Elementary, My Dear Watson: Non-Invasive Neural Keyword Spotting in the LibriBrain Dataset

Gereon Elvers
PNPL
Department of Engineering Science
University of Oxford, UK
gereon.elvers@tum.de

Gilad Landau
PNPL
Department of Engineering Science
University of Oxford, UK
gilad@robots.ox.ac.uk

Oiwi Parker Jones
PNPL
Department of Engineering Science
University of Oxford, UK
oiwi@robots.ox.ac.uk

Abstract

Non-invasive brain-computer interfaces (BCIs) are beginning to benefit from large, public benchmarks. However, current benchmarks target relatively simple, foundational tasks like Speech Detection and Phoneme Classification, while application-ready results on tasks like Brain-to-Text remain elusive. We propose Keyword Spotting (KWS) as a practically applicable, privacy-aware intermediate task. Using the deep 52-hour, within-subject LibriBrain corpus, we provide standardized train/validation/test splits for reproducible benchmarking, and adopt an evaluation protocol tailored to extreme class imbalance. Concretely, we use area under the precision-recall curve (AUPRC) as a robust evaluation metric, complemented by false alarms per hour (FA/h) at fixed recall to capture user-facing trade-offs. To simplify deployment and further experimentation within the research community, we are releasing an updated version of the `pnpl` library with word-level dataloaders and Colab-ready tutorials. As an initial reference model, we present a compact 1-D Conv/ResNet baseline with focal loss and top-k pooling that is trainable on a single consumer-class GPU. The reference model achieves $\sim 13\times$ the permutation-baseline AUPRC on held-out sessions, demonstrating the viability of the task. Exploratory analyses reveal: (i) predictable within-subject scaling—performance improves log-linearly with more training hours—and (ii) the existence of word-level factors (frequency and duration) that systematically modulate detectability.

1 Introduction

The arrival of large and readily available datasets has begun to supply non-invasive brain-computer-interface (BCI) research with the kind of "common yard-stick" that ImageNet [Russakovsky et al., 2015] provided for computer vision. Among current non-invasive datasets for decoding speech, LibriBrain [Özdogan et al., 2025] is the *deepest* (i.e., *largest within-subject*) with 52 hours of magnetoencephalography (MEG) recorded from a single participant. This dataset forms the foundation for the 2025 PNPL Competition [Landau et al., 2025], an open machine-learning competition that has catalyzed progress on two foundational decoding tasks: Speech Detection and Phoneme Classification. Progress on these tasks can be seen by looking at the online leaderboards (https://libribrain.com/). For example, in just two months F1 macro scores on the Speech

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

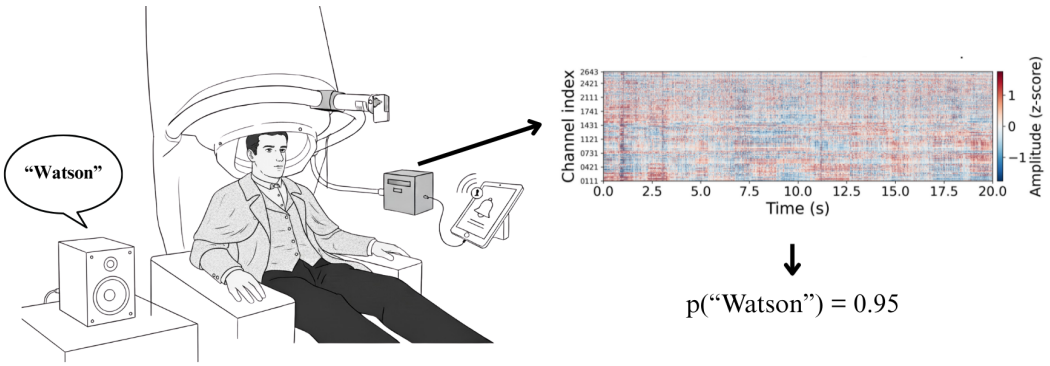

Figure 1: KWS task setup. The participant listens to a Sherlock Holmes audiobook while the model spots a chosen keyword (e.g., "Watson") from MEG signals.

Detection task advanced rapidly from about 68% to a new state-of-the-art on the (extended) public leaderboard of 96%. Our aim in this paper is to build on this success by introducing a new standardized task: Word Detection (a.k.a. Keyword Spotting). Substantively, we provide the same kinds of supporting infrastructure for this task (e.g., data loader, reference model, reproducible metrics, public leaderboard) which led directly to accelerated improvements in Speech Detection and Phoneme Classification.

As Landau et al. [2025] explain, the two tasks in the 2025 PNPL Competition were selected for their simplicity. Over time, the idea was to increase the complexity, and utility, of benchmark decoding tasks. Keyword Spotting (KWS) is an exciting landmark as it represents the first decoding task on this curriculum with practical utility for BCIs. In the established domain of voice computing, Keyword Spotting is commonly used to detect "wake words" (e.g., "Hey Siri", "Alexa", "OK Google"). Wake words like these can be used to indicate that subsequent speech should be interpreted as a command, or they can be used themselves as commands. In the emerging domain of *brain* computing, even a single wake word (e.g., "help") could be profoundly meaningful to someone with severe paralysis. Only slightly further along the curriculum, a small set of working keywords (e.g., "hungry", "tired", "thirsty", "toilet", "pain") would transform their quality of life. The benchmark task established in this paper is intended, ultimately, to lead to such an outcome. For clarity, and to contrast the use of acoustic inputs, we use the term *Neural Keyword Spotting* to denote Keyword Spotting from continuous brain data, a task represented schematically in Figure 1.

2 Related Work

2.1 Existing Tasks for Non-Invasive Speech Decoding

As mentioned in the introduction, the introduction of a rigorous benchmark a recent development in non-invasive speech decoding. Here a *benchmark* is a set of resources used by the community to measure progress. Benchmarks go beyond typical published work by including the following to support the measurement of progress.

- **Standardized data**: Publicly available, with well-defined train and holdout splits.
- **Evaluation metrics**: Agreed measure of success with public leaderboard to track progress.
- **Reference model**: Reproducible implementation with open weights and training code.

Prior to the release of LibriBrain [Özdogan et al., 2025], together with a standard Python library (`pnpl`) for loading predefined data splits [Landau et al., 2025], a number of open datasets [e.g., Schoffelen et al., 2019, Nastase et al., 2022, Gwilliams et al., 2023] were starting to reappear across large-scale studies [e.g., Défossez et al., 2023, d'Ascoli et al., 2024, Ridge and Parker Jones, 2024, Jayalath et al., 2025b]. However, data splits were not generally replicated making it difficult to compare methods. The same model architectures and weights were neither generally shared nor used as baselines and there were no public leaderboards.

A rich set of speech decoding tasks have nonetheless emerged. These include the following tasks.

