	947	cried	374	anything	248	moor	197
two	816	find	373	london	246	miss	196
little	768	make	371	doubt	245	oh	195
sir	765	asked	369	men	245	mcmurdo	195
may	762	round	364	went	244	passed	193
come	754	great	362	answered	243	mrs	193
know	732	morning	357	dear	243	lady	193
time	697	day	356	work	242	moment	192
room	693	end	352	open	239	gave	190
came	637	take	351	told	231	read	189
must	635	young	351	rather	228	looking	188
think	614	every	349	four	228	hall	188
face	598	right	348	clear	228	leave	187
back	597	still	345	whole	226	indeed	186
house	583	friend	344	black	226	taken	186
watson	560	sherlock	337	since	225	white	186
way	555	took	331	dr	225	father	185
good	544	done	327	business	224	another	184
hand	532	side	323	thing	222	better	184
never	526	head	319	police	221	road	184
last	512	small	312	years	220	hear	181
night	510	give	310	lay	220	chair	181
door	500	look	310	fellow	218	cannot	180
case	490	quite	310	always	217	set	178
might	490	let	307	behind	217	lestrade	178
nothing	489	street	307	gone	216	words	178
much	485	light	307	sure	214	fear	178
long	482	hands	304	five	214	ask	176
shall	481	ever	304	certainly	213	hour	175
say	471	put	302	many	209	part	173
made	467	name	302	together	208	strange	173
like	460	place	300	word	206	among	173
first	456	seen	298	woman	206	really	172
found	454	window	297	instant	205	front	172
old	447	enough	293	point	205	perhaps	170
eyes	445	something	293	brought	204	heart	167
yes	441	without	287	within	204	public	166
tell	438	far	286	dark	204	wife	166
heard	429	got	284	stood	204	crime	165

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the properties of the dataset as described in Section 3, the experimental findings detailed in Section 4, and the limitations discussed in Section 5.2.

   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: We discuss limitations in Section 5.2. Ethical and privacy considerations are discussed in Appendix A.1.

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 do not present theoretical 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: Aside from providing the code we also detail all information necessary to reproduce of our experimental results in Appendix E, D, and, F. We also provide simple access to the dataset ensuring that our results are easily reproducible.

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: We provide execution-ready scripts to recreate the training runs of the experimental results. The link to these is available in Section 4.

   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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail information necessary to understand the main results in Section 4. We describe data splits in Section B.6. We provide the full training and test details in Appendix E, D, and 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]

   Justification: We report error bars that represent standard error across multiple seeds in Section 4. For the phoneme classification and speech detection experiments we report statistical significance compared to a naive baseline as measured by an exact permutation test.

   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]

Justification: We provide details on computer resources need to reproduce the results in Appendix D.3, E.3, and F.2.

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]

Justification: We discuss ethical considerations at length in A. The NeurIPS Code of Ethics was reviewed and adhered to.

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]

Justification: We provide an extensive discussion of potential positive and negative societal impacts in Section 5, and Appendix A.

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: There are no immediate risks for misuse associated with the data and models we are publishing. However, we are intentionally releasing them under a CC BY-NC license to retain the ability to prevent potential unethical commercial use in the future.

    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: The baseline word classification results rely on the dataset published by Armeni et al. (2022). We have appropriately credited the authors in the main text and explicitly referenced the dataset's license (RU-DI-HD-1.0) in Appendix F.

    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 provide extensive documentation alongside our code and dataset on GitHub and Hugging Face including information on licensing, training, and limitations.

    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: [Yes]

    Justification: In Appendix B.2 and B.3 we give a detailed account of instructions given to participants.

    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: [Yes]

    Justification: The participant provided informed consent for data collection and approved the sharing of pseudonymised data for research purposes, in accordance with the University of Oxford's ethical oversight.

    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: This research does not involve LLMs as any important, original, or non-standard components.

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.