Benchmarking Non-Invasive Speech BCIs: Lessons Learned from the 2025 PNPL Competition

Gereon Elvers[1] GEREON.ELVERS@TUM.DE
Gilad Landau[1] GILAD@ROBOTS.OX.AC.UK
Francesco Mantegna[1] FRANCESCO@ROBOTS.OX.AC.UK
Miran Özdogan[1]
Tasha Kim[1]
Teyun Kwon[1]
SungJun Cho[1,2]
Benjamin Ballyk[1]
Luisa Kurth[1]
Dulhan Jayalath[1]
Pratik Somaiya[1]
Chetan Gohil[2]
Brendan Shillingford[3]
Greg Farquhar[3]
Minqi Jiang[4]
Caglar Gulcehre[5]
Xabier de Zuazo[6]
Hamza Abdelhedi[7]
Yorguin Mantilla Ramos[7]
Karim Jerbi[7]
Mark Woolrich[2]
Natalie Voets[8]
Oiwi Parker Jones[1] OIWI@ROBOTS.OX.AC.UK

[1] *PNPL 🍍, University of Oxford*
[2] *OHBA, University of Oxford*
[3] *Google DeepMind*
[4] *Meta AI*
[5] *École Polytechnique Fédérale de Lausanne*
[6] *University of the Basque Country – UPV/EHU*
[7] *Mila, Quebec AI Institute & CoCo Lab, Université de Montréal*
[8] *NDCN, University of Oxford*

Editors: Tao Qin, Kun Zhang, and Jes Frellsen

Abstract

We present a retrospective on the 2025 PNPL Competition, an open machine-learning benchmark for decoding speech from non-invasive brain recordings, which ran from June to September 2025 and culminated in a workshop at NeurIPS that December. The competition used the LibriBrain dataset — over 50 hours of within-subject MEG with standard data splits — and challenged participants to tackle two tasks: Speech Detection (binary frame-level classification) and Phoneme Classification (39-class ARPAbet prediction). Across 155

registered teams and 6,041 submissions, top F1-macro scores reached 95.6% for Speech Detection and 73.6% for Phoneme Classification, exceeding prior baselines of 68% and 44%, respectively, by wide margins. High-performing submissions converged on a common architectural template — temporal CNN or TCN backbones with task-specific recurrent or attention-based heads — whilst task-specific strategies such as temporal smoothing and ensemble voting proved critical for competitive performance. Hidden track evaluations suggest that main-track Speech Detection models generalised poorly to balanced class distributions. This could imply a reliance on class priors rather than robust neural decoding, highlighting a useful lesson: evaluation distributions should reflect intended deployment conditions. We further reflect on design choices, organisational lessons, and plans for future competitions in this series.

Keywords: brain decoding, speech detection, phoneme classification, MEG, benchmark

1. Introduction

By 2025, non-invasive speech decoding had reached an inflection point: deep learning methods were producing their first substantial results on MEG and EEG data, yet progress remained fragmented across incompatible datasets and splits, unreproducible models and training pipelines, and a field divided between neuroscientists and machine learning researchers. The stakes were not academic. Invasive brain-computer interfaces have demonstrated that decoding intended speech from neural signals is achievable (Moses et al., 2021) — with word error rates falling from around 20% in 2023 (Willett et al., 2023) to 2.5% in 2024 (Card et al., 2024). Brain surgery, however, remains a prohibitive barrier. A non-invasive route to speech decoding would transform the clinical picture, and MEG, with its combination of high temporal and spatial resolution, currently offers the best available signal to build a bridge from surgical systems to consumer-grade EEG devices. The primary challenge is that non-invasive decoding is substantially harder than invasive decoding; and the field has lacked the shared *infrastructure* needed to make rapid, reproducible progress.

This infrastructure gap inspired work to begin at Oxford in 2023, culminating in the first of multiple large-scale data releases that we are planning: LibriBrain (Özdogan et al., 2025), over 50 hours of within-subject MEG from a single participant listening to audiobooks, with standard train, validation, and test splits designed for reproducibility. LibriBrain's within-subject design was a deliberate choice, as deep single-subject data has been shown to yield better decoding performance per recording hour than shallow multi-subject data (d'Ascoli et al., 2025), and models trained on LibriBrain substantially outperform models trained on earlier datasets even at equivalent data volumes (Jayalath et al., 2025a). Isolated analyses using similar datasets existed before (Défossez et al., 2023; d'Ascoli et al., 2025; Jayalath et al., 2025b), but without standardised evaluation it has been difficult to know what generalisable lessons to draw. The precedent from computer vision here is instructive: shared benchmarks such as ImageNet (Russakovsky et al., 2015) and, more recently, the Natural Scenes Dataset in fMRI (Allen et al., 2022), have accelerated progress precisely by creating a common ground for comparison. In 2025, LibriBrain provided the foundation to do the same for non-invasive speech decoding.

The 2025 PNPL Competition (Landau et al., 2025b) was designed from the ground up with the LibriBrain data collection to build on this foundation, with accessibility as its guiding philosophy: bringing the machine learning community into an area that has histor-

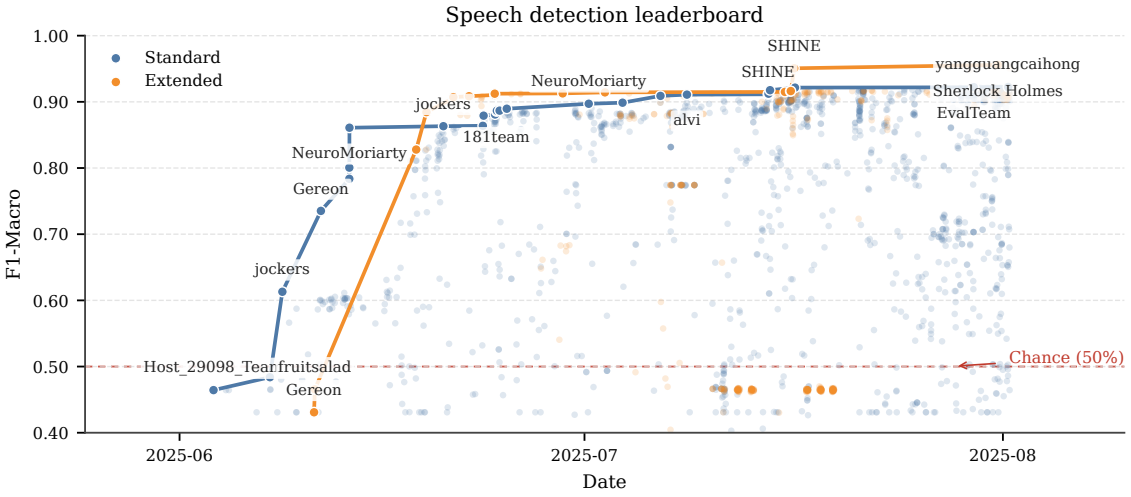

Figure 1: Leaderboard for Speech Detection (Phase 1). The x- and y-axes show submission dates and F1-macro scores (higher is better). Each circle is a submission, with blue and orange lines outlining improvements in the best models over time for the Standard and Extended Tracks. Balanced binary chance here is 50% (dashed red line). The final rankings, which used independent holdout data, are in Table 1; see Appendix A for the Phoneme Classification leaderboard.

