CoEEG: Benchmarking Collaborative Learning in EEG Decoding

Anonymous authors
Paper under double-blind review

Abstract

Electroencephalography (EEG) decoding is constrained by between-subject variability, scarce per-user labeled data, and privacy restrictions that scatter records across clinics and devices. In this work, we build **CoEEG** to jointly compare model families and collaborative learning protocols on three EEG paradigms, Motor Imagery (MI), SSVEP, and P300. Our benchmark covers three model families, traditional machine learning with paradigm-specific priors, deep models trained from scratch, and EEG foundation models. We compare ten widely used models under four protocols that span non-collaborative and collaborative training. We find that collaboration helps EEG decoding, but not uniformly. On SSVEP, nine of ten models benefit, consistent with a decision boundary that is largely shared across subjects. On P300, every model benefits only when each client keeps its classification head private. On MI, the gains concentrate in models that perform poorly on their own. Across the six federated cells the from-scratch family has the higher family mean, though a foundation model posts the single best score in two of them. The sharing boundary follows a measured factor, how far representations align, and an inferred one, whether the decision boundary is shared. A private head splits motor imagery and never lifts the weakest clients, who lose 20.7 percentage points (pp) on SSVEP against 9.3 pp for the strongest. Once parameter count, memory, and communication are included, lightweight models are the better performance–cost choice in most settings, and the largest foundation model transmits up to 378.73 GiB. Because the value of collaborative learning depends jointly on the (1) paradigm, (2) model, and (3) sharing boundary, **CoEEG** supports holistic evaluation and model selection for different EEG decoding tasks.
Page: https://anonymous.4open.science/w/coeegBench.
Code: https://anonymous.4open.science/r/coeegBench.

1 Introduction

Brain–computer interfaces (BCIs) decode brain signals into commands, most commonly from non-invasive electroencephalography (EEG) recorded at the scalp (Wolpaw et al., 2002). EEG distributions vary across individuals because of anatomy, brain state, hardware, and electrode contact. A rule learned for one subject may therefore fail for another, yet each user typically contributes only a few hundred labeled trials. Recordings are also dispersed across hospitals, laboratories, and personal devices under privacy and compliance constraints. The open question is how to increase this supervision without moving raw data.

Federated learning (McMahan et al., 2017) provides such a scheme: clients train locally and exchange model parameters with a server while retaining their raw data. The mapping to EEG is direct: each recorded subject becomes a client whose trials stay local. The first question is when such collaboration improves EEG decoding under realistic user-level heterogeneity, including for clients with weak independent-training baselines.

Collaboration does not remove user heterogeneity. Under non-independent and identically distributed (non-IID) clients, one global model may impose a decision boundary that fits no user well. Personalized federated learning instead partitions the model into shared and private parts. One common design shares the feature extractor while keeping classification heads client-specific

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

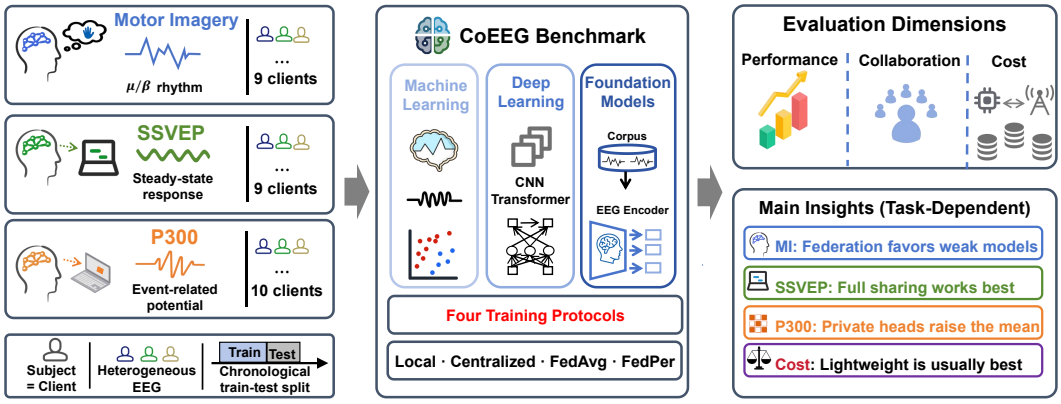

Figure 1: **Overview of CoEEG benchmark. Left**: three EEG paradigms (Motor Imagery, SSVEP, P300) with real subjects as clients and raw data staying at the clients (except the `Centralized` contrast, which pools data). **Middle**: three model families and the four training protocols. **Right**: evaluation dimensions and main insights.

(Arivazhagan et al., 2019). This motivates a separate question: *when should the decision layers of an EEG decoder remain local rather than be shared across users?*

Model choice forms another dimension. EEG decoders span traditional machine learning with paradigm-specific priors, deep networks trained from scratch, and foundation models pre-trained on multi-source EEG (Appendix A). We distinguish the two neural families by training provenance rather than architecture: foundation models transfer pre-trained representations, whereas the other neural models learn them from random initialization.

Recent EEG foundation-model benchmarks evaluate models under centralized, cross-subject, and within-subject few-shot settings (Lee et al., 2025; Liu et al., 2026). EEG federated-learning studies instead focus on new training or aggregation strategies over a limited set of tasks and neural backbones (Hang et al., 2023; Liu et al., 2025; Jia et al., 2024; Baykara et al., 2025). Neither line characterizes model–protocol interactions when the participating subjects are evaluated on their own held-out data. Model scale also matters: prior benchmarks show that many EEG foundation models are substantially larger than compact task-specific networks (Lee et al., 2025; Liu et al., 2026), while exact implementation-specific counts can vary (Appendix F). Federated protocols transmit shared parameters every round, so predictive performance must be assessed together with communication and deployment cost.

We therefore study these dimensions jointly in a subject-as-client setting. Each recorded subject forms one participating client, the corresponding trials remain local, and evaluation uses later, held-out data from the same subject (Figure 1). We organize the benchmark around four questions. **RQ1 (Collaboration):** across MI, SSVEP, and P300 decoding, when does collaboration outperform independent training? **RQ2 (Models):** what advantages and limitations emerge within and across model families, paradigms, and training protocols? **RQ3 (Personalization):** when should the decision layers of an EEG decoder remain local rather than be shared across users? We operationalize this question by comparing full-model sharing with a client-private classification head. **RQ4 (Cost):** how do parameter count, communication, and peak memory change the performance–cost trade-off?

To our knowledge, this is the first EEG decoding benchmark to jointly compare model families and collaboration protocols under a participating-subject-as-client evaluation on temporally held-out data from every client. Our contributions are threefold: (i) One evaluation protocol covering motor imagery, SSVEP, and P300, with temporal splits, fixed-final-round results, paired statistics, and worst-client analysis. (ii) Matched `Local`, `Centralized`, `FedAvg`, and `FedPer` comparisons for models trained from scratch and pre-trained foundation models, together with paradigm-specific traditional baselines evaluated under `Local` and `Centralized` and a cross-client representation analysis. (iii) Parameter count, peak memory, and theoretical communication reported alongside predictive performance. The central finding is that the value of collaboration depends jointly on the EEG paradigm, model, and sharing boundary.