Brain-to-Text (B2T) takes variable-length neural sequences (e.g., EEG, MEG, fMRI) as input and outputs text transcripts, typically evaluated with word error rate (WER) or semantic similarity metrics such as the BERTScore [Zhang et al., 2020]. B2T is the analogue of ASR and represents a long-term goal, though its difficulty has motivated the development of what we call intermediate tasks, which lie between more foundational tasks like Speech Detection and full B2T. Non-invasive B2T has been explored with EEG and MEG [Duan et al., 2023, Jo et al., 2024, Yang et al., 2024b,c,a], with semantic metrics often reported in place of WER, although recent work shows that competitive WERs are beginning to be achievable non-invasively [Jayalath et al., 2025a]. In fMRI, the coarse temporal resolution makes word-level alignment unlikely, though remarkable paraphrases have been produced which retain some semantic similarities to the ground truth speech [e.g., Tang et al., 2023].

Word Classification uses fixed-length neural segments aligned to individual words, producing categorical labels from a closed vocabulary. A number of recent works have focused on vocabularies of 250 words [d'Ascoli et al., 2024, Özdogan et al., 2025, Jayalath et al., 2025a], though recent models can also impute out-of-vocabulary items with an external LLM [e.g., Anthropic, 2025] if the predicted word is "unknown" [Jayalath et al., 2025a].

Phoneme Classification operates on shorter neural segments aligned to phonemes, predicting categorical labels over the phoneme inventory [Özdogan et al., 2025, Landau et al., 2025]. Relatedly, Phonetic Feature Classification outputs binary labels for broader phonological features such as voicing. Phonetic features group together multiple phoneme classes (e.g., voiced /b, v, z/ vs. unvoiced /p, f, s/) and can therefore be more data-efficient [Gwilliams et al., 2022, Jayalath et al., 2025b].

Segment Identification is a matching task. Given paired speech and brain data (e.g., cut continuous data into 3 second segments), the task is to correctly match audio and brain segments [Défossez et al., 2023, Tang et al., 2023]. This task is only applicable when speech and brain data are temporally aligned, limiting its utility for BCIs.

Speech Detection works on potentially open-ended neural recording. The aim is to identify when subjects were processing speech. The use of the term *processing* here is deliberate, as subjects could for example be listening to speech [Özdogan et al., 2025, Landau et al., 2025] or speaking aloud [Dash et al., 2020]. There is a contrast between Speech Detection and Classification, though it is perhaps subtle. In Speech Classification, fixed-duration inputs are assigned to a class (e.g., speech or non-speech) [Jayalath et al., 2025b]. Speech Classification models can be repurposed for Detection by applying them in sliding windows; but the task definitions remain formally distinct.

2.2 Keyword Spotting

Keyword spotting in the traditional audio domain (also referred to as wake-word detection) is a mature, highly-imbalanced detection problem optimizsed for very low false-alarm (FA) rates at fixed recall. Early small-footprint CNN and CRNN systems established the modern operating regime (e.g., 0.5 FA/h at acceptable FRR) under tight on-device constraints [e.g., Sainath and Parada, 2015, Arík et al., 2017]. Large benchmarks like Speech Commands [Warden, 2018]) and efficient architectures like MatchboxNet [Maidina et al., 2020] and Keyword Transformer [Berg et al., 2021] further drove accuracy/latency trade-offs for embedded devices, while industrial deployments (e.g., Apple's "Hey Siri") codified evaluation practices around FA/h and user-centric thresholds [Apple Siri Team, 2017].

On the invasive brain side, Milsap et al. [2019] introduced neural KWS with ECoG, showing low-latency, high-specificity detection using matched-filter templates spanning motor and auditory speech representations. Recent intracortical studies push to large-vocabulary online decoding and inner-speech control, but their goals (continuous B2T, WER/CER) and signal quality differ materially from non-invasive KWS [Willett et al., 2023, Metzger et al., 2023, Kunz et al., 2025].

Non-invasive technologies (EEG/MEG) have dramatic benefits over surgical implants in terms of safety and scalability. The application of keyword spotting is motivated by two converging strands. First, segment identification decoders trained to predict self-supervised speech representations from brain signals reliably retrieve the matching few-second stimulus among large candidate sets and generalise across participants - evidence that non-invasive signals carry phonetic/lexical detail at the granularity needed for lexical identification [Défossez et al., 2023, d'Ascoli et al., 2024].

Second, converging MEG/EEG results show sensitivity to phoneme sequence structure and higher-level linguistic content, and recent deep models capture meaningful portions of the speech-to-language transform in these signals [Gwilliams et al., 2022, Tezcan et al., 2023, Desai et al., 2021]. Against this recent progress, LibriBrain allows testing whether the same brain-speech representations that enable segment retrieval also support lexical selectivity for pre-specified words in its long-form, naturalistic stories. Due to its larger scale, it also allows building on prior EEG-based KWS pilots, which have largely remained at small-lexicon trialwise classification/onset detection [Sakthi et al., 2021].

3 Methods

3.1 Dataset

The following summary closely follows the original LibriBrain description. For full details, see [Özdogan et al., 2025]. In brief, the dataset covers over 52 hours of within-subject MEG data recorded on a 306-channel MEGIN Triux™ Neo system (102 magnetometers, 204 planar gradiometers). Recordings were acquired at 1 kHz and minimally preprocessed (head-motion correction; Maxwell filter; 50/100 Hz notch filter; 0.1-125 Hz band-pass filter) before downsampling to 250 Hz (4 ms samples), yielding data of shape $C \times T$ with C=306. Each session is paired with an `events.tsv` file listing onset/duration (s) for speech, word, and phoneme segments, all produced by forced-alignment [Ochshorn and Hawkins, 2015] and then manually corrected.

The release spans 93 sessions (3,139 min; 52.32 h) with 466,230 word tokens (16,892 unique) and 1,511,732 phoneme tokens. See Figures 6 and 7 for an overview of the dataset. Word frequencies are Zipfian, providing keywords across a wide base-rate spectrum (short, frequent function words vs. longer, rarer content/proper names). These properties suit event-referenced keyword detection with extreme class imbalance.

3.2 Task Definition

We cast neural keyword spotting (KWS) from MEG as an event-referenced detection task using LibriBrain word onsets [Özdogan et al., 2025]. This can be formalized as follows: First, we fix a small keyword set $\mathcal{V}$ (minimally $|\mathcal{V}| = 1$). For each keyword $k \in \mathcal{V}$, let $d_{\max}(k)$ be the maximum duration of any instance of k in the corpus. Given that we may want to extract brain recordings that start and end before and after audio event boundaries (e.g., because neural processing continues after the presentation of a stimulus), we can select fixed pre/post buffers $\beta^- \geq 0$ and $\beta^+ \geq 0$. These offsets can then be used to define $D(\mathcal{V})$, which is the total window duration (in seconds) of neural data to extract around any keyword in $\mathcal{V}$:

$$D(\mathcal{V}) = \beta^- + \max_{k \in \mathcal{V}} d_{\max}(k) + \beta^+.$$

Concretely, for each word token with onset t_i and string s_i, we extract a $306 \times D(\mathcal{V})$ window starting at $t_i - \beta^-$. This guarantees that any instance of any $k \in \mathcal{V}$ fits fully inside the window while allowing for a longer window duration if further context can improve detection. The binary label is

$$y_i \;=\; \mathbb{1}\{\, s_i \in \mathcal{V} \,\} \in \{0, 1\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

In contrast to KWS, full-vocabulary Brain-to-Text aims to identify $w \in \mathcal{W}$ (hundreds of thousands of words), which introduces severe long-tail sparsity and requires calibrating thresholds across many classes. KWS is a practical, fixed-lexicon task: it asks only whether any member of a small, predefined set $\mathcal{V}$ occurred. This offers a simple, reliable trigger with clean control over latency and false alarms—useful both on its own (assistive or hands-free commands) and as a stepping stone towards richer decoders.