ically demanded considerable domain expertise, by abstracting human neuroimaging data into standard tensor formats and providing end-to-end tooling. Across two phases and 155 registered teams, it produced 6,041 submissions on the benchmark tasks of Speech Detection and Phoneme Classification, culminating in a workshop at NeurIPS 2025 in San Diego. This paper is a retrospective on that competition. We synthesise patterns across competition submissions in Section 3.4, draw organisational lessons in Section 4, and conclude with plans for the 2026 PNPL Competition as the community advances towards word-level and ultimately full brain-to-text applications of non-invasive speech decoding.

2. Competition Overview

This section details the competition structure, tasks, evaluation metrics, and materials provided to participants. Complete specifications are available in the competition paper (Landau et al., 2025b) and on the competition website.[1]

2.1. Dataset and Splits

The competition used the LibriBrain dataset (Özdogan et al., 2025): over 50 hours of non-invasive MEG from a single participant listening to audiobooks, with time-locked speech/non-speech and phoneme annotations. Data were divided into train (51.57 hours), validation (0.36 hours), and test (0.38 hours) partitions, with additional holdout sessions reserved for (1) public leaderboard updates, (2) final rankings, and (3) hidden tasks (Table 2,

1. https://libribrain.com/

Section 3.4.4; also Appendix B). To reduce information leakage via slow features (Harris, 2021), all holdout sessions were acquired on different days from the training data.

2.2. Tasks

Two supervised decoding tasks were defined. **Speech Detection**: given $\mathbf{X} \in \mathbb{R}^{306\times T}$ (sensors $\times$ time points), predict for each $t \in \{1,\dots,T\}$ at 4 ms resolution whether speech is present ($y_t = 1$) or absent ($y_t = 0$). The general train, validation, test, and competition holdout data contained more speech than non-speech frames. **Phoneme Classification**: given a fixed window $\mathbf{X} \in \mathbb{R}^{306\times T'}$ (sensors $\times$ window size) aligned to a phoneme onset, predict which of 39 ARPAbet classes it corresponds to. Training and public evaluation data for this task used averages over 100 repetitions per phoneme to improve the signal-to-noise ratio (for decoding results with different amounts of averaging, see Özdogan et al., 2025).

2.3. Tracks and Rules

Each task was offered in a **Standard Track** (LibriBrain data only, emphasising algorithmic innovation) and an **Extended Track** (additional data and pre-training permitted), with separate leaderboards for each combination. Prize eligibility rules were updated between phases in response to Extended Track participation patterns; full details are given in Section 4.2.

The competition also included a **Hidden Track** for each task. These were not originally advertised, however this policy was updated during the competition (see Section 4.4). For speech detection, the hidden track used balanced speech and non-speech frames; for phoneme classification, single-trial (unaveraged) examples were used. See Section 3.4.4 for hidden track results, and Section 4.3 for a discussion of their implications.

2.4. Evaluation and Submission Protocol

All submissions were evaluated using the macro-averaged F1 score

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{F1}_k, \quad \mathrm{F1}_k = \frac{2\cdot \mathrm{TP}_k}{2\cdot \mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k},$$

where TP_k, FP_k, and FN_k denote true positives, false positives, and false negatives for each class k, with $K = 2$ for Speech Detection and $K = 39$ for Phoneme Classification. This metric was chosen over alternatives such as micro-averaged F1 (which weights classes by frequency) to penalise models that ignore minority classes, encouraging balanced performance across all categories. We generally report F1-macro scores as percentages.

Participants submitted predictions as CSV files via the EvalAI platform. Submissions were scored automatically against the public leaderboard holdout set, with results displayed on the competition website in real time. Final rankings were computed on an independent holdout set to control for overfitting on leaderboard feedback. For prize-eligible entries, Standard Track teams in prize positions were contacted to share training code to verify that no external datasets were used; Extended Track teams were required to confirm use of additional data beyond LibriBrain.

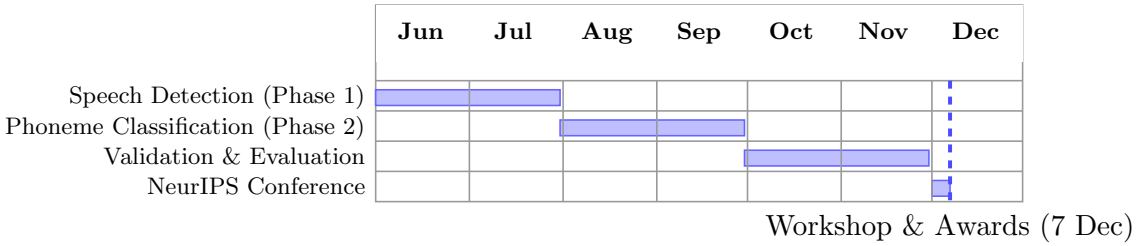

Figure 2: 2025 competition timeline: two competition phases, validation and evaluation, and the NeurIPS conference (2–7 December) with workshop/awards on 7 December.

2.5. Timeline

The competition was run over two phases as shown in Figure 2. Final verification was completed before the NeurIPS workshop on 7 December 2025.

2.6. Resources and Support

To lower barriers to entry, we provided: (1) a **Python library** (`pnpl`) with automatic dataset downloading and PyTorch-compatible data loaders; (2) **reference models** achieving F1-macro scores of 68% (Speech Detection) and 44% (Phoneme Classification); (3) four Google Colab **tutorial notebooks**,[2] executable with free GPU access, that provide code to train baseline models for both tasks, as well as examples for making competition submissions; (4) **blog posts** on problem context, baselines, and modelling approaches (Elvers et al., 2025; Landau et al., 2025a; Mantegna et al., 2025; Kwon et al., 2025); (5) a **Discord server** with 248 registered members; and (6) a **submission platform** with automatic evaluation and real-time leaderboard updates.[3] All materials remain publicly available.