2 CoEEG Benchmark

2.1 EEG Tasks and Settings

Let a dataset contain K subjects, each subject corresponding to a client k whose local data are split, following the original temporal order of sessions, runs, blocks, or trials, into an earlier training part and a later test part that never participates in training. The training data of client k are

$$\mathcal{D}_k^{\text{train}} = \{(\mathbf{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}, \qquad \mathbf{x}_i^{(k)} \in \mathbb{R}^{C\times T}, \tag{1}$$

where $n_k = |\mathcal{D}_k^{\text{train}}|$ is the number of training trials of client k, $\mathbf{x}_i^{(k)}$ is its i-th trial with C channels and T samples, and $y_i^{(k)}$ is the class label. The test data $\mathcal{D}_k^{\text{test}}$ are the later part, so the test set measures decoding performance on future data of the same user. BNCI2014_001 and Nakanishi2015 contain nine clients each, and BNCI2014_009 contains ten clients. All three datasets are trial-wise classification tasks.

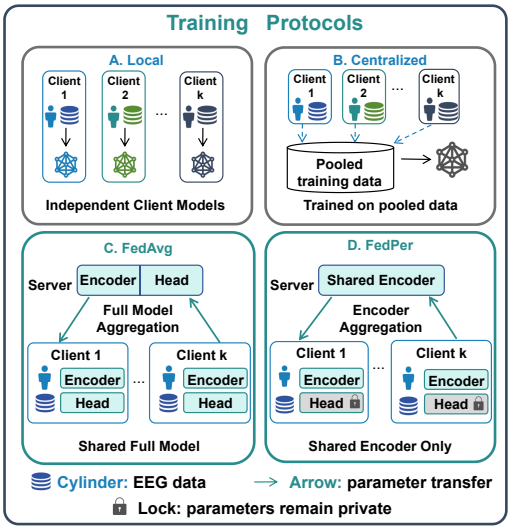

Figure 2: **EEG training protocols**. `Local` and `Centralized` compare with 2 `Fed` protocols.

2.2 Training Protocols

With model parameters $\mathbf{w}$ and loss ℓ, the empirical risk of client k is

$$F_k(\mathbf{w}) = \frac{1}{n_k} \sum_{(\mathbf{x},y)\in\mathcal{D}_k^{\text{train}}} \ell(f_{\mathbf{w}}(\mathbf{x}), y), \tag{2}$$

and the global objective is $F(\mathbf{w}) = \sum_{k=1}^{K} \frac{n_k}{N} F_k(\mathbf{w})$, where $N = \sum_{k=1}^{K} n_k$ is the total number of training trials. We compare four training protocols (Figure 2), all of them well known and widely used. Our aim is not to propose a new protocol but to compare established options under a common evaluation, covering two comparison settings and the two federated protocols in most common use, so that the benchmark stays close to standard practice:

- **`Local`:** each client trains only on its own data (the status-quo baseline without collaboration);
- **`Centralized`:** all training data are pooled into one model. This is infeasible under privacy constraints like EEG and serves only as a comparison setting;
- **`FedAvg` (federated averaging):** at round t each client initializes from the global parameters $\mathbf{w}_t$, optimizes locally to $\mathbf{w}_{t+1}^{(k)}$, and the server aggregates proportionally to sample counts (McMahan et al., 2017),

$$\mathbf{w}_{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} \mathbf{w}_{t+1}^{(k)}; \tag{3}$$

- **`FedPer` (personalized federated learning):** following (Arivazhagan et al., 2019), parameters are split into a shared representation $\boldsymbol{\phi}$ and client-private classification heads $\mathbf{h}_k$. Only the representation is aggregated,

$$\boldsymbol{\phi}_{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} \boldsymbol{\phi}_{t+1}^{(k)}, \qquad \mathbf{h}_k \text{ stays at client } k. \tag{4}$$

All four protocols use the same training setup: 200 federated rounds with one local epoch per client per round, and 200 epochs for `Local` and `Centralized`. Every client therefore traverses its local data equally often, and all clients participate in every round. Evaluation is on each client's own test set. Traditional machine-learning baselines rely on spatial filters, covariance, or template statistics, for which parameter averaging is undefined. They are reported only under `Local` and `Centralized`.

2.3 BASELINE MODELS

We evaluate three families with 13 models in total, divided by **how the discriminative representation is obtained** (the discriminative representation refers to the feature extracted from raw EEG before the classification head, which captures class-relevant information and yields compact within-class and separable between-class representations). **(1) Traditional machine learning (ML, one per paradigm)**: CSP+LDA for BNCI2014_001, TRCA for Nakanishi2015, and xDAWN+LDA for BNCI2014_009. All three classify with paradigm-specific priors plus hand-crafted features, exploiting known neurophysiological regularities such as band-power changes in motor imagery and frequency-locked SSVEP responses. **(2) Deep models trained from scratch (DL)**: five models trained in a supervised manner from random initialization, organized into a CNN subgroup (EEGNet, ShallowConvNet, LMDA-Net) and a Transformer subgroup (EEG-Conformer, Deformer). **(3) EEG foundation models (FM)**: five models pre-trained on large-scale multi-source EEG and then fine-tuned, namely CBraMod, EEGMamba, EEGPT, LaBraM, and Neuro-GPT. Foundation models start from officially released pre-trained weights, are fine-tuned with all parameters updated, and follow their official input adaptation (channel mapping, temporal patching). All classification heads are replaced with output layers matching the task's class count.

The partition rests on training provenance rather than network structure. The benchmark we extend (Liu et al., 2026) uses the same partition, and CNN and Transformer only organize the DL family internally. We run no separate statistical inference on these subgroups. The ML methods differ by paradigm and are deliberately chosen standard baselines, not statements about generalization.

2.4 EVALUATION METRICS

Each dataset uses the primary metric of Table 1. For each model, protocol, and seed (42/43/44), the metric is computed per client, averaged across clients, then across seeds. Single-model results are cross-seed mean $\pm$ std, and protocol-level results equally weight the ten neural models. Primary metrics have different scales across datasets, so we compare only directions, rankings, and interaction effects, never absolute scores.

To answer whether collaboration is broadly beneficial rather than merely positive on average, Section 3 complements protocol group means with **two-level paired statistics**: *model-level* descriptions (per-model differences $\Delta(\cdot - \texttt{Local})$ and $\Delta(\texttt{FedPer} - \texttt{FedAvg})$, plus the proportion of benefiting models) and *subject-level* inference (a two-sided Wilcoxon signed-rank test over the 9–10 independent subjects, the only independent sampling unit in our design). We additionally report a worst-client statistic and a tercile analysis over clients grouped by their independent-training (`Local`) score, with precise definitions and full test tables in Appendix C.

System cost is measured by parameter count, theoretical federated communication (FP32 shared-parameter bytes $\times$ ($2\times$ rounds $+1$) $\times$ clients, validated against per-run measurements in Appendix F.1), and per-run peak GPU memory. Wall-clock time serves only as a conditional reference (Appendix F). Appendix E additionally measures cross-client representation similarity with linear centered kernel alignment (CKA) (Kornblith et al., 2019), a normalized 0-to-1 similarity between two clients' representations.