By default, we adopt LibriBrain's session-level train/val/test split [Özdogan et al., 2025]. For a chosen set of keywords $\mathcal{V}$, we verify that positives occur in both validation and test. If not, we replace them with the two sessions containing the most positives for $\mathcal{V}$. This ensures sufficient positive examples for reliable metric computation, particularly important given the extreme class imbalance in keyword detection. For a session S containing word tokens with strings $\{s_i\}_{i=1}^{n_S}$, let

$$c_S = \sum_{i=1}^{n_S} \mathbb{1}\{s_i \in \mathcal{V}\}$$

be the count of keyword instances in session S. Validation and test are set to the two sessions maximizing c_S.

3.3 Metrics

We use area under the precision-recall curve (AUPRC) as the primary metric. This is well suited to keyword spotting because its base rate equals the empirical prevalence of positives, so gains are easy to interpret. It also summarises the trade-off we actually care about under heavy imbalance—how many of the system's alarms are correct (precision) as we demand more coverage (recall). Finally, unlike alternative metrics (e.g., AUROC), AUPRC is not overly optimistic when precision is too low to be usable. We supplement AUPRC with additional metrics like AUROC to provide a comprehensive view. For scenario-grounded reporting, we translate any validation-selected threshold (precision P, recall R) into hourly rates under an assumed event frequency λ (keywords/hour):

$$\text{FA/h} = R\,\lambda\Big(\tfrac{1}{P} - 1\Big), \qquad \text{Misses/h} = \lambda(1-R), \qquad \text{Detections/h} = \lambda R.$$

We consider two illustrative use cases: assistive access ($\lambda \approx 2$/h) and hands-free control ($\lambda \approx 10$/h). Thresholds τ are chosen on validation either (a) by maximising recall under a false-alarm budget, or (b) by minimising FA/h subject to a target recall; selected τ are then frozen for testing. For clarity, FA/h can also be computed directly from test-set coverage (false positives per hour of labelled windows); unless otherwise noted, we report the scenario-translated FA/h using (P, R, λ).

3.4 Reference Model

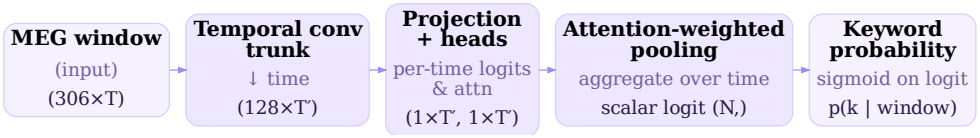

Figure 2: Reference model overview. A $306 \times T$ MEG window passes through a temporal convolutional trunk (with time downsampling) to produce a $128 \times T'$ representation. A projection stage feeds two temporal heads that emit per-time logits and attention scores. An attention-weighted pooling aggregates over time to a scalar logit, which is mapped to the keyword probability via a sigmoid.

Our reference system ingests 306-channel MEG windows of length T and processes them with a compact temporal convolutional trunk that includes a residual block and a time-downsampling layer, yielding a $128 \times T'$ representation (temporal CNNs are strong sequence/biosignal decoders; residual connections stabilize deeper stacks and enlarge receptive fields efficiently [Bai et al., 2018, Schirrmeister et al., 2017, Lawhern et al., 2018, He et al., 2016]). A projection stage produces a 512-channel sequence, from which two 1×1 temporal heads compute (i) per-time logits and (ii) attention scores normalised along time. The final output is a scalar logit obtained by attention-weighted summation of the per-time logits (a learned MIL-style pooling well-suited to brief events within longer windows [Ilse et al., 2018, Kong et al., 2020, McFee et al., 2018]; in MEG, this lets the model emphasise time-locked acoustic/lexical responses such as M100/N400 components [Gage et al., 1998, Halgren et al., 2002, Hari and Salmelin, 2012]). Training uses focal loss with a small pairwise ranking term: focal down-weights abundant easy negatives and focuses gradient on rare, hard positives under extreme imbalance [Lin et al., 2017], while the pairwise (logistic) ranking aux loss encourages correct ordering of positives above negatives, supporting PR/Average-Precision-aligned selection [Burges, 2010, Yue et al., 2007, Davis and Goadrich, 2006, Saito and Rehmsmeier, 2015]. Batches are class-balanced by oversampling positives, and we apply light temporal jitter and additive noise (both standard, effective regularizers for EEG/MEG time-series [Buda et al., 2018, Lashgari et al., 2020, He et al., 2021, Rommel et al., 2022]). We optimise with AdamW [Loshchilov and Hutter, 2019] and select checkpoints by validation AUPRC (preferred under heavy class imbalance [Saito and Rehmsmeier, 2015, Davis and Goadrich, 2006]).

4 Results

Where possible, all experiments use the standard train/validation/test splits provided by the `pnpl` dataset (using the logic described in Section 3.4) and are fully reproducible (see Appendix A). Unless noted, values are seed-averages over three runs. Error bars are standard errors across seeds. For Table 1 we report standard errors approximated from 95% bootstrap CIs (4,000 resamples).

4.1 Model Performance

We first establish that the dataset carries usable signal for keyword detection by evaluating a single model on the held-out test set ($n = 4660$, positives $= 24$; base rate $= 0.00515$). Given the absence of prior publicly reproducible MEG keyword spotting benchmarks, we evaluate against permutation-derived random baselines to demonstrate the presence of meaningful signal rather than competitive performance. While overall performance is modest, threshold-free metrics indicate clear signal (AUPRC $\approx 13.4\times$ the permutation baseline; AUROC ≈ 0.80). Full results are provided in Table 1. Beyond threshold-free metrics, we include an operational snapshot. At a target recall of ~ 0.10, the scenario-translated FA/h for the assistive case ($\lambda = 2$/h) is ~ 2.19 (SE ≈ 1.63), corresponding to ~ 13 alerts per correct detection. For reference, directly counting false positives per hour under the labelled test coverage yields $\sim 16.3\,\mathrm{FA/h}$ (seed-avg; SE ≈ 12.1). Under FA/h budgets (scenario scale, $\lambda = 2$/h), the model achieves recall ~ 0.14 at 2.0 FA/h and ~ 0.08 at 0.5 FA/h (seed-averaged). These numbers contextualise the ranking metrics and set a baseline for future improvements. The right panel of Fig. 4 shows the mean recall–FA/h operating curve with per-seed traces. For this snapshot, operating points are chosen on the test PR curves for presentation; in deployment we would select thresholds on validation and freeze them before testing.