3. Participant Outcomes

3.1. Participation Overview

The competition attracted 155 registered teams from at least 15 countries, with participants representing academia, industry, and independent researchers. Speech Detection (Phase 1) attracted 93 unique teams and 1,958 submissions; Phoneme Classification (Phase 2) engaged 62 teams and 4,083 submissions. We identified 7 teams that competed in both phases (see Appendix C for participation statistics).

Beyond competition submissions, community engagement with the supporting resources has been encouraging: the `pnpl` library was downloaded 14,736 times between release and March 2026 (Appendix D). Over the same period, the dataset was downloaded 15,086 times, and at its peak, the Discord server hosted 248 members. In total, 11 system-description papers (from 10 groups) were submitted to the NeurIPS 2025 PNPL Workshop by the following teams: BrainWatson (Suzuki et al., 2025), I_love_silksong (Li et al., 2026; Liang et al., 2026), neural2speech (de Zuazo et al., 2025), NTHU HMI LAB (Chien et al., 2025),

2. https://neural-processing-lab.github.io/2025-libribrain-competition/participate/
3. https://eval.ai/web/challenges/challenge-page/2504/overview

Task	Track	Score	Prize	USD	Winning Team	Paper
Speech Detection	Extended	95.6%	1st Extended	$1,200	Sherlock Holmes	
Speech Detection	Extended	95.1%	2nd Extended	$1,000	SHINE	
Speech Detection	Standard	92.18%	1st Standard	$1,200	Parameter Team	
Speech Detection	Standard	92.15%	2nd Standard	$1,000	TheLimbicLad	
Speech Detection	Standard	92.12%	3rd Standard	$800	alvi	
Speech Detection	Extended	90.74%	Special Prize	$300	I_love_silksong	
Speech Detection	Hidden	51.44%	Hidden Prize	$600	NeuroAI	
Phoneme Classification	Standard	73.6%	1st Standard	$1,200	neural2speech	
Phoneme Classification	Standard	73%	2nd Standard	$1,000	BrainWatson	
Phoneme Classification	Standard	65.7%	3rd Standard	$800	sigint	
Phoneme Classification	Extended	60.4%	Special Prize	$300	I_love_silksong	
Phoneme Classification	Hidden	5.38%	Hidden Prize	$600	September Labs	
			Total Prizes	**$10,000**		

Table 1: Winning teams and prize allocations. Scores reported as macro-averaged F1, sorted by task from best (highest) on competition holdout data. All prizes in USD. Where available, PDF icons in the 'Paper' column link to workshop papers.

Null-1 (2025), QuantumFiremen (2025), Saidnesh (2025), September Labs (Stepanov et al., 2025), Sherlock Holmes Team (2025), and SHINE (Xu et al., 2026). See Table 1 for rankings.

3.2. Track Participation and Rule Adaptation

The Standard Track consistently drew more participants than the Extended Track, with Standard-to-Extended ratios of 3.4:1 for Speech Detection and 3.1:1 for Phoneme Classification. Ultimately, the Extended Track did not attract sufficient competitive depth in Phase 1 to continue justifying ranked prizes in Phase 2, as we discuss in Section 4.2.

3.3. Performance Summary

Table 1 summarises all prize-winning submissions sorted by task and F1-macro score.

Speech Detection produced top scores exceeding 95% F1-macro in the Extended Track and 92% in the Standard Track, against a reference model baseline of 68% and balanced binary chance at 50%. The top three Standard Track submissions clustered between 92.12% and 92.18% — a range of just 0.06 percentage points — suggesting convergence on a performance ceiling within the constraints of the competition data. Extended Track submissions achieved higher absolute scores, with two teams — Sherlock Holmes (95.6%) and SHINE (95.1%) — demonstrating clear benefits from the additional resources, though one approach required a closed-set audio database and therefore has different applicability assumptions from open-vocabulary decoding (see Sherlock Holmes Team, 2025).

Phoneme Classification was substantially harder, consistent with its role as the more demanding task in the curriculum. The best Standard Track score was 73.6% F1-macro (neural2speech), well above the reference baseline (44%) and far above chance (2.6% for 39 balanced classes), but with room for further improvement. The gap between first and second place (0.6 percentage points) was narrower than between second and third place (7.3

percentage points), suggesting that neural2speech and BrainWatson converged on similar approaches whilst a larger gap separated these leaders from the rest of the field. Notably, Extended Track submissions did not outperform the Standard Track for Phoneme Classification: the best Extended score (60.4%) trailed all three Standard Track prize winners. This contrasts with Speech Detection, where the Extended Track submissions dominated.

3.4. Patterns Across Submitted Systems

The following analysis draws on prize-winning workshop papers, leaderboard trajectories, and structured information collected from high-performing teams. Because the number of machine learning systems available for analysis is small and many design choices were correlated across teams, the observations below should be read as empirical patterns and practical hypotheses rather than statistically established causal conclusions. We return to this caveat formally in Section 3.4.5.

3.4.1. A common architectural template

Despite the diversity of teams and backgrounds, high-performing submissions converged on a strikingly similar set of approaches. Most treated the MEG recording as a high-dimensional multichannel time series and built models accordingly, rather than adapting architectures from computer vision or natural language processing without modification.

The common template used the 306-channel MEG time series as input, applied a temporal convolutional backbone — typically using dilated convolutions or TCN-style blocks — and attached a task-specific prediction head. Dilated convolutions are well-suited to MEG: by increasing the dilation factor exponentially across layers, the receptive field grows to cover long temporal windows (8–30 s for Speech Detection) without a proportional increase in parameters. This architecture was used by the Parameter Team, SHINE, neural2speech, BrainWatson, and others, yielding a mean F1-macro of 87% across this group.

Within this shared template, task-specific routing diverged meaningfully. For Speech Detection, all three Extended Track prize-winning teams added bidirectional LSTM heads to capture long-range context in both temporal directions. For Phoneme Classification, the best-performing submission (neural2speech) used a Conformer-based head (Gulati et al., 2020), which combines self-attention with convolutional modules in a macaron-style architecture (FFN → attention → convolution → FFN), combining sensitivity to local spectral structure with broader contextual integration (de Zuazo et al., 2025). The second strongest Speech Detection submission (SHINE) used WaveNet-style layers as a variant on the same convolutional theme (Xu et al., 2026).

3.4.2. Task-specific drivers of performance

Whilst both tasks used the same LibriBrain recordings, they rewarded meaningfully different modelling choices.