2.5 IMPLEMENTATION DETAILS

Datasets and preprocessing. The three datasets cover the motor-imagery (MI), SSVEP, and P300 paradigms (Table 1). Data are obtained through MOABB (Chevallier et al., 2024) and frozen into a unified version, split within each client in original temporal order. Each model applies model-level preprocessing consistent with its reference implementation (its own sampling rate, frequency band, and normalization). The per-dataset specifications are given in Appendix F. Within a dataset, all models share the same frozen data, time windows, and client partition.

Training configuration. All neural models use a unified setup: AdamW (Loshchilov & Hutter, 2019) with cosine learning-rate scheduling (complete hyperparameters in Appendix F). The main result is fixed to the final round actually trained (the *fixed-final-round* protocol), without any model selection based on test curves. Traditional baselines are deterministic and run once.

Table 1: **Datasets, client partitions, and primary metrics**. BNCI2014_001 (Tangermann et al., 2012), Nakanishi2015 (Nakanishi et al., 2015), and BNCI2014_009 (Brunner et al., 2015; Riccio et al., 2013) are obtained through MOABB (Chevallier et al., 2024). Metrics: balanced accuracy (BA, the mean of per-class recalls), accuracy, and target-class AUPRC (area under the precision–recall curve, computed as the non-interpolated average precision of the target-class predicted probability). BNCI2014_009 has a roughly 1:5 class imbalance. Only the first recording session of each subject is used, and all splits are performed within each client in original temporal order.

Dataset	**Paradigm**	**Clients/Channels/Classes**	**Temporal split** (train/test trials)	**Metric** (chance)
BNCI2014_001	MI	9 / 22 / 4	Runs 1–5 / Run 6 (240/48 trials)	BA (25.0%)
Nakanishi2015	SSVEP	9 / 8 / 12	Blocks 1–10 / Blocks 11–15 (120/60 trials)	Accuracy (8.3%)
BNCI2014_009	P300	10 / 16 / 2	first 80% / last 20% (in time, 460/116 trials)	target-class AUPRC (16.7%)

Table 2: **Main fixed-final-round results** (%). Bold marks the highest full-precision mean per dataset and protocol, and ties are bolded jointly. "Rank" is the average rank of each model over the 12 cells (*3 datasets* × *4 protocols*), and ML exclusive. "Cent." abbreviates `Centralized`, the ML row is a single deterministic run, and "–" denotes not available.

		BNCI2014_001 (MI)				**Nakanishi2015 (SSVEP)**				**BNCI2014_009 (P300)**			
		BA (%)				Accuracy (%)				AUPRC (%)			
Model	Rank	Local	Cent.	FedAvg	FedPer	Local	Cent.	FedAvg	FedPer	Local	Cent.	FedAvg	FedPer
ML													
Paradigm-specific	–	**62.5**	43.5	–	–	**99.1**	98.1	–	–	24.0	57.0	–	–
DL													
EEG-Conformer	4.00	53.1	42.7	62.9	59.5	68.6	95.2	96.8	92.3	69.9	67.7	70.9	71.0
Deformer	4.17	46.3	40.4	41.1	52.5	**99.1**	**98.4**	96.4	**94.9**	72.0	66.0	69.4	78.1
EEGNet	4.58	49.8	40.1	54.2	57.3	77.1	94.9	97.0	85.1	66.9	60.3	**74.7**	**80.2**
LMDA	4.58	51.2	47.8	57.2	**59.8**	62.8	91.1	88.9	70.6	74.1	61.4	73.2	78.1
ShallowConvNet	5.42	60.4	50.9	57.1	56.4	50.8	69.0	71.4	48.0	70.4	68.0	71.4	71.3
FM													
CBraMod	3.83	60.4	**55.9**	52.2	58.5	84.5	97.1	**97.1**	91.7	65.2	63.7	66.0	77.1
Neuro-GPT	4.92	43.8	55.4	**67.4**	50.5	37.5	92.5	92.8	57.8	**74.1**	**72.3**	71.2	75.6
LaBraM	6.08	35.0	42.4	47.7	43.2	84.4	95.2	95.6	91.6	65.4	63.1	67.6	75.3
EEGPT	7.75	31.2	37.7	33.6	31.9	70.1	98.0	93.5	78.7	49.7	59.6	61.1	57.3
EEGMamba	9.67	36.1	35.0	31.7	34.4	22.5	89.6	65.6	42.5	41.7	47.7	43.2	42.1

3 Experimental Results

Table 2 summarizes the core results: the fixed-final-round primary metrics of 13 models across three datasets and four protocols, comparable only within the columns of each dataset. Bold marks the highest full-precision mean per dataset and protocol. Cross-seed standard deviations and the tie convention are reported in Table 10 (Appendix B.2).

3.1 RQ1: When Does Collaboration Outperform Independent Training?

> **Collaboration outperforms independent training on all three paradigms, though the protocols that do so differ by paradigm.**

On Nakanishi2015, `Centralized`, `FedAvg`, and `FedPer` all significantly outperform independent training (subject-level $p \approx 0.004$): nine of ten models benefit under `Centralized` and `FedAvg` and eight of ten under `FedPer`, with the locally saturated Deformer the sole exception under the first two. On BNCI2014_001, `FedPer` significantly outperforms independent training, `FedAvg` improves more weakly (Appendix C), and `Centralized` harms seven of ten models. On BNCI2014_009, `FedPer` is the only protocol that benefits every model, while `FedAvg` benefits seven of ten models yet is not subject-level significant ($p = 0.770$) and `Centralized` benefits two. Table 3 gives the protocol means and Table 15 the paired statistics and benefiting counts (full tests in Appendix C).

Table 3: **Protocol-level fixed-final-round results** (ten neural models equally weighted, %). Each dataset is scored with its primary metric from Table 1. Mean: cross-model mean ± std. The worst-client statistic is defined in Appendix C.1. Bold marks the highest protocol-level mean per dataset.

Dataset	Statistic	Local	Cent.	FedAvg	FedPer
BNCI2014_001 (MI)	Mean	46.7 ± 10.3	44.8 ± 7.3	$\mathbf{50.5 \pm 11.9}$	50.4 ± 10.4
	Worst client	31.0 ± 5.5	30.8 ± 8.7	32.8 ± 10.3	32.8 ± 5.5
Nakanishi2015 (SSVEP)	Mean	65.7 ± 23.3	$\mathbf{92.1 \pm 8.6}$	89.5 ± 11.5	75.3 ± 19.6
	Worst client	31.8 ± 28.5	75.6 ± 18.7	68.6 ± 19.5	42.8 ± 23.7
BNCI2014_009 (P300)	Mean	64.9 ± 10.8	63.0 ± 6.7	66.9 ± 9.2	$\mathbf{70.6 \pm 11.9}$
	Worst client	38.8 ± 9.2	46.9 ± 6.4	50.6 ± 7.8	44.6 ± 8.6