Metric	**Baseline**	**Model** ($\pm$ SE)	**% improvement**	**p-value**
F1	0.010	**0.107** $\pm$ 0.038	+970%	$\approx 1.00 \times 10^{-5}$
F1-Macro	0.431	**0.542** $\pm$ 0.028	+25.8%	$\approx 1.00 \times 10^{-5}$
Accuracy	**0.995**	0.955 $\pm$ 0.033	-4.0%	n.s.
MCC	0.000	**0.119** $\pm$ 0.027	n/a	$\approx 1.00 \times 10^{-5}$
AUROC	0.500	**0.804** $\pm$ 0.017	+60.8%	$\approx 2.00 \times 10^{-5}$
AUPRC	0.007	**0.094** $\pm$ 0.032	+1243%	$\approx 2.00 \times 10^{-5}$

Table 1: Performance compared to random baselines derived from permutation nulls. Thresholded metrics use threshold $\tau = 0.5$. Standard errors are approximated from the 95% bootstrap CIs via normality (SE $\approx (\mathrm{CI_{hi}} - \mathrm{CI_{lo}})/3.92$).

4.2 Keyword Choice

An important consideration in KWS is the choise of keyword(s). In LibriBrain, as in many real-world corpora, longer words are rarer: word length in phonemes is negatively correlated with token frequency (Spearman $r = -0.28$ with $\log$ frequency; $p = 2.7 \times 10^{-185}$; Fig. 3 left). This matters because length can reduce false alarms while frequency controls how many positives we can realistically train on. To navigate this trade-off, we selected the most frequent word at each phoneme length and measured %ΔAUPRC over the empirical base rate. The length-%ΔAUPRC relation is non-monotonic (Fig. 3 right): among a 12-item shortlist spanning 1-12 phonemes, the 5-phoneme *watson* yields the largest %ΔAUPRC, whereas several longer items (e.g., 9-12 phonemes) underperform despite greater duration.

A controlled comparison across three similarly frequent keywords of different lengths (*walk*, *surely*, *excellent*; 3/5/8 phonemes) shows no detectable difference in %ΔAUPRC within our precision (overlapping SEMs; Fig. 4). This indicates that, once frequency is matched, mere length is not the primary driver of detectability in MEG KWS. The pattern is consistent with established constraints on neural speech processing and KWS: benefits accrue less from duration per se and more from properties that improve time-locking and reduce lexical competition—salient acoustic onsets and early stress (stronger M100/M200), an early uniqueness point (UP)—i.e., the keyword becomes lexically unique after only a few initial phonemes (a small UP index relative to its length)—a sparse phonological neighborhood, and moderate frequency with lower contextual predictability [Gage et al., 1998, Leminen et al., 2011, Vitevitch and Luce, 1999, Halgren et al., 2002, Chen et al., 2014].

Empirically, *watson* may profit from prosodic prominence in narrative speech and an early uniqueness point, outweighing any gains attributable to length alone; *watson* may also benefit from attentional saliency, being a word that the subject consistently paid attention to.

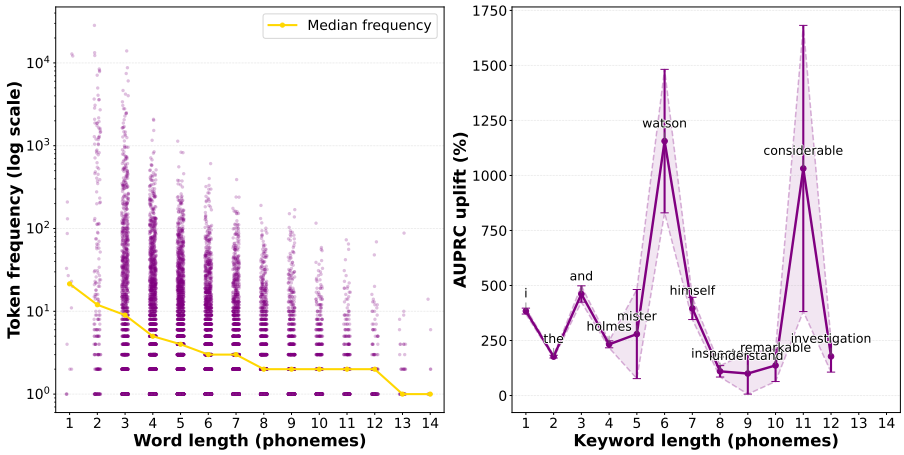

Figure 3: Keyword choice trade-offs. (Left) Relationship between word length in phonemes and token frequency across the full LibriBrain corpus (points are unique words; y-axis is log-scaled for readability; the line shows the median frequency per length). (Right) %ΔAUPRC over the base rate for the shortlisted keywords as a function of their phoneme count.

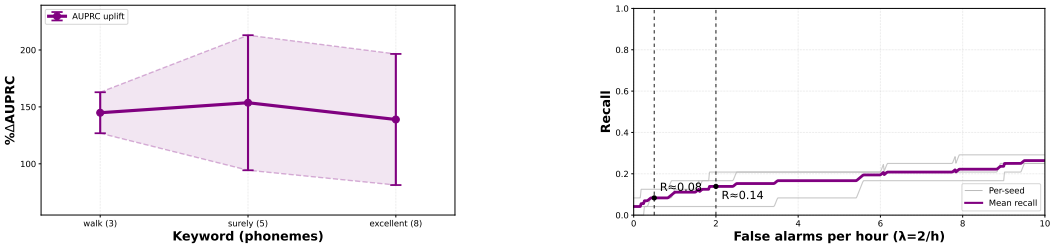

Figure 4: Left: %ΔAUPRC over the empirical base rate for three similarly frequent keywords with 3, 5, and 8 phonemes, showing no significant difference. Right: Recall–FA/h operating curve (assistive, $\lambda = 2$/h) with budget markers at 0.5 and 2.0 FA/h.

4.3 Data Scaling

To understand how keyword detection performance scales with available training data, we systematically varied the fraction of the 52-hour corpus used for training while keeping the validation and test sets fixed. For these scaling runs we used 0 s pre-onset and +0.25 s post-onset windows (per-instance window length of 1.05 s). Because training uses many overlapping windows around labelled events, the total windowed duration processed exceeds the 52 h of unique recordings: at 10% this corresponds to ≈14 h of windowed data (≈5.2 h unique), and at 100% to ≈143 h of windowed data.

As shown in Figure 5, AUPRC improves approximately log-linearly as we increase the training fraction from 10% to 100%, consistent with established within-subject scaling laws in neural decoding [d'Ascoli et al., 2024, Sato et al., 2024]. Notably, even with just 10% of the training data (≈14 h windowed; ≈5.2 h unique), the model achieves meaningful performance above chance, suggesting that keyword detection remains feasible even in scenarios with limited recording time. Permutation tests confirm that AUPRC is not above chance at 5% ($p = 0.108$), but is already significant at 10% ($p = 0.0156$; one-sided, 10,000 draws), and remains strongly significant thereafter (20% $p = 6.0 \times 10^{-4}$; 40–100% $p \leq 2 \times 10^{-4}$).