Speech Detection. The largest single performance improvement came not from architecture but from post-processing. Model predictions at the 250 Hz MEG sampling rate oscillated between speech and non-speech labels every 4 ms — physiologically implausible,

given that speech segments last several seconds. Enforcing temporal coherence via smoothing yielded a mean F1-macro gain of approximately +21 percentage points across teams that applied it. Methods included median filters (window size $w = 40$), moving averages, and two-threshold hysteresis, all of which suppress high-frequency label switching whilst preserving genuine speech onset and offset transitions.

Phoneme Classification. The two largest performance drivers were ensemble voting and sample averaging. Combining predictions from multiple models trained with different random seeds, resolving disagreements by majority vote, produced the largest single gain of approximately +13 percentage points in F1-macro. All eight top-performing phoneme teams also exploited the 100-trial averaged data provided by the competition to improve the signal-to-noise ratio. A third factor was instance-level normalisation, which was critical for holdout generalisation in the winning team (neural2speech). A fourth factor was the utilisation of class imbalances across phoneme classes. Although F1-macro gives each class equal influence on the final evaluation score, standard example-averaged training losses give more influence to frequent phonemes. The top two submissions addressed this with inverse-square-root or temperature-scaled class weighting in the loss function.

3.4.3. Approaches that did not transfer

The clearest case of an approach that did not transfer was the use of pretrained vision backbones. Some teams reshaped the 306-channel MEG sensor array into image-like arrays compatible with architectures such as EfficientNet (Tan and Le, 2020). These approaches underperformed substantially, plausibly because the pretrained features were poorly matched to MEG waveform structure and image-like reshaping distorted the sensor array topology.

3.4.4. Evaluation robustness and hidden tracks

The hidden track evaluations tested generalisation beyond the training distributions. For balanced Speech Detection, the best score was 51.44% F1-macro — only slightly above balanced binary chance (50%) — despite the same models achieving above 90% on the holdout for the main tracks. For single-trial Phoneme Classification, the best hidden score was 5.38% F1-macro, above chance (2.6%) but well below the averaged-trial performance. This drop is expected given that single-trial MEG is a substantially harder problem than 100-trial averages (see analysis of averaging in Özdogan et al., 2025). These results admit several interpretations, which we discuss in the context of evaluation design in Section 4.3.

3.4.5. Statistical caveats

The number of high-performing submissions available for systematic comparison was small — approximately seven per track — which is insufficient for reliable inference about the independent contribution of any single design choice. Permutation tests across architectural features identified only one factor, the choice of Transformer architectures in Speech Detection, as approaching conventional significance ($p = 0.095$), and none of the features we explored survived Benjamini–Hochberg false discovery rate correction (all adjusted $p > 0.45$).

4. Organisational Outcomes

4.1. Competition Materials Lowered the Barrier to Entry

A central goal of the competition was to make non-invasive neural decoding accessible to AI researchers without prior neuroimaging experience. All the resources we provided — automatic data downloading, PyTorch-compatible data loaders, Google Colab notebooks with free GPU access, and blog posts covering problem context and modelling approaches — appear to have achieved this. The dataset was downloaded 1348 times in the first month and 155 teams registered, representing academia, industry, and independent researchers.

We set the reference models at a level that was deliberately more achievable rather than highly optimised, reasoning that an approachable baseline would encourage broader participation and deeper engagement. In practice engagement was indeed strong: the baseline was beaten by a large number of teams. Lessons from other competitions suggest that compute grants may also be as valuable as cash prizes for driving participation, particularly amongst student teams (e.g., Milani et al., 2020).

4.2. Even Small Friction Can Affect Participation

The two-track design balanced two goals. The Standard Track was intended to create a space where algorithmic innovation could compete on equal footing, without scale alone determining the winner. The Extended Track, by contrast, explicitly invited the bitter lesson (Sutton, 2019): teams were free to bring unlimited additional data, pre-training, and compute to see how far scaling could push performance. We assumed that teams competing on the Extended Track would be comfortable integrating external datasets into the `pnpl` library themselves, so the results surprised us.

In Phase 1, only two Extended Track submissions outperformed the Standard Track winners (Table 1), falling short of the three required to justify separate ranked prizes; a third group received a Special Prize rather than a ranked prize, so that future readers would not interpret limited Extended Track competition as evidence against scaling. By Phase 2, we had responded by replacing ranked Extended Track prizes with a single Special Prize, awarded to the best Extended Track effort without requiring it to outperform Standard Track entries.

Why did the bitter lesson not prevail and scaling dominate? One hypothesis is friction: because the Standard Track was comprehensively supported end-to-end, participants could focus entirely on modelling. By contrast, the Extended Track required integrating additional datasets into the `pnpl` library, and may have felt comparatively effortful. Compute resources were likely a second barrier. To address this, future competitions could explore compute partnerships and provide equivalent tooling for Extended Tracks, including ready-made loaders for external datasets. A growing number of efforts are emerging to support this (e.g., King et al., 2026), motivated in part by work on foundation models for the brain (e.g., Jiang et al., 2024; Yang et al., 2024; Csaky et al., 2024; Xiao et al., 2025; Huang et al., 2025; Cho et al., 2026; Jayalath and Parker Jones, 2026), so the 2025 PNPL Competition may simply have been ahead of the curve.

4.3. Evaluation Distributions Should Reflect Deployment

While the core evaluation methodology proved to be effective (e.g., automatic scoring via online platform using the macro-F1 score on independent leaderboard and final holdout splits), a lesson emerged from the hidden tracks. For Speech Detection, speech labels constituted approximately 78% of frames in the training, validation, test, and competition holdout data. This is a reasonable property for a dataset collected to study speech processing; given that MEG recording time is expensive, it makes sense to maximise the proportion of time in which a task stimulus is present. However, this meant that every evaluation split shared the same class imbalance, and there was no incentive not to exploit the empirical prior. As discussed in Section 3.4.4, the best Speech Detection models achieved above 90% F1-macro on the imbalanced holdout whilst barely exceeding chance on the balanced hidden evaluation. This is consistent with models learning to predict the dominant class (though we have not ruled out other explanations, such as atypically low signal quality in the balanced holdout session). Similarly, evaluating phoneme classification on averaged rather than single-trial samples meant that models were not required to be robust to individual sample noise.

These observations motivate a future design principle: evaluation distributions should reflect the intended deployment context, not the data collection process. In assistive communication, non-speech is likely to outweigh intended speech, and averaging across very large numbers of samples is likely infeasible. Future competitions would benefit from evaluating across a range of class distributions.