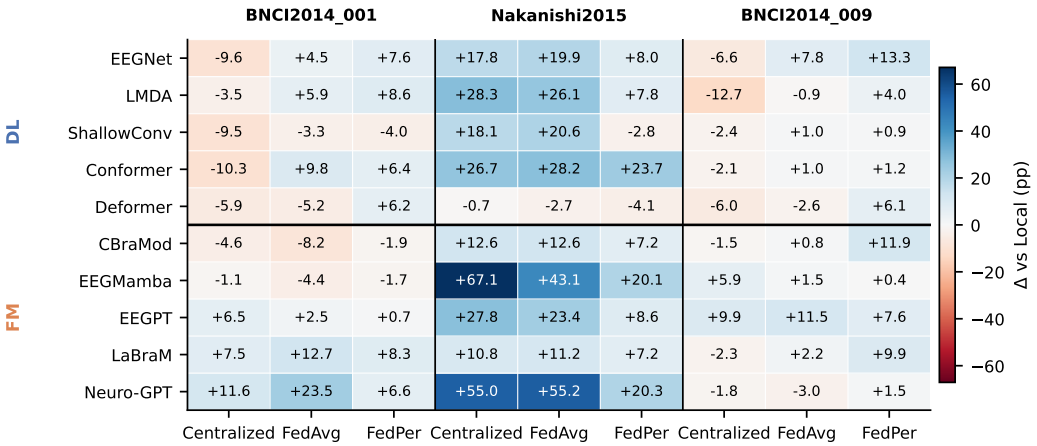

Figure 3: **Per-model protocol comparison** (heatmap, Δ against `Local`). DL methods are top five rows, while FM are bottom five. Columns include three datasets × three collaboration protocols.

Averages hide how unevenly collaboration acts. On BNCI2014_001, federated gains concentrate on models with weak locally trained baselines (Neuro-GPT, LaBraM), while the locally strongest (ShallowConvNet, CBraMod) are harmed: a shared prior helps weak models but does not guarantee surpassing strong locally trained baselines. On the `Centralized` side of BNCI2014_009 the pattern reverses: all five from-scratch models are harmed, whereas the template-based xDAWN+LDA gains 33.0 pp, consistent with P300 response templates transferring across users while discriminatively trained parameters do not. Whether `Centralized` training is feasible therefore depends on both task structure and model type. The worst clients nevertheless improve on SSVEP and P300, mainly through `Centralized` and `FedAvg`, and only marginally on motor imagery through the federated protocols (Figure 3, with per-client distributions in Figure 6 and Appendix B.1).

The per-client terciles (Appendix C.4) confirm the tilt toward difficult users: the `FedAvg−Local` gain falls from $+34.1$ pp in the weakest SSVEP tercile to $+13.8$ pp in the strongest, and from $+8.3$ pp for the weakest to -1.4 pp for the strongest on BNCI2014_009. Collaboration therefore pays off unevenly, so the next question is which model family to choose.

3.2 RQ2: What Advantages and Limitations Emerge Within and Across Model Families?

> **The from-scratch family is the more reliable choice in this model pool, leading on family mean in all six federated cells, though a foundation model posts the single best score in two of them.**

Table 4 (Appendix B.1) shows that the DL family mean exceeds the FM family mean in all six federated cells, by $+8.0/+13.4$ pp on MI, $+1.2/+5.7$ pp on SSVEP, and $+10.1/+10.3$ pp on P300 under `FedAvg`/`FedPer`; a leave-one-model-out check (Table 5) holds in four of the six cells. This

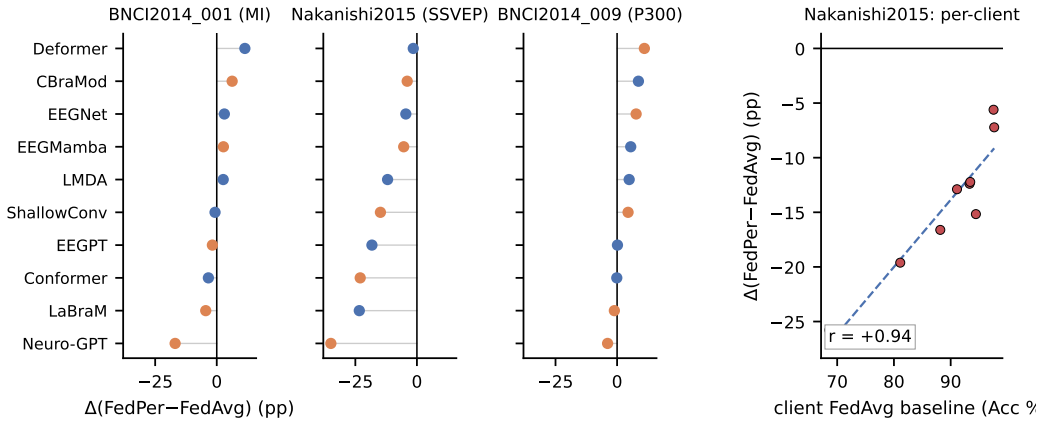

Figure 4: **Analysis of decision layers** (RQ3). (a) Per-model $\Delta(\texttt{FedPer} - \texttt{FedAvg})$ in % (● DL, ● FM). (b) On Nakanishi2015, per-client $\Delta(\texttt{FedPer} - \texttt{FedAvg})$ against `FedAvg` (Pearson $r = +0.94$).

is a family-level tendency rather than a per-cell guarantee: the single best model in two of the six cells is still a foundation model. Across the twelve dataset–protocol cells no family is excludable, since DL holds the highest family mean in eight and the three best combinations span all three families.

The FM family is the most internally heterogeneous: on the Nakanishi2015 `Local` column it spans 22.5% (EEGMamba) to 84.5% (CBraMod), a range above 60 pp, and its best- and worst-ranked models are both FM while the five DL models cluster between 4.0 and 5.4 (Table 2). This is consistent with the centralized-setting observation that specialists remain competitive while foundation-model performance varies across tasks and domains (Lee et al., 2025).

Two family-level regularities remain (Table 6, Appendix B.1). **Traditional methods with strong paradigm priors are highly competitive**: CSP+LDA at 62.5% under BNCI2014_001 `Local` exceeds all neural models, and TRCA is nearly saturated on Nakanishi2015. In such regimes the case for complex neural models rests on room for optimization and extension rather than current decoding performance. **Pre-trained representations are more robust under heterogeneous `Centralized` training**: on MI and P300 none of the five from-scratch models benefits from `Centralized`, whereas 2–3 FM members do (Figure 3). This robustness does not carry over to federation: on BNCI2014_009 the DL family mean exceeds the FM family mean under all four protocols (Table 4), suggesting that under limited data and class imbalance the faster adaptation of compact models matters more than pre-training, a mechanism we do not measure directly.

3.3 RQ3: When Should EEG Decision Layers Remain Local?

> **The classification head should stay shared on SSVEP; on P300 a private head raises the mean without reaching significance and lowers the worst client; on MI the choice is per-model. The sharing boundary follows a measured factor, how far representations align, and an inferred one, whether the decision boundary is shared.**

The `FedAvg`–`FedPer` contrast tests *minimal personalization*: the two differ only in whether the classification head participates in aggregation. We call the choice of which parameters are aggregated the *sharing boundary*, which for `FedPer` falls at the classification head, whose partition of feature space we call the *decision boundary*. Figure 4 shows the per-model differences $\Delta(\texttt{FedPer} - \texttt{FedAvg})$ (exact values in Appendix B.4) and, on SSVEP, their client-level association.