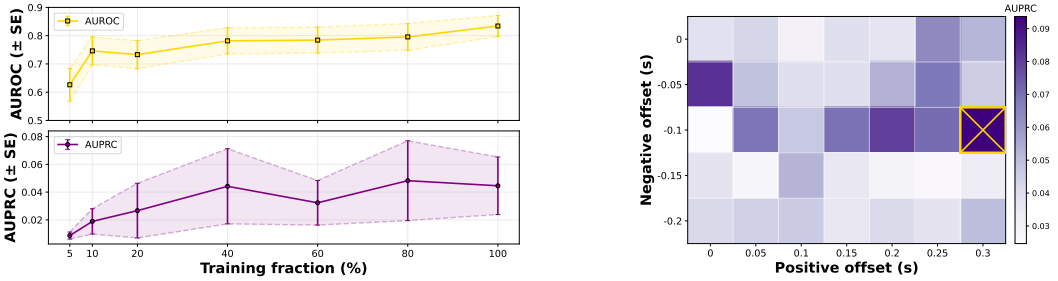

Figure 5: Left: Keyword spotting scales with the amount of training data; AUPRC improves roughly log-linearly as a larger fraction of the 52-hour corpus is used. Right: Effect of temporal offsets around the keyword onset. The X marks where AUPRC peaks with a modest pre-onset context ($\sim$0.1s) and a slightly longer post-onset window ($\sim$0.3s).

4.4 Sample Length

Adding pre- and post-onset offsets modestly improves detection. Averaged over all non-zero offsets, AUPRC increases by $\sim 25\%$ relative to the 0/0 baseline (absolute +0.0097). This improvement is statistically significant (paired per-seed mean +0.0099 AUPRC, SE 0.0028; 95% CI [0.0045, 0.0154]; one-sided $p < 10^{-3}$). We observe the best AUPRC near (neg=0.1s, pos=0.3s), but no clear monotonic trend across offsets; performance is relatively flat in a small neighbourhood around this setting and declines for very short or overly long windows, suggesting a bias–variance trade-off between providing sufficient context and diluting signal with unrelated activity.

5 Conclusion

We introduce a reproducible MEG keyword-spotting task on LibriBrain, demonstrate meaningful signal, and release task specifications, a modified `pnpl` library, baseline model, and tutorial materials.

5.1 Practical Utility

Despite the achieved improvement in metrics, performance is not yet sufficient for reliable hands-free use. In an assistive scenario (λ=2 h^{-1}), at recall $\approx$0.10 the system yields $\approx$2.2 false alarms per hour (about 13 alerts per correct detection). Priorities for future work thus include: (i) stronger ranking, (ii) calibration and principled threshold selection, and (iii) deployment strategies that suppress false alarms (multi-confirmation, small ensembles, cascaded detectors with context-aware priors).

5.2 Limitations & Future Work

Limitations. This study reports results for a single participant on a single corpus. Generalisability across participants remains an open question for Neural Keyword Spotting, though generalization is becoming less of a problem in decoding than it was [Csaky et al., 2023, Défossez et al., 2023, d'Ascoli et al., 2024, Jayalath et al., 2025b,a]. Validation and test sessions were selected to maximise positives, stabilising metrics at the expense of a mild base-rate bias. Test sets contain few positives, so thresholded metrics carry substantial uncertainty. We evaluate a compact model only, without exploring richer encoders, self-supervised pretraining, streaming inference, or multi-keyword training. Finally, we use event-referenced windows; continuous-stream detection with latency and explicit false-alarm accounting remains open.

Future work. Building on the success of competitions like the 2025 LibriBrain Competition [Landau et al., 2025] or Brain-to-text '25 [Card et al., 2025], we will release a leaderboard system for the keyword detection task as part of a larger-scale competition later this year. We also plan to extend this work to additional datasets, including inner speech and multi-subject recordings. Further analysis, such as phoneme-informed keyword selection and a deeper examination of temporal offsets, should clarify where gains are available building on the exploratory results presented here.

Acknowledgments and Disclosure of Funding

We would like to thank to the organizers and anonymous reviewers for their efforts and constructive feedback. We also acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558) and the NVIDIA Corporation for contributing 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).

References

Anthropic. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, 2025. Accessed: 2025-04-22.

Apple Siri Team. Hey Siri: An on-device DNN-powered voice trigger for Apple's personal assistant. Apple Machine Learning Journal, 2017. URL https://machinelearning.apple.com/research/hey-siri.

Sercan Ö. Arík, Markus Kliegl, Rewon Child, Joel Hestness, Andrew Gibiansky, Chris Fougner, Ryan Prenger, and Adam Coates. Convolutional recurrent neural networks for small-footprint keyword spotting. In *Interspeech*, pages 1606–1610, 2017. doi: 10.21437/Interspeech.2017-1737.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Axel Berg, Mark O'Connor, and Miguel Tairum Cruz. Keyword transformer: Auditory attention for small-footprint keyword spotting. *arXiv preprint arXiv:2104.00769*, 2021.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. doi: 10.1016/j.neunet.2018.07.011.

Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

Nicholas Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. Brain-to-text '25. Kaggle, 2025. URL https://kaggle.com/competitions/brain-to-text-25.

Guoguo Chen, Carolina Parada, and Georg Heigold. Small-footprint keyword spotting using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4087–4091, 2014. doi: 10.1109/ICASSP.2014.6854370.

Richard Csaky, Mats W. J. van Es, Oiwi Parker Jones, and Mark Woolrich. Group-level brain decoding with deep learning. *Human Brain Mapping*, 44:6105–6119, 2023. doi: 10.1002/hbm.26500. URL https://doi.org/10.1002/hbm.26500.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Decoding individual words from non-invasive brain recordings across 723 participants. *arXiv preprint arXiv:2412.17829*, 2024.

Debadatta Dash, Paul Ferrari, Satwik Dutta, and Jun Wang. Neurovad: Real-time voice activity detection from non-invasive neuromagnetic signals. *Sensors*, 20(8):2248, 2020. doi: 10.3390/s20082248.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 233–240, 2006. doi: 10.1145/1143844.1143874.

Maansi Desai, Jade Holder, Cassandra Villarreal, Nat Clark, Brittany Hoang, and Liberty S Hamilton. Generalizable EEG encoding models with naturalistic audiovisual stimuli. *Journal of Neuroscience*, 41(43):8946–8962, 2021.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete EEG waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*, 2023.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain activity with self-supervised learning. *Nature Machine Intelligence*, 5(11):1262–1276, 2023. doi: 10.1038/s42256-023-00732-3.

Nicole Gage, David Poeppel, Timothy P. L. Roberts, and Gregory Hickok. Auditory evoked m100 reflects onset acoustics of speech sounds. *Brain Research*, 814:236–239, 1998. doi: 10.1016/S0006-8993(98)01058-0.

Laura Gwilliams, Jean-Rémi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, 13(1):6606, 2022. doi: 10.1038/s41467-022-34326-1.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. MEG-MASC: A magnetoencephalography dataset for naturalistic language comprehension. *Scientific Data*, 10(1):1–16, 2023. doi: 10.1038/s41597-023-02170-x.

Eric Halgren, Rupali Dhond, Niels Christensen, Cyma Van Petten, Ksenija Marinkovic, Jeffrey Lewine, and Anders M. Dale. N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage*, 17(3):1101–1116, 2002. doi: 10.1006/nimg.2002.1268.

Riitta Hari and Riitta Salmelin. Magnetoencephalography: From SQUIDs to neuroscience. *NeuroImage*, 61(2):386–396, 2012. doi: 10.1016/j.neuroimage.2011.11.074.

Chao He, Jialu Liu, Yuesheng Zhu, and Wencai Du. Data augmentation for deep neural networks model in EEG classification task: A review. *Frontiers in Human Neuroscience*, 15:765525, 2021. doi: 10.3389/fnhum.2021.765525.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Dulhan Jayalath, Gilad Landau, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text. *arXiv preprint arXiv:2505.13446*, 2025a.

Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The Brain's Bitter Lesson: Scaling speech decoding with self-supervised learning. *International Conference on Machine Learning (ICML)*, 2025b.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are EEG-to-text models working? *arXiv preprint*, 2024. URL https://arxiv.org/abs/2405.06459.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Erin M. Kunz, Benyamin Abramovich Krasa, Foram Kamdar, Donald T. Avansino, Nick Hahn, Seonghyun Yoon, Akansha Singh, Samuel R. Nason-Tomaszewski, Nicholas S. Card, Justin J. Jude, Brandon G. Jacques, Payton H. Bechefsky, Carrina Iacobacci, Leigh R. Hochberg, Daniel B. Rubin, Ziv M. Williams, David M. Brandman, Sergey D. Stavisky, Nicholas AuYong, Chethan Pandarinath, Shaul Druckmann, Jaimie M. Henderson, and Francis R. Willett. Inner speech in motor cortex and implications for speech neuroprostheses. *Cell*, 188(17):4658–4673.e17, 2025. doi: 10.1016/j.cell.2025.06.015.

Gilad Landau, Miran Özdogan, Gereon Elvers, Francesco Mantegna, Pratik Somaiya, Dulhan Jayalath, Luisa Kurth, Teyun Kwon, Brendan Shillingford, Greg Farquhar, Minqi Jiang, Karim Jerbi, Hamza Abdelhedi, Yorguin Mantilla Ramos, Caglar Gulcehre, Mark Woolrich, Natalie Voets, and Oiwi Parker Jones. The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. *NeurIPS, Competition Track*, 2025. https://arxiv.org/abs/2506.10165.

Elnaz Lashgari, Dehua Liang, and Uri Maoz. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346:108885, 2020. doi: 10.1016/j.jneumeth.2020.108885.

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

Aleksi Leminen, Mikko Leminen, Teija Kujala, and Yury Shtyrov. A combined EEG and MEG study of spoken word recognition time-locked to the uniqueness point. *Frontiers in Human Neuroscience*, 5:66, 2011. doi: 10.3389/fnhum.2011.00066.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Asmaa Maidina et al. Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition. *arXiv preprint arXiv:2004.08531*, 2020.

Brian McFee, Vincent Lostanlen, Justin Salamon, Mark Cartwright, and Juan Pablo Bello. Adaptive pooling operators for weakly labeled sound event detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018. doi: 10.1109/ICASSP.2018.8462220.

Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023. doi: 10.1038/s41586-023-06443-4.

Griffin Milsap, Maxwell Collard, Christopher Coogan, Qinwan Rabbani, Yujing Wang, and Nathan E. Crone. Keyword spotting using human electrocorticographic recordings. *Frontiers in Neuroscience*, 13:60, 2019. doi: 10.3389/fnins.2019.00060.

Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, Gina Choe, Ariel Goldstein, Tamara Vanderwal, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. The narratives collection: Shared annotated datasets for naturalistic language processing in the neuroimaging of story listening. *Scientific Data*, 9(1):1–23, 2022. doi: 10.1038/s41597-022-01735-6.

Robert M. Ochshorn and Max Hawkins. Gentle: A robust yet lenient forced aligner built on Kaldi, 2015. URL https://lowerquality.com/gentle/.

Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. *NeurIPS, Datasets & Benchmarks Track*, 2025. URL https://arxiv.org/abs/2506.02098.

Jeremy Ridge and Oiwi Parker Jones. Resolving domain shift for representations of speech in non-invasive brain recordings. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2410.19986.

Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort. Data augmentation for learning predictive models on EEG: a systematic comparison. *Journal of Neural Engineering*, 19 (6), 2022. doi: 10.1088/1741-2552/aca220.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Tara N. Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *Interspeech*, pages 1478–1482, 2015. doi: 10.21437/Interspeech.2015-352.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Madhumitha Sakthi, Maansi Desai, Liberty Hamilton, and Ahmed Tewfik. Keyword-spotting and speech onset detection in EEG-based brain computer interfaces. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 519–522. IEEE, 2021.

Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of EEG data. *arXiv preprint arXiv:2407.07595*, 2024.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730.

Jan-Mathijs Schoffelen et al. A 204-subject multimodal human neuroimaging dataset to study language processing. *Scientific Data*, 6(1):1–13, 2019. doi: 10.1038/s41597-019-0020-y.

Jerry Tang, Alexander LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. doi: 10.1038/s41593-023-01304-9.

Filiz Tezcan, Hugo Weissbart, and Andrea E. Martin. A tradeoff between acoustic and linguistic feature encoding of continuous speech in the human brain. *eLife*, 12:e82386, 2023. doi: 10.7554/eLife.82386.

Michael S. Vitevitch and Paul A. Luce. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3):374–408, 1999. doi: 10.1006/jmla.1998.2618.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Robert Weide. The CMU pronouncing dictionary, 1998. URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict. release 0.6.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620 (7976):1031–1036, 2023. doi: 10.1038/s41586-023-06377-x.

Yiqian Yang, Yiqun Duan, Hyejeong Jo, Qiang Zhang, Renjing Xu, Oiwi Parker Jones, Xuming Hu, Chin-teng Lin, and Hui Xiong. NeuGPT: Unified multi-modal neural GPT. *arXiv preprint arXiv:2410.20916*, 2024a.

Yiqian Yang, Yiqun Duan, Qiang Zhang, Renjing Xu, and Hui Xiong. NeuSpeech: Decode neural signal as speech. *arXiv preprint*, 2024b. URL https://arxiv.org/abs/2403.01748.

Yiqian Yang, Hyejeong Jo, Yiqun Duan, Qiang Zhang, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. MAD: Multi-alignment MEG-to-text decoding. *arXiv preprint arXiv:2406.01512*, 2024c.

Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–278, Amsterdam, The Netherlands, 2007. ACM. doi: 10.1145/1277741.1277790.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A Open Resources: Code, Tutorial, and Leaderboard

A.1 Updated `pnpl` datasets

We release an updated version of the open source `pnpl` library (Özdogan et al. [2025]) to support word-level tasks. This allows both full signal-to-word and single/multi-keyword tasks to be performed using similar syntax to the existing `LibriBrainSpeech` and `LibriBrainPhoneme` classes:

Word-level task:

```
from pnpl.datasets import LibriBrainWord

dataset = LibriBrainWord(
    data_path="./data/",
    partition="train",
    tmin=0.0,
    tmax=0.8,
)
```

Single-keyword task:

```
from pnpl.datasets import LibriBrainWord

dataset = LibriBrainWord(
    data_path="./data/",
    partition="train",
    keyword_detection="watson",
)
```

Multi-keyword task:

```
from pnpl.datasets import LibriBrainWord

dataset = LibriBrainWord(
    data_path="./data/",
    partition="train",
    keyword_detection=["sherlock", "holmes"],
)
```

For the single- or multi-keyword task, sample length is inferred from the longest keyword duration and can be extended with the `positive_buffer` and `negative_buffer` arguments. Overwrites using `tmin` and `tmax` are of course possible. For full signal-to-word, that is rarely the intended behaviour, so these options are disabled and a reasonable default is used instead.

Similarly, the `keyword_detection` variant will verify that the keyword(s) are present in the dataset and, if not, default to highest prevalence sessions as validation and test sets, while the signal-to-word variant will use the default validation and test sets.

The library is available on PyPI[1] and on GitHub[2].

[1] https://pypi.org/project/pnpl/

[2] https://github.com/neural-processing-lab/pnpl

A.2 Tutorial Notebook

To encourage further exploration within the community, we also release a tutorial in the format of a Jupyter Notebook. Within the compute limits of the Colab Free Tier (`T4` GPU), the notebook allows for training a model around 10% of the LibriBrain dataset, reaching significantly above chance performance in under 30 minutes. The notebook is available in the `tutorial` folder of the `keyword-experiments` repository[3].

A.3 Experiment Code

Finally, to allow for full reproducibility, we release the code for the experiments and analysis conducted in this paper. The code is available in the `experiments` folder of the `keyword-experiments` repository[4].

B Dataset Figures

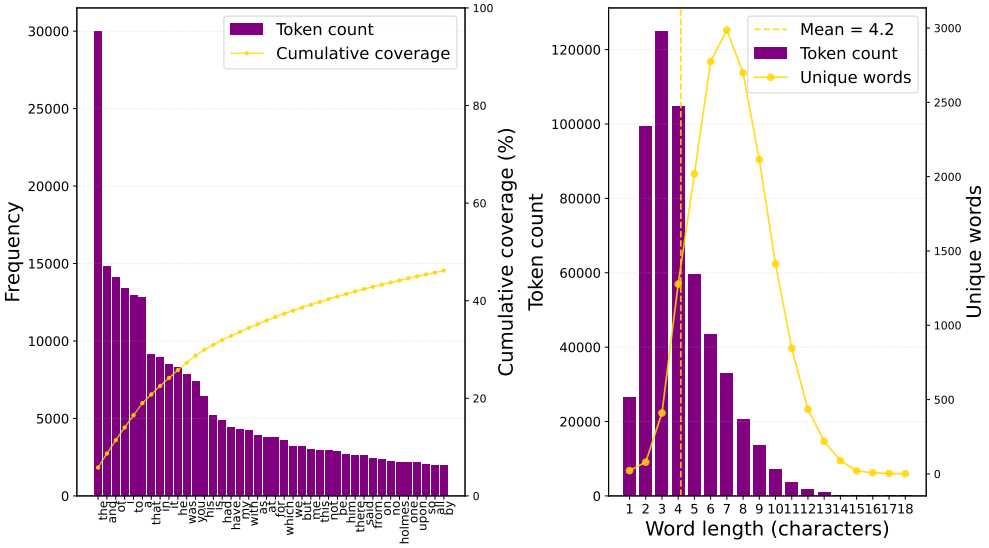

Figure 6: Overview of the LibriBrain dataset. (A) 40 most common words and their coverage of the dataset. (B) Word length distribution and the number of unique words for each length.

C Dataset Tables

C.1 Data Scaling

[3] http://github.com/neural-processing-lab/libribrain-keyword-experiments
[4] http://github.com/neural-processing-lab/libribrain-keyword-experiments

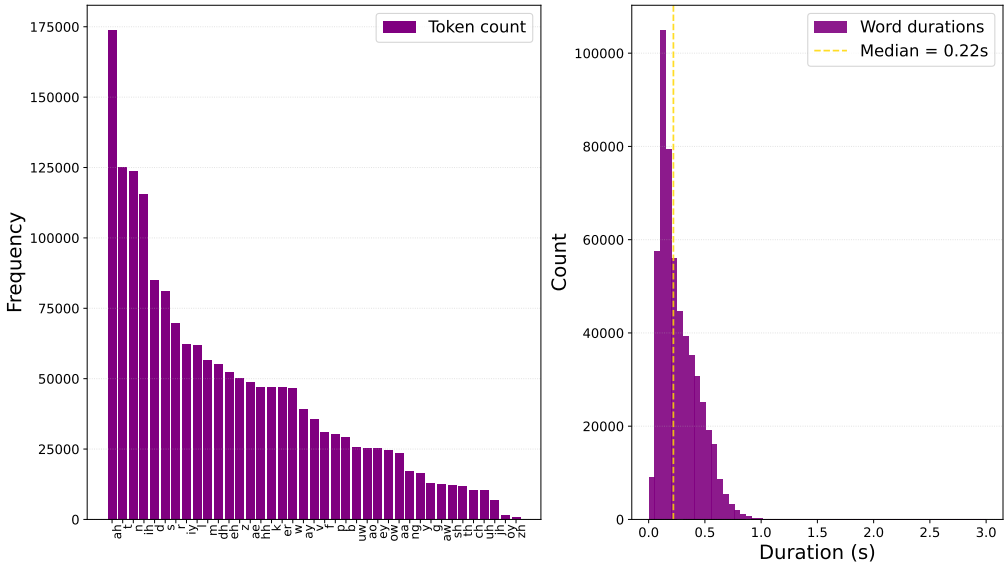

Figure 7: (C) Phoneme distribution across the corpus (39 ARPAbet phonemes from [Weide, 1998]). (D) Word duration distribution with median line.

Training fraction	**AUPRC** ($\pm$ SE)	**AUROC** ($\pm$ SE)	**AUPRC p-value**	**AUPRC/base rate** ($\times$)
5%	0.009 ± 0.003	0.626 ± 0.058	0.108	1.69
10%	0.019 ± 0.009	0.746 ± 0.049	0.0156	3.67
20%	0.027 ± 0.020	0.733 ± 0.050	6.0×10^{-4}	5.18
40%	0.044 ± 0.027	0.782 ± 0.046	$< 5 \times 10^{-5}$	8.59
60%	0.032 ± 0.016	0.784 ± 0.046	2.0×10^{-4}	6.28
80%	0.048 ± 0.029	0.796 ± 0.047	$< 5 \times 10^{-5}$	9.37
100%	0.045 ± 0.021	0.834 ± 0.037	$< 5 \times 10^{-5}$	8.66

Table 2: Detailed scaling results for keyword detection across training fractions (seed-averaged over three runs). Standard errors are approximated from 95% bootstrap CIs as $\mathrm{SE} \approx (\mathrm{CI}_{\mathrm{hi}} - \mathrm{CI}_{\mathrm{lo}})/3.92$. P-values are from one-sided permutation tests of the seed-average AUPRC against the null. The base rate for the fixed test set is 0.00515.

C.2 Keyword Choice

Keyword	**Base rate**	**AUPRC**	**AUROC**	**Acc**	**Best F1**
and	0.039	0.218 ± 0.014	$\mathbf{0.825} \pm 0.002$	0.756 ± 0.010	$\mathbf{0.292} \pm 0.008$
the	0.077	0.213 ± 0.005	0.728 ± 0.006	0.673 ± 0.004	0.278 ± 0.006
i	0.039	0.191 ± 0.005	0.784 ± 0.004	0.708 ± 0.015	0.279 ± 0.006
watson	0.005	$\mathbf{0.065} \pm 0.017$	0.759 ± 0.006	0.952 ± 0.034	0.149 ± 0.036
holmes	0.008	0.028 ± 0.001	0.791 ± 0.013	0.820 ± 0.066	0.072 ± 0.005
himself	0.003	0.013 ± 0.001	0.758 ± 0.021	0.993 ± 0.001	0.062 ± 0.008
considerable	0.001	0.012 ± 0.007	0.684 ± 0.066	0.997 ± 0.002	0.041 ± 0.025
inspector	0.004	0.008 ± 0.001	0.593 ± 0.020	0.994 ± 0.001	0.040 ± 0.012
mister	0.002	0.007 ± 0.004	0.505 ± 0.051	0.998 ± 0.000	0.054 ± 0.022
remarkable	0.001	0.003 ± 0.001	0.658 ± 0.070	0.991 ± 0.008	0.010 ± 0.004
understand	0.001	0.002 ± 0.001	0.523 ± 0.136	$\mathbf{0.999} \pm 0.000$	0.006 ± 0.003
investigation	0.001	0.002 ± 0.001	0.665 ± 0.047	0.998 ± 0.001	0.011 ± 0.007

Table 3: Seed-averaged per-keyword metrics (absolute units): base rate (positive prevalence), AUPRC, AUROC, Accuracy, and Best F1 (per-seed best across thresholds). Means and $\pm$ SEM are computed across seeds. Bold marks the best improvement vs base rate for AUPRC, and the highest mean for other columns.

C.3 Operating Points

Scenario	Metric	Value	SE
Assistive ($\lambda = 2$/h), target recall ≈ 0.10	FA/h	2.194	1.629
Assistive ($\lambda = 2$/h), FA/h budget 2.0	Recall	0.139	0.050
Assistive ($\lambda = 2$/h), FA/h budget 0.5	Recall	0.083	0.024
Labelled test coverage	FP/h	16.3	12.1

Table 4: Operating-point snapshot (best-AUPRC buffer: neg=0.1s, pos=0.3s). Values are seed-averages $\pm$ SE ($n = 3$). Scenario-scale metrics use the assistive case ($\lambda = 2$/h).

C.4 Keyword Length (matched frequency)

Keyword	Chars	Base rate	AUPRC	%ΔAUPRC over base
walk	4	0.00056	0.00136 ± 0.00010	144.9 ± 18.0
surely	6	0.00056	0.00141 ± 0.00031	153.7 ± 59.3
excellent	9	0.00056	0.00133 ± 0.00033	138.9 ± 57.6

Table 5: Matched-frequency keyword comparison (seed-averaged over three runs). Results show no significant differences (overlapping SEMs).

C.5 Temporal Offsets

Neg [s]	Pos [s]	AUPRC (mean $\pm$ SE)	AUROC (mean $\pm$ SE)	Seeds
0.00	0.00	0.039 ± 0.009	0.811 ± 0.011	3
0.00	0.05	0.043 ± 0.003	0.813 ± 0.008	3
0.00	0.10	0.030 ± 0.007	0.779 ± 0.015	3
0.00	0.15	0.039 ± 0.009	0.799 ± 0.021	3
0.00	0.20	0.037 ± 0.003	0.781 ± 0.011	3
0.00	0.25	0.064 ± 0.007	0.820 ± 0.002	3
0.00	0.30	0.052 ± 0.006	0.821 ± 0.007	3
0.05	0.00	0.083 ± 0.029	0.799 ± 0.025	3
0.05	0.05	0.049 ± 0.010	0.781 ± 0.010	3
0.05	0.10	0.039 ± 0.006	0.780 ± 0.013	3
0.05	0.15	0.040 ± 0.013	0.758 ± 0.002	3
0.05	0.20	0.053 ± 0.007	0.812 ± 0.009	3
0.05	0.25	0.069 ± 0.021	0.800 ± 0.010	3
0.05	0.30	0.045 ± 0.014	0.796 ± 0.003	3
0.10	0.00	0.025 ± 0.005	0.730 ± 0.011	3
0.10	0.05	0.069 ± 0.031	0.826 ± 0.005	3
0.10	0.10	0.047 ± 0.011	0.789 ± 0.005	3
0.10	0.15	0.070 ± 0.003	0.787 ± 0.020	3
0.10	0.20	0.080 ± 0.029	0.836 ± 0.006	3
0.10	0.25	0.071 ± 0.007	0.787 ± 0.014	3
0.10	**0.30**	$\mathbf{0.094 \pm 0.032}$	$\mathbf{0.804 \pm 0.017}$	**3**
0.15	0.00	0.038 ± 0.012	0.764 ± 0.027	3
0.15	0.05	0.027 ± 0.005	0.781 ± 0.028	3
0.15	0.10	0.053 ± 0.014	0.760 ± 0.010	3
0.15	0.15	0.036 ± 0.003	0.779 ± 0.013	3
0.15	0.20	0.030 ± 0.003	0.795 ± 0.014	3
0.15	0.25	0.027 ± 0.006	0.789 ± 0.017	3
0.15	0.30	0.034 ± 0.007	0.797 ± 0.017	3
0.20	0.00	0.029 ± 0.000	0.764 ± 0.000	1
0.20	0.05	0.045 ± 0.010	0.819 ± 0.009	3
0.20	0.10	0.046 ± 0.015	0.810 ± 0.004	3
0.20	0.15	0.036 ± 0.003	0.801 ± 0.020	3
0.20	0.20	0.042 ± 0.003	0.792 ± 0.010	3
0.20	0.25	0.037 ± 0.009	0.822 ± 0.011	3
0.20	0.30	0.050 ± 0.012	0.836 ± 0.008	3

Table 6: Seed-averaged performance across temporal offsets around the keyword onset. Values are mean $\pm$ standard error across seeds. The row with the highest %ΔAUPRC over the 0/0 baseline is typeset in bold.