4.4. Disclose Hidden Tracks from the Start

As we elaborate in Appendix B, the nature of the hidden evaluation data was not initially disclosed — we had intended it as a surprise to reveal at the NeurIPS workshop. However, as results came in, several keen-eyed teams noticed that the statistical properties of the competition holdout appeared to shift and raised the alarm on the community Discord server. We refer to this episode jokingly as 'distribution gate'. Once community interest was sufficiently widespread we publicly confirmed the existence of a hidden evaluation track.

The engagement this generated was positive: teams were reasoning carefully about evaluation methodology and discussing together, both goals of the competition. It also motivates another lesson for future editions: tell participants at the outset if a competition holdout includes a hidden partition; that does not have to ruin the surprise of what it's for.

5. Conclusions

The 2025 PNPL Competition is an existence proof that open machine-learning benchmarking can drive rapid progress on non-invasive speech decoding. Starting from reference model baselines of 68% and 44% F1-macro, the machine learning community reached 95.6% and 73.6% across 155 registered teams and 6,041 submissions. Comprehensive materials and a task curriculum contributed to broad participation, and the competition generated organisational lessons for future editions. Together these advance the broader agenda: moving the curriculum forward to word-level and ultimately full text-level decoding, with evaluation conditions that reflect the needs of the people who will benefit from non-invasive speech BCIs.

Acknowledgements

We thank all competition and workshop participants whose contributions made this experiment a success; we're excited to continue building this community together. We also thank Pillar VC for sponsoring $10,000 in prizes and for hosting the social event at NeurIPS. We would like to thank NeurIPS for supporting the workshop on the day and for providing complimentary conference registrations for participants. We also gratefully acknowledge Google for providing free GPU access via Google Colab, which was essential for lowering the computational barrier to entry. PNPL is supported by the MRC (MR/X00757X/1), Royal Society (RG\R1\241267), NSF (2314493), NFRF (NFRFT-2022-00241), SSHRC (895-2023-1022), and ARIA (SCNI-SE01-P004).

References

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x. URL https://doi.org/10.1038/s41593-021-00962-x.

Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, and David M. Brandman. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024. doi: 10.1056/NEJMoa2314132.

S.-Y. Chien, B.-Y. Mao, Y.-N. Chang, and P.-C. Kuo. Evaluating impact of distribution shifts on preprocessing and modeling choices for speech decoding from MEG signals, 2025. URL https://libribrain.com/workshop-papers/NTHU%20HMI%20LAB/libribrain-nthuhmilab-2025.pdf. Presented at the 2025 PNPL Workshop at NeurIPS.

SungJun Cho, Chetan Gohil, Rukuang Huang, Oiwi Parker Jones, and Mark Woolrich. A systematic evaluation of sample-level tokenization strategies for MEG foundation models. *arXiv preprint arXiv:2602.16626*, 2026. URL https://arxiv.org/abs/2602.16626.

Richard Csaky, Mats W. J. van Es, Oiwi Parker Jones, and Mark Woolrich. Foundational GPT model for MEG. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2404.09256.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16:10521, 2025. doi: 10.1038/s41467-025-65499-0. URL https://doi.org/10.1038/s41467-025-65499-0.

Xabier de Zuazo, Ibon Saratxaga, and Eva Navas. MEGConformer: Conformer-based MEG decoder for robust speech and phoneme classification, 2025. URL https://arxiv.org/abs/2512.01443. Presented at the 2025 PNPL Workshop at NeurIPS.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Ma-*

chine Intelligence, 5(10):1097–1107, 2023. doi: 10.1038/s42256-023-00714-5. URL https://www.nature.com/articles/s42256-023-00714-5.

Gereon Elvers, Gilad Landau, Dulhan Jayalath, Francesco Mantegna, and Oiwi Parker Jones. Building a collaborative foundation for non-invasive speech BCIs: The 2025 PNPL competition, 2025. URL https://neural-processing-lab.github.io/2025-libribrain-competition/blog/building-collaborative-foundation-non-invasive-speech-bcis/. Blog post.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, 2020. URL https://arxiv.org/abs/2005.08100.

Kenneth D. Harris. Nonsense correlations in neuroscience. *bioRxiv*, 2021. doi: 10.1101/2020.11.29.402719.

Rukuang Huang, SungJun Cho, Chetan Gohil, Oiwi Parker Jones, and Mark Woolrich. MEG-GPT: A transformer-based foundation model for magnetoencephalography data. *arXiv preprint arXiv:2510.18080*, 2025. URL https://arxiv.org/abs/2510.18080.

Dulhan Jayalath and Oiwi Parker Jones. MEG-XL: Data-efficient brain-to-text via long-context pre-training. *arXiv preprint arXiv:2602.02494*, 2026. URL https://arxiv.org/abs/2602.02494.

Dulhan Jayalath, Gilad Landau, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text. *International Conference on Machine Learning (ICML), Workshop on Generative AI and Biology*, 2025a. URL https://arxiv.org/abs/2505.13446. arXiv preprint arXiv:2505.13446.

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. URL https://proceedings.mlr.press/v267/jayalath25a.html.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Jean-Rémi King, Corentin Bel, Linnea Evanson, Julien Gadonneix, Sophia Houhamdi, Jarod Lévy, Josephine Raugel, Andrea Santos Revilla, Mingfang Zhang, Julie Bonnaire, Charlotte Caucheteux, Alexandre Défossez, Théo Desbordes, Pablo Diego-Simón, Shubh Khanna, Juliette Millet, Pierre Orhan, Saarang Panchavati, Antoine Ratouchniak, Alexis Thual, Teon L. Brooks, Katelyn Begany, Yohann Benchetrit, Marlène Careil, Hubert Banville, Stéphane d'Ascoli, Simon Dahan, and Jérémy Rapin. NeuralSet: A high-performing Python package for Neuro-AI. *arXiv preprint arXiv:2605.03169*, 2026. URL https://arxiv.org/abs/2605.03169.

Teyun Kwon, SungJun Cho, Gereon Elvers, Francesco Mantegna, and Oiwi Parker Jones. Language-inspired approaches to phoneme classification, 2025. URL https://neural-processing-lab.github.io/2025-libribrain-competition/blog/language-inspired-approaches-phoneme-classification. Blog post.

Gilad Landau, Gereon Elvers, Miran Özdogan, and Oiwi Parker Jones. The speech detection task and the reference model, 2025a. URL https://neural-processing-lab.github.io/2025-libribrain-competition/blog/speech-detection-reference-model. Blog post.

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*, 2025b. URL https://arxiv.org/abs/2506.10165.

Songyi Li, Linze Zheng, Jinghua Liang, and Zifeng Zhang. MEBM-Speech: Multi-scale enhanced BrainMagic for robust MEG speech detection. *NeurIPS, 2025 PNPL Workshop*, 2026. URL https://arxiv.org/abs/2603.02255. Team: I_love_silksong.