The three datasets give three different answers. **On BNCI2014_009 (P300), personalization benefits**: the mean is $+3.7$ pp with seven of ten models positive (Table 14), `FedPer` is the best protocol for seven of ten models and has the highest group mean of the four protocols, and it is the only proto-

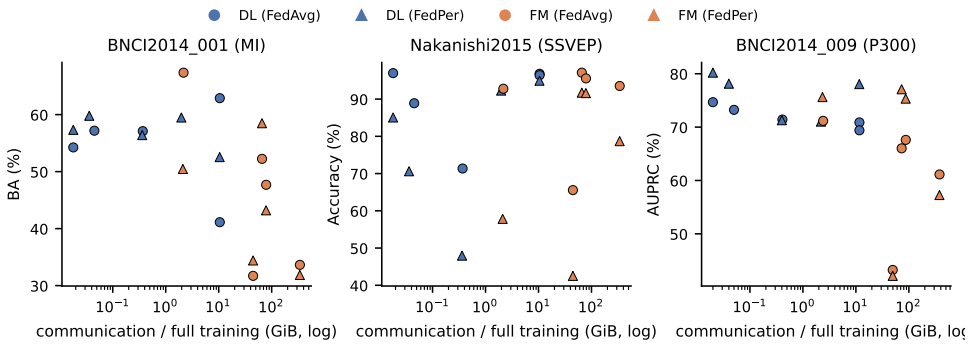

Figure 5: **Analysis of performance-communication trade off** (RQ4). The x-axis is the communication cost of training (log scale), the y-axis the fixed-final-round primary metric. Circles (○) denote `FedAvg` and triangles (△) `FedPer`, with • DL and • FM.

col under which every neural model matches or beats its own `Local` (Table 15). **On BNCI2014_001 (motor imagery) the effect splits by backbone**: five models gain and five lose, averaging -0.1 pp, because the model that gains most from aggregation (Neuro-GPT) loses most from a private head while the Deformer, which aggregation harms, gains most. Models with large collaboration gains should therefore not split off a private head, whereas models with strong local performance can benefit from one. **On Nakanishi2015 (SSVEP) personalization is uniformly harmful**: all ten models fall relative to `FedAvg`, with all nine subjects negative ($p = 0.004$).

On Nakanishi2015, the private head does not help the users who need it most. The `FedPer` worst-client mean of Table 3 is **never** better than `FedAvg` on any dataset, and on Nakanishi2015 the `FedPer`$-$`FedAvg` difference correlates positively with the client's `FedAvg` baseline ($r = +0.94$): the weakest tercile loses 20.7 pp on average against 9.3 pp for the strongest, and the worst client falls from 68.6% under `FedAvg` to 42.8% under `FedPer`. On the other two datasets the terciles show no systematic tilt (Appendix C.4, Appendix C.3), and on BNCI2014_009 the weakest tercile in fact gains $+7.1$ pp. On Nakanishi2015 the pattern suggests that client difficulty stems from insufficient shareable information rather than a mismatched local decision boundary.

The measured factor is cross-client representation alignment, quantified with linear CKA over the clients' penultimate representations. The medians over the ten neural models are 0.45, 0.58, and 0.76 under `FedAvg` on MI, SSVEP, and P300, against 0.17, 0.22, and 0.32 under `Local` (Appendix E). If alignment were the only factor, the most alignable paradigm should benefit least from a private head. Yet on BNCI2014_009 (P300), the most alignable paradigm, a private head still helps seven of ten models over a shared one. This points to a second, inferred factor, whether the decision boundary is shared across subjects. It cannot be measured because each client is evaluated on its own held-out trials and there is no common evaluation set. Under that inference, a private head helps when representations align but boundaries differ across subjects (P300), hurts when boundaries agree (SSVEP), and splits motor imagery by backbone.

3.4 RQ4: How Do Predictive Performance and System Overheads Trade Off?

> **Lightweight models are the better performance–cost choice in most of the evaluated settings.**

On all three datasets, the extra performance gained by scaling up the model mostly fails to cover its communication, memory, and deployment costs. The lightweight models (EEGNet, LMDA, ShallowConvNet) occupy the low-communication–high-performance region in every dataset–protocol combination, and the best combination on BNCI2014_009, EEGNet+`FedPer`, has only 1,554 parameters, 0.02 GiB of theoretical communication, and 0.6 GB of peak memory. The cost spread among the best combinations is extreme: on BNCI2014_001 Neuro-GPT+`FedAvg` transmits 2.24 GiB, two orders of magnitude below the largest foundation model (EEGPT, 341.74 GiB).

When lightweight models reach comparable or higher performance, the larger model no longer offers an advantage. Figure 5 shows the resulting landscape, and Appendix F gives the full cost values.

4 Discussion and Limitations

4.1 Training Dynamics and the Final-Round Protocol

The best round (the highest cross-client mean score of each run) differs from the fixed final round by 2.21 pp on BNCI2014_001, 0.59 pp on Nakanishi2015, and 3.58 pp on BNCI2014_009. Because it is selected from test curves and carries optimism bias, we use the fixed final round throughout. Appendix D reports the training curves and the full degradation statistics.

4.2 Practical Guidance

The family means of RQ2 make from-scratch models the safer default under federation when no validation signal is available. Protocol effects otherwise depend on the specific model–task combination (Table 15, Figure 3), so a sensible procedure is to identify the signal structure reusable across users and then select along the performance–cost front. The three paradigms give directly usable guidance: for SSVEP-like stimulus-locked paradigms, evaluate protocols that share the full model and traditional template methods first. For P300-like sparse event-related potentials, evaluate compact protocols that keep the classification head local first. For motor-imagery-like spontaneous-rhythm paradigms, verify shared-representation quality and individual adaptation simultaneously. The classification head is not necessarily the optimal private module, so future personalization algorithms should explore private components beyond it.

4.3 Limitations

This work covers three datasets, ten neural models, one traditional baseline per paradigm, and 9–10 clients, so it cannot represent all foundation models, recording hardware, clinical tasks, or larger federations. The unified schedule (a learning rate of 10^{-3}, 200 rounds) strains high-capacity models on small clients, where EEGMamba and Neuro-GPT nearly memorize Nakanishi2015 `Local` while testing at 22–38%. The family-level comparisons of RQ2 therefore describe this unified-default track over this model pool, not per-model fair adaptation. In both federated protocols, BatchNorm running statistics stay client-local, while the two differ only in whether the classification head is aggregated. Client-level analyses pool multiple models, so individual models may deviate. The experiments use no separate validation set, so the best-round analysis is diagnostic only, and BNCI2014_009 is scored with target-class AUPRC, which may differ from some prior work. The CKA analysis uses penultimate representations from a single seed, compares clients whose trials are not paired, and is valid only within-model, cross-protocol. The paradigm differences are empirical associations, explained at the representation level rather than through neurophysiological causation.

5 Conclusion

In this work, we present a new collaborative learning benchmark for EEG decoding, **CoEEG** for short. We compare three model families and four training protocols in a subject-level collaborative EEG setting, with paired statistics, worst-client analysis, and representation-level evidence. **(1) Collaboration helps, but its benefit depends on task structure.** Sharing is close to universally beneficial on SSVEP and concentrates on the weakest clients, and P300 improves all models only when the classification head stays private. Motor imagery lies in between. **(2) The from-scratch family is the more reliable default in this model pool.** It leads on family mean in all six federated cells, though a foundation model posts the single best score in two of them. **(3) The sharing boundary follows a measured factor, how far representations align, and an inferred one, whether the decision boundary is shared.** Private-head personalization improves the P300 mean, splits motor imagery, reduces SSVEP performance, and never lifts the worst-client mean over `FedAvg`. **(4) When parameter count, memory, and communication are included, lightweight models are usually the better performance–cost choice.**

AI USE STATEMENT

In this work, we used generative AI tools for manuscript writing assistance: drafting parts of the paper, translating author-prepared drafts between Chinese and English, editing the manuscript to improve readability, and refining word choice and phrasing. Additionally, we used generative AI tools to assist with implementing and debugging experiment and analysis code and with generating data plots. We have not used generative AI tools for the other tasks that require disclosure, and the remaining such tasks are not applicable to this work. All AI-assisted work was reviewed and revised by the authors. Reported numbers were verified against the experimental records, and references were checked against their primary sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly available human EEG datasets (BNCI2014_001, Nakanishi2015, BNCI2014_009) obtained through the MOABB framework. No new human-subject data were collected. We redistribute no raw recordings; access and preprocessing scripts are provided instead, and any reuse remains subject to the terms of the original dataset releases. Because only model parameters are exchanged, no raw EEG leaves a client under the federated protocols, but this is an architectural property rather than a formal privacy guarantee: we evaluate no membership-inference, reconstruction, or differential-privacy mechanism. The benchmark is intended for research on decoding models and federated protocols and does not provide clinical decision support.

REPRODUCIBILITY STATEMENT

All datasets are publicly available and obtained through MOABB. Section 2.5 specifies the datasets and preprocessing, and Section 2.4 defines all metrics, aggregation conventions, and statistical tests. The training configuration and system environment are specified in Appendix F. Appendices B–F report per-seed and per-client results, secondary metrics, full statistical tests, training dynamics, representation analyses, and measured system costs. Every table and statistic in the paper is computed by scripts from the per-round evaluation records and per-client final-round metrics. An anonymized code package is included as supplementary material. The code and a companion project page with an overview of the benchmark and its results are available at `https://anonymous.4open.science/r/coeegBench/` and `https://anonymous.4open.science/w/coeegBench/`.

REFERENCES

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

C. A. Baykara, S. R. Pandey, A. B. Ünal, H. Lee, and M. Akgün. Federated learning for epileptic seizure prediction across heterogeneous EEG datasets. *arXiv preprint arXiv:2508.08159*, 2025.

C. Brunner et al. BNCI horizon 2020: Towards a roadmap for the BCI community. *Brain-Computer Interfaces*, 2(1):1–10, 2015. doi: 10.1080/2326263X.2015.1008956.

S. Chevallier et al. The largest EEG-based BCI reproducibility study for open science: The MOABB benchmark. *arXiv preprint arXiv:2404.15319*, 2024.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 2089–2099, 2021.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-GPT: Towards a foundation model for EEG. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024. doi: 10.1109/ISBI56570.2024.10635453.

Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. EEG-Deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE*

Journal of Biomedical and Health Informatics, 29(3):1909–1918, 2025. doi: 10.1109/JBHI.2024.3504604.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Wenlong Hang, Jiaxing Li, Shuang Liang, Yuan Wu, Baiying Lei, Jing Qin, Yu Zhang, and Kup-Sze Choi. FedEEG: Federated EEG decoding via inter-subject structure matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. doi: 10.1109/ICASSP49357.2023.10095564.

He He and Dongrui Wu. Spatial filtering for brain computer interfaces: A comparison between the common spatial pattern and its variant. *arXiv preprint arXiv:1808.06533*, 2018.

Tianwang Jia, Lubin Meng, Siyang Li, Jiajing Liu, and Dongrui Wu. Federated motor imagery classification for privacy-preserving brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3442–3451, 2024. doi: 10.1109/TNSRE.2024.3457504.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations (ICLR)*, 2025.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 5132–5143, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 3519–3529, 2019.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659.

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.

Na Lee, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Are large brainwave foundation models capable yet? Insights from fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 32878–32888. PMLR, 2025.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG-FM-Compass: Progress, benchmarking, and future directions for EEG foundation models. *National Science Review*, 2026. doi: 10.1093/nsr/nwag466.

Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. mixEEG: Enhancing EEG federated learning for cross-subject EEG classification with tailored mixup. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 1273–1282, 2017.

Zhengqing Miao, Meirong Zhao, Xin Zhang, and Dong Ming. LMDA-Net: A lightweight multi-dimensional attention network for general EEG-based brain–computer interfaces and interpretability. *NeuroImage*, 276:120209, 2023. doi: 10.1016/j.neuroimage.2023.120209.

Masaki Nakanishi, Yijun Wang, Yu-Te Wang, and Tzyy-Ping Jung. A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. *PLOS ONE*, 10(10):e0140703, 2015. doi: 10.1371/journal.pone.0140703.

Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, 65(1):104–112, 2018. doi: 10.1109/TBME.2017.2694818.

Angela Riccio, Luca Simione, Francesca Schettini, Alessia Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, Donatella Mattia, and Febo Cincotti. Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Frontiers in Human Neuroscience*, 7: 732, 2013. doi: 10.3389/fnhum.2013.00732.

Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xDAWN algorithm to enhance evoked potentials: Application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009. doi: 10.1109/TBME.2009.2012869.

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.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

M. Tangermann et al. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280. Curran Associates, Inc., 2024. doi: 10.52202/079017-1239.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. EEGMamba: An EEG foundation model with mamba. *Neural Networks*, 192:107816, 2025b. doi: 10.1016/j.neunet.2025.107816.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002. doi: 10.1016/S1388-2457(02)00057-3.

Daniel Adu Worae and Aarthy Nagarajan. NEXUS-MI: Communication-aware federated personalization for gateway-coordinated motor-imagery brain-computer interfaces. *arXiv preprint arXiv:2609.09786*, 2026.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Related Work

The development of EEG decoding methods can be summarized as three complementary lines. Traditional machine-learning methods explicitly exploit known signal regularities of each paradigm: CSP with LDA for motor imagery (He & Wu, 2018), TRCA with template matching for SSVEP (Nakanishi et al., 2018), and xDAWN for P300 spatial filtering (Rivet et al., 2009). These methods have few parameters and train stably. For each paradigm we select one competitive traditional method as a baseline against which the practical benefit of complex neural models can be judged. Deep models trained from scratch learn discriminative representations directly with neural networks: EEGNet realizes a compact architecture with depthwise-separable temporal–spatial convolutions (Lawhern et al., 2018), ShallowConvNet turns the spectral–spatial priors of CSP into learnable shallow convolutions (Schirrmeister et al., 2017), LMDA-Net balances performance and interpretability with lightweight attention (Miao et al., 2023), EEG-Conformer combines a convolutional front end with self-attention (Song et al., 2023), and Deformer designs a decoding structure around the temporal hierarchy of EEG (Ding et al., 2025).