Jinghua Liang, Zifeng Zhang, Songyi Li, and Linze Zheng. MEBM-Phoneme: Multi-scale enhanced brainmagic for end-to-end MEG phoneme classification. *NeurIPS, 2025 PNPL Workshop*, 2026. URL https://arxiv.org/abs/2603.02254. Team: I_love_silksong.

Francesco Mantegna, Gereon Elvers, and Oiwi Parker Jones. Brain-inspired approaches to speech detection, 2025. URL https://neural-processing-lab.github.io/2025-libribrain-competition/blog/brain-inspired-approaches-speech-detection. Blog post.

Stephanie Milani, Nicholay Topin, Brandon Houghton, William H. Guss, Sharada P. Mohanty, Keisuke Nakata, Oriol Vinyals, and Noboru Sean Kuno. Retrospective analysis of the 2019 MineRL competition on sample efficient reinforcement learning. In *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 203–214. PMLR, 2020.

David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021. doi: 10.1056/NEJMoa2027540. URL https://doi.org/10.1056/NEJMoa2027540.

Null-1. Team Null-1 phoneme classification system for the NeurIPS 2025 PNPL Competition (task 2), 2025. URL https://libribrain.com/workshop-papers/Null-1/libribrain-null1-2025.pdf. Presented at the 2025 PNPL Workshop at NeurIPS.

QuantumFiremen. All-for-one: Combining small convolutional models for MEG-based speech detection, 2025. URL https://libribrain.com/workshop-papers/

QuantumFiremen/libribrain-quantumfiremen-2025.pdf. Presented at the 2025 PNPL Workshop at NeurIPS.

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. doi: 10.1007/s11263-015-0816-y.

Saidnesh. Libri phoneme classification, 2025. URL https://libribrain.com/workshop-papers/Saidnesh/libribrain-saidnesh-2025.pdf. Presented at the 2025 PNPL Workshop at NeurIPS.

Sherlock Holmes Team. Bypassing direct reconstruction: Speech detection from MEG via large-scale audio retrieval, 2025. URL https://libribrain.com/workshop-papers/Sherlock%20Holmes/libribrain-sherlockholmes-2025.pdf. Presented at the 2025 PNPL Workshop at NeurIPS.

Ihor Stepanov, Aleksandr Smechov, Alexander Yavorskyi, and Shivam Chaudhary. DeMEGa: Decoding listened phonemes via MEG signals by stacking conformers with DeBERTa-style disentangled attention, 2025. URL https://libribrain.com/workshop-papers/September%20Labs/libribrain-septemberlabs-2025.pdf. Presented at the 2025 PNPL Workshop at NeurIPS.

Richard Sutton. The bitter lesson. Incomplete Ideas (blog), 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Shuntaro Suzuki, Chia-Chun Dan Hsu, Yu Tsao, and Komei Sugiura. MEGState: Phoneme decoding from magnetoencephalography signals. *NeurIPS, 2025 PNPL Workshop*, 2025. URL https://arxiv.org/abs/2512.17978. Team: BrainWatson.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2020.

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*, 620:1031–1036, 2023. doi: 10.1038/s41586-023-06377-x.

Qinfan Xiao, Ziyun Cui, Chi Zhang, SiQi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=cjHQj0tCy6.

Xiran Xu, Yujie Yan, Xihong Wu, and Jing Chen. SHINE: Sequential hierarchical integration network for EEG and MEG, 2026. URL https://arxiv.org/abs/2602.23960. Presented at the 2025 PNPL Workshop at NeurIPS.

Yiqian Yang, Yiqun Duan, Hyejeong Jo, Qiang Zhang, Renjing Xu, Oiwi Parker Jones, Xiaoqian Hu, Chin-Teng Lin, and Hui Xiong. NeuGPT: Unified multi-modal neural GPT. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2410.20916.

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.

Appendix A. Leaderboard for Phoneme Classification

Figure 3 shows the leaderboard trajectory for Phoneme Classification (Phase 2). The same pattern of steady improvement that we saw in the Speech Detection leaderboard (Figure 1) is visible. Note that the dates along the x-axes of the two leaderboards differ, as Speech Detection was run first in Phase 1 before Phoneme Classification in Phase 2 (Figure 2).

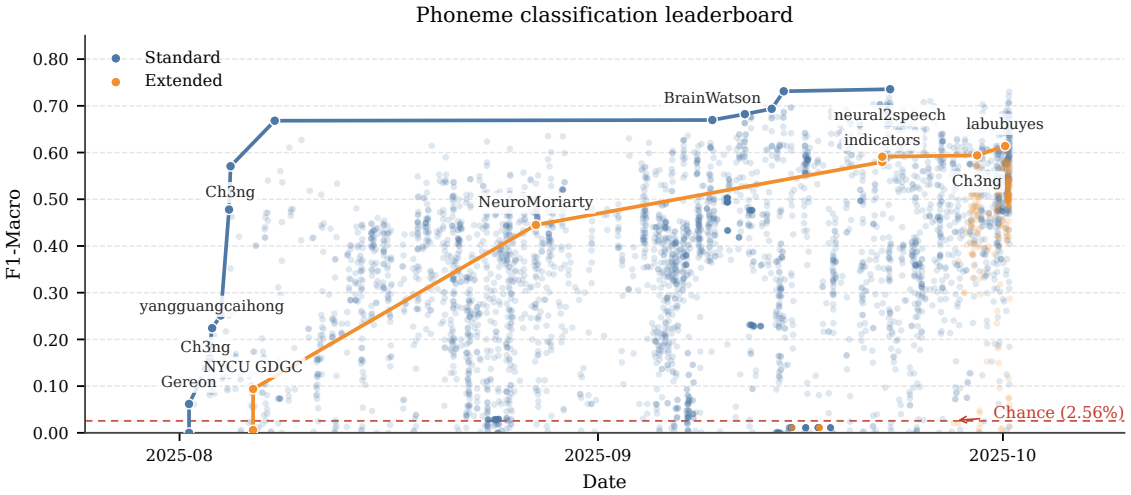

Figure 3: Leaderboard for Phoneme Classification (Phase 2). The x- and y-axes show submission dates and F1-macro scores (higher is better). Each circle is a submission, with blue and orange lines outlining improvements in the best models over time for the Standard and Extended Tracks. For reference, chance here is 2.56% (1/39 classes). The reference model baseline was 44%.

Appendix B. Data Splits Revealing Hidden Tasks

Table 2 provides a full overview of the data splits used in the competition, including the structure of the competition holdout set which was excluded from the original competition paper (Landau et al., 2025b). The hidden tasks used balanced classes for Speech Detection and single trials for Phoneme Classification. See Section 3.4.4 for discussion of the hidden track results, and Section 4.4 for additional discussion.

Table 2: Overview of data splits. The competition holdout comprises three disjoint subsets: leaderboards, final rankings, and hidden tasks (blue box). Hidden task details were withheld from the original competition paper (Landau et al., 2025b). They were used to evaluate balanced classes for Speech Detection and single trials for Phoneme Classification (Section 3.4.4).

	Train	**Validation**	**Test**	**Competition Holdout**		
				Leaderboard	Final Rankings	Hidden Tasks
Data	Public	Public	Public	Public	Public	Public
Labels	Public	Public	Public	Private	Private	Private
Hours	51.57	0.36	0.38	Private	Private	Private

Appendix C. Participation Statistics

In addition to aggregate participation counts, Figures 4 and 5 show cumulative submissions over time for Phase 1 and Phase 2, respectively.

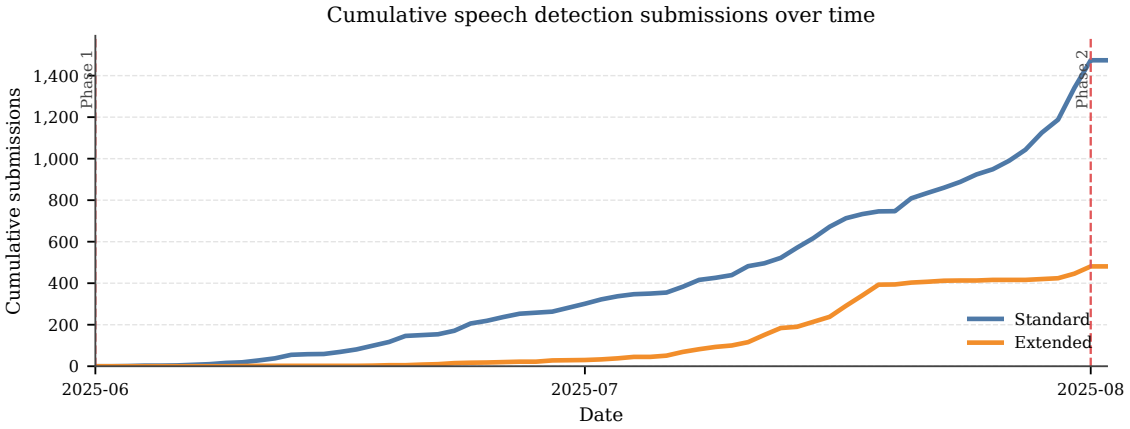

Figure 4: Cumulative submissions over time for Phase 1: Speech Detection.

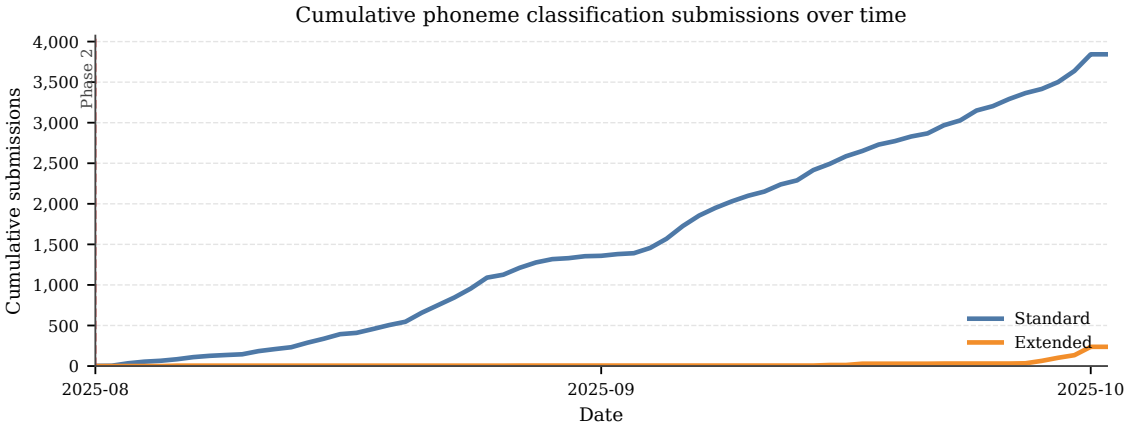

Figure 5: Cumulative submissions over time for Phase 2: Phoneme Classification.

Table 3 reports additional participation statistics across tasks and tracks.

	Standard	Extended	Total
Speech Detection (Phase 1)			
Teams	72	21	93
Submissions	1,476	482	1,958
Avg submissions/team	20.5	23.0	21.0
Phoneme Classification (Phase 2)			
Teams	47	15	62
Submissions	3,845	238	4,083
Avg submissions/team	81.8	15.9	65.9
Both phases combined			
Unique teams			155
Teams in both phases			7
Total submissions			6,041

Table 3: Participation statistics by task and track. Average submissions per team calculated as total submissions divided by the number of teams in that track.

Appendix D. Python Library Downloads

The `pnpl` Python library was downloaded 14,736 times between its release in April 2025 and March 2026. Figure 6 shows daily and cumulative downloads over this period, with red dashed lines marking the start of each competition phase and the NeurIPS workshop. Downloads peaked around the start of Phase 2 (August 2025), with a secondary peak in January 2026 following the NeurIPS workshop. Summary statistics and monthly breakdowns are provided in Tables 4 and 5.

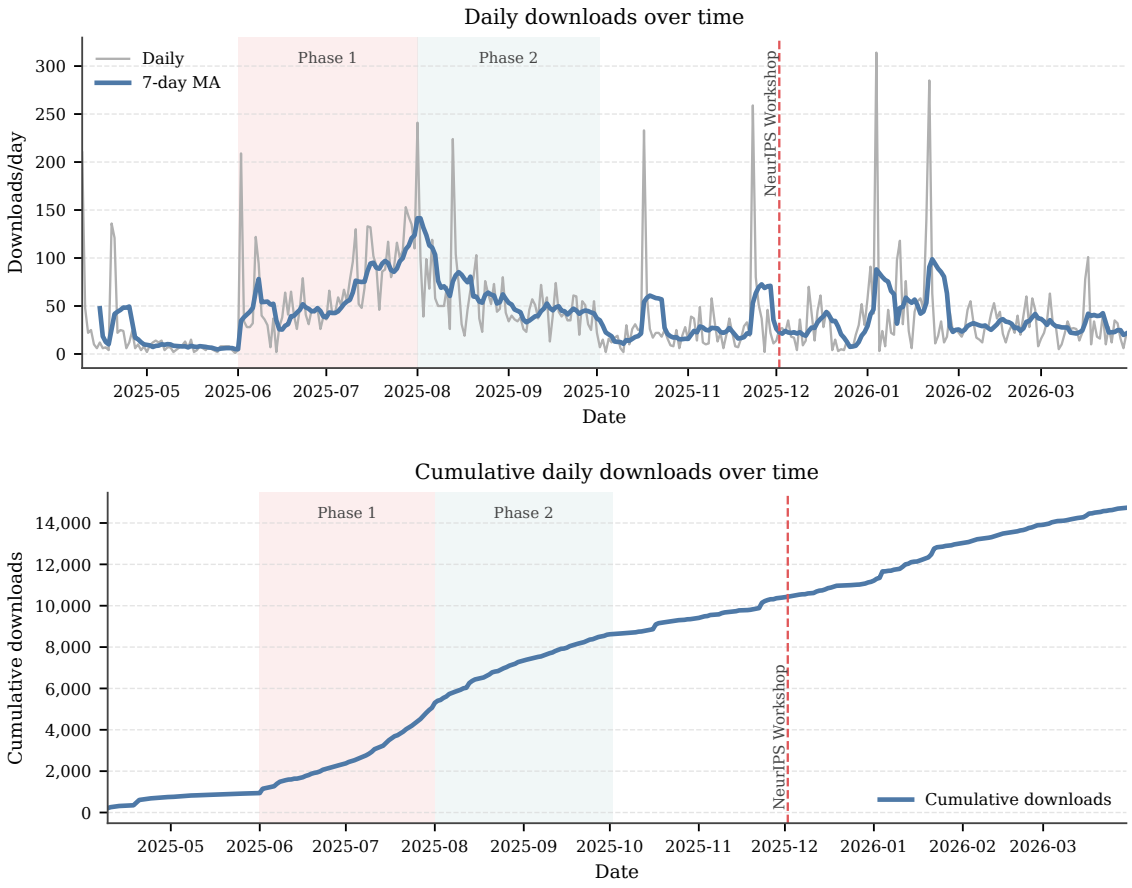

Figure 6: Daily (top) and cumulative (bottom) `pnpl` library downloads from release in April 2025 through March 2026. Red dashed lines indicate the start of Phase 1, Phase 2, and the NeurIPS workshop. The 7-day moving average (blue) is shown alongside raw daily counts (grey).

Table 4: Summary statistics for `pnpl` daily downloads (full history since release at time of writing).

Metric	**Value**
Date range	2025-04-09 to 2026-03-30
Days observed	350
Total downloads	14,736
Mean downloads/day	42.10
Median downloads/day	30.0
Std. deviation	44.06
p10 / p25 / p75 / p90	7.0 / 15.0 / 51.0 / 91.1
95% CI for mean	[37.49, 46.72]

Table 5: Monthly `pnpl` downloads (full history since release at time of writing).

Month	**Downloads**	**MoM**
2025-04	748	—
2025-05	188	−74.9%
2025-06	1,405	+647.3%
2025-07	2,722	+93.7%
2025-08	2,254	−17.2%
2025-09	1,280	−43.2%
2025-10	799	−37.6%
2025-11	997	+24.8%
2025-12	762	−23.6%
2026-01	1,859	+144.0%
2026-02	884	−52.4%
2026-03	838	−5.2%

Appendix E. Dataset Downloads

The `LibriBrain` dataset reached 16,412 recorded downloads by 28 April 2026. Figure 7 shows cumulative dataset downloads over time based on available snapshot counts, with red dashed lines marking the start of each competition phase and the NeurIPS workshop. Downloads grew especially quickly during the first competition months: from 757 recorded downloads on 13 June 2025 to 6,160 by 8 August 2025, and to 9,736 by 3 October 2025. By the end of March 2026, the dataset had surpassed 15,000 downloads, indicating sustained usage beyond the active competition period. Snapshot-level growth statistics are provided in Table 6.

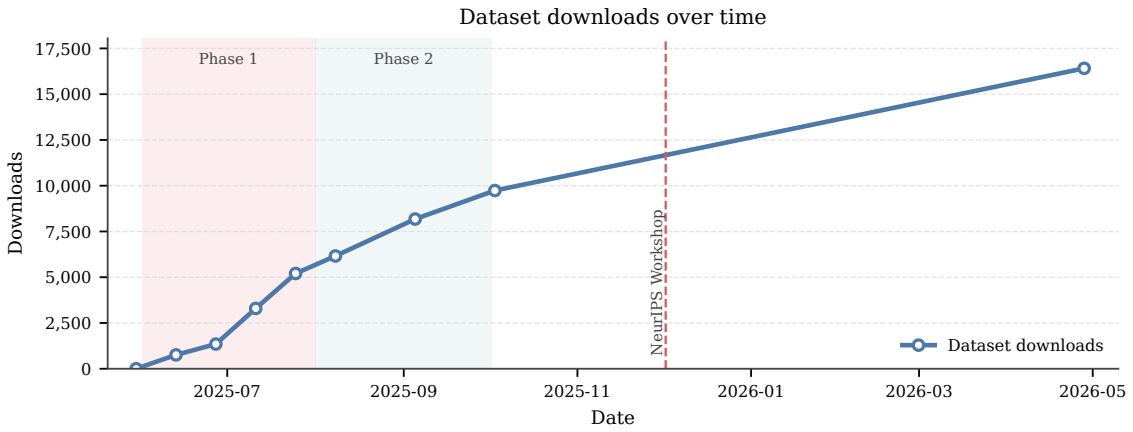

Figure 7: Cumulative `LibriBrain` dataset downloads from June 2025 through April 2026. Red dashed lines indicate the start of Phase 1, Phase 2, and the NeurIPS workshop. Markers indicate available download-count snapshots.

Table 6: Cumulative `LibriBrain` dataset downloads and growth between available snapshots.

Date	Downloads	Days since previous	Increase	Avg./day
2025-06-13	757	—	—	—
2025-06-27	1,348	14	591	42.2
2025-07-11	3,295	14	1,947	139.1
2025-07-25	5,204	14	1,909	136.4
2025-08-08	6,160	14	956	68.3
2025-09-05	8,178	28	2,018	72.1
2025-10-03	9,736	28	1,558	55.6
2026-03-28	15,086	176	5,350	30.4
2026-04-28	16,412	31	1,326	42.8