EEG foundation models follow the large-model recipe: pre-train on large-scale multi-source EEG with self-supervised objectives, then transfer by fine-tuning. BENDR introduced wav2vec-style contrastive prediction to neural signals (Kostas et al., 2021), LaBraM learns cross-dataset representations by predicting discretized tokens (Jiang et al., 2024), CBraMod models spatial and temporal dependencies in parallel via criss-cross attention (Wang et al., 2025a), EEGPT uses a masked Transformer to learn task-general representations (Wang et al., 2024), Neuro-GPT explores GPT-style autoregressive pre-training (Cui et al., 2024), and EEGMamba adopts a Mamba-style state-space model (Gu & Dao, 2023; Wang et al., 2025b) for efficient long sequences. Related efforts include BIOT for multi-modal biosignals (Yang et al., 2023) and NeuroLM, which unifies language and EEG signals (Jiang et al., 2025). Two recent benchmark studies (Lee et al., 2025; Liu et al., 2026) systematically compared foundation models with specialist baselines, the former under centralized subject-independent cross-validation and the latter under leave-one-subject-out and within-subject few-shot settings. Both found that linear probing is often insufficient, specialists trained from scratch remain competitive on many tasks, and larger models are not necessarily better. We adopt their training-source-based family partition and push this comparison to user-level federated settings with a system-cost dimension.

Federated learning iterates between local updates and server aggregation, optimizing a shared model without pooling raw data (McMahan et al., 2017). `FedAvg`, which averages all model parameters weighted by local sample counts, is the most basic scheme. Under highly non-IID clients the unified global model suffers from client drift, which SCAFFOLD-type methods characterize and correct from an optimization perspective (Karimireddy et al., 2020). Such methods still converge to a single global model and complement the personalized line below. Personalized federated learning partitions parameters into shared and private parts: `FedPer` shares the representation layers and keeps classification heads local (Arivazhagan et al., 2019), FedBN keeps client-private normalization statistics (Li et al., 2021), and FedRep gives a theoretical account of the "shared representation + local head" split (Collins et al., 2021). We adopt `FedAvg` and `FedPer` as the two representative points on the sharing boundary, so that "unified model versus local adaptation" becomes a controlled comparison rather than being entangled with more complex personalization algorithms.

EEG-specific federated learning methods have begun to appear: FedEEG matches inter-subject class-center structure to correct client drift (Hang et al., 2023), mixEEG designs an EEG-aware mixup augmentation for cross-subject classification (Liu et al., 2025), FedBS uses client-private batch normalization and sharpness-aware minimization to reduce between-user discrepancies (Jia et al., 2024), and federated learning has been applied to cross-dataset seizure prediction in a simulated multi-site setting (Baykara et al., 2025). These works focus on improving aggregation or training strategies on a few tasks and models, typically evaluating on random splits, held-out subjects, or held-out institutions rather than future data from participating clients. Closer to our setting, NEXUS-MI personalizes a shared backbone with subject-local heads under gateway communication constraints and tests later-session data of each subject, but studies synchronization policy rather than the interaction of model families and sharing protocols (Worae & Nagarajan, 2026). Complementary to them, we propose no new algorithm. We build a measurement benchmark spanning three paradigms, ten neural models, and one paradigm-specific traditional baseline per paradigm, answering a more fundamental question: how do collaboration, personalization, and model selec-

tion interact in EEG? Unlike existing EEG benchmarks (Lee et al., 2025; Liu et al., 2026), we take each real subject as a client, fix four protocols, evaluate each participating client on temporally held-out data from the same subject, and measure the interaction between protocols and models while reporting system costs alongside predictive performance.

B Complete Results and Robustness

B.1 Supplementary Tables and Figures for the Main Text

This subsection collects supplementary views referenced from the main text: per-client score distributions (Figure 6), family-level means (Table 4), and `Centralized` success counts (Table 6).

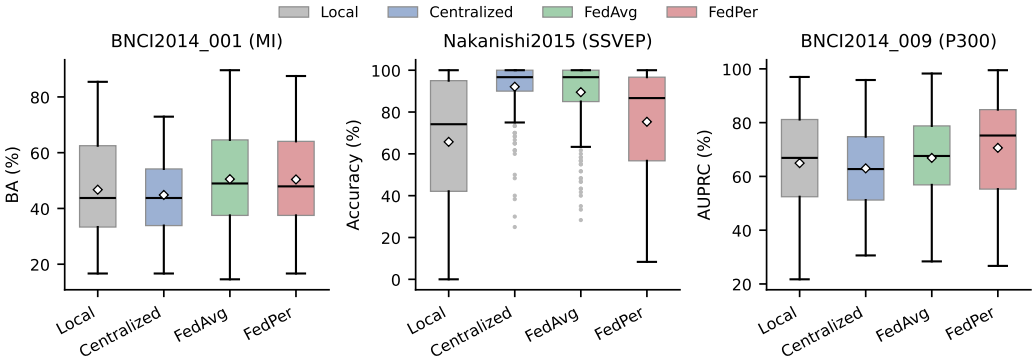

Figure 6: **Per-client primary-metric distributions** under the four protocols. Panels from left to right: BNCI2014_001 (BA), Nakanishi2015 (Accuracy), BNCI2014_009 (AUPRC). Each box pools the final-round scores of 10 neural models $\times$ 3 seeds $\times$ all clients (about 270–300 points, where the diamond marks the mean and matches the mean row of Table 3). Boxes span the interquartile range (IQR), the thick line is the median, whiskers extend to 1.5$\times$IQR, and gray points are outliers. These are descriptive distributions of a pooled, dependent sample, not confidence intervals, and the spread they show mixes between-model, between-seed, and between-client variation.

Table 4: **Family means** (%). ML is the paradigm-specific baseline (single run). DL and FM are within-family means of their five members. Bold marks the highest value within each dataset and protocol.

Dataset	Family	Local	Centralized	FedAvg	FedPer
BNCI2014_001 (MI)	ML	**62.5**	43.5	–	–
	DL	52.2	44.4	**54.5**	**57.1**
	FM	41.3	**45.3**	46.5	43.7
Nakanishi2015 (SSVEP)	ML	**99.1**	**98.1**	–	–
	DL	71.7	89.7	**90.1**	**78.2**
	FM	59.8	94.5	88.9	72.5
BNCI2014_009 (P300)	ML	24.0	57.0	–	–
	DL	**70.7**	**64.7**	**71.9**	**75.7**
	FM	59.2	61.3	61.8	65.5

Table 5: **Leave-one-model-out sensitivity of the family comparison**. Each row removes one member of a family at a time and recomputes that family's mean; the range spans the five removals of that family. "Sign change" counts removals after which the DL–FM gap changes sign. Computed from the tabulated one-decimal means of Table 2.

Dataset	Protocol	Gap	DL LOMO range	FM LOMO range	Sign change
BNCI2014_001	`FedAvg`	$+8.0$	52.4–57.9	41.3–50.2	0/10
BNCI2014_001	`FedPer`	$+13.4$	56.4–58.3	40.0–46.7	0/10
Nakanishi2015	`FedAvg`	$+1.2$	88.4–94.8	86.9–94.8	4/10
Nakanishi2015	`FedPer`	$+5.7$	74.0–85.7	67.7–80.0	1/10
BNCI2014_009	`FedAvg`	$+10.1$	71.2–72.6	59.5–66.5	0/10
BNCI2014_009	`FedPer`	$+10.3$	74.6–76.9	62.6–71.3	0/10

The equal-weight family means of Table 4 are averages over five chosen members, not samples from either model population, so we also report a leave-one-model-out (LOMO) check (Table 5). Four of the six federated cells keep the DL lead under every single-member removal. The two Nakanishi2015 cells do not: there the `FedAvg` margin of $+1.2$ pp reverses both when the weakest FM member (EEGMamba, 65.6%) is removed, which lifts the FM mean to 94.8%, and when any of the three strongest DL members (EEG-Conformer, Deformer, EEGNet) is removed, which lowers the DL mean to 88.4–88.5%, and the `FedPer` margin of $+5.7$ pp reverses only when EEGMamba is removed (FM 80.0% against DL 78.2%). The family *medians* are steadier than the means and keep the DL lead in all six cells (MI $+9.4$/$+14.1$, Nakanishi2015 $+2.9$/$+6.4$, BNCI2014_009 $+5.4$/$+2.8$ pp under `FedAvg`/`FedPer`). The SSVEP margin should therefore be read as composition-dependent, and the RQ2 claim is scoped accordingly: it holds under this unified-default track and over this model pool, tabulated per cell, rather than as a population-level ordering.

Table 6: **Family $\times$ `Centralized` success counts**. "X/5 benefit" counts the models of a family whose `Centralized` beats their own `Local`. The mean Δ values, including the single-model ML reference, are in pp.

Dataset	DL family (5 models)	FM family (5 models)	ML reference
BNCI2014_001 (MI)	0/5 benefit, mean -7.8	3/5 benefit, mean $+4.0$	CSP+LDA -19.0
Nakanishi2015 (SSVEP)	4/5 benefit, mean $+18.0$	5/5 benefit, mean $+34.7$	TRCA -0.9
BNCI2014_009 (P300)	0/5 benefit, mean -6.0	2/5 benefit, mean $+2.1$	xDAWN+LDA $+33.0$

B.2 Per-Seed Final-Round Primary Metrics

Tables 7, 8, and 9 give the cross-client mean final-round primary metric for each (model $\times$ protocol $\times$ seed) on BNCI2014_001, Nakanishi2015, and BNCI2014_009. They are the per-seed expansion of the mean $\pm$ standard deviations in Table 10.

Table 7: **BNCI2014_001 (MI), per-seed BA** (%). "–" marks deterministic ML baselines (single run). The traditional baselines are deterministic and run once. Their values are listed in the Seed 42 column only for layout alignment and do not involve a random seed.

Model	Family	Method	Seed 42	Seed 43	Seed 44
CSP+LDA	ML	Local	62.5	–	–
CSP+LDA	ML	Centralized	43.5	–	–
EEGNet	DL	Local	51.6	49.5	48.1
EEGNet	DL	Centralized	38.7	41.7	40.0
EEGNet	DL	FedAvg	52.8	55.1	54.9
EEGNet	DL	FedPer	56.3	60.0	55.8
LMDA	DL	Local	49.8	52.8	51.2
LMDA	DL	Centralized	48.8	47.2	47.2
LMDA	DL	FedAvg	57.4	57.6	56.5
LMDA	DL	FedPer	60.0	58.3	61.1
ShallowConvNet	DL	Local	59.7	62.0	59.5
ShallowConvNet	DL	Centralized	49.8	51.2	51.9
ShallowConvNet	DL	FedAvg	54.9	58.6	57.9
ShallowConvNet	DL	FedPer	55.1	58.3	55.8
EEG-Conformer	DL	Local	54.6	52.5	52.1
EEG-Conformer	DL	Centralized	46.1	40.3	41.9
EEG-Conformer	DL	FedAvg	62.5	63.4	62.7
EEG-Conformer	DL	FedPer	62.0	57.4	59.0
Deformer	DL	Local	47.2	44.4	47.2
Deformer	DL	Centralized	41.4	41.4	38.4
Deformer	DL	FedAvg	41.0	42.1	40.3
Deformer	DL	FedPer	53.2	51.2	53.2
CBraMod	FM	Local	59.5	60.6	61.1
CBraMod	FM	Centralized	55.8	57.9	53.9
CBraMod	FM	FedAvg	52.8	52.5	51.4
CBraMod	FM	FedPer	58.8	59.7	56.9
EEGMamba	FM	Local	38.4	36.6	33.3
EEGMamba	FM	Centralized	33.6	37.3	34.3
EEGMamba	FM	FedAvg	32.9	30.3	31.9
EEGMamba	FM	FedPer	34.7	36.3	32.2
EEGPT	FM	Local	32.4	31.0	30.1
EEGPT	FM	Centralized	39.6	33.6	39.8
EEGPT	FM	FedAvg	36.3	29.4	35.2
EEGPT	FM	FedPer	34.5	25.9	35.2
LaBraM	FM	Local	35.2	36.6	33.1
LaBraM	FM	Centralized	44.9	39.6	42.8
LaBraM	FM	FedAvg	49.3	46.3	47.5
LaBraM	FM	FedPer	45.1	45.8	38.7
Neuro-GPT	FM	Local	44.4	42.6	44.4
Neuro-GPT	FM	Centralized	56.7	52.5	56.9
Neuro-GPT	FM	FedAvg	68.1	65.7	68.3
Neuro-GPT	FM	FedPer	51.4	47.7	52.3

Table 8: **Nakanishi2015 (SSVEP), per-seed accuracy** (%). The traditional baseline (TRCA) is deterministic and run once. Its values are listed in the Seed 42 column only for layout alignment and do not involve a random seed.

Model	Family	Method	Seed 42	Seed 43	Seed 44
TRCA	ML	Local	99.1	–	–
TRCA	ML	Centralized	98.1	–	–
EEGNet	DL	Local	74.1	83.1	74.1
EEGNet	DL	Centralized	92.0	96.1	96.5
EEGNet	DL	FedAvg	96.5	96.9	97.6
EEGNet	DL	FedPer	84.3	83.7	87.2
LMDA	DL	Local	57.0	66.1	65.2
LMDA	DL	Centralized	89.4	92.0	91.9
LMDA	DL	FedAvg	88.9	89.8	88.0
LMDA	DL	FedPer	70.9	65.0	75.9
ShallowConvNet	DL	Local	48.9	51.3	52.2
ShallowConvNet	DL	Centralized	68.9	69.3	68.7
ShallowConvNet	DL	FedAvg	71.7	72.8	69.6
ShallowConvNet	DL	FedPer	47.0	48.5	48.3
EEG-Conformer	DL	Local	70.6	66.7	68.5
EEG-Conformer	DL	Centralized	95.7	94.8	95.2
EEG-Conformer	DL	FedAvg	95.9	97.2	97.2
EEG-Conformer	DL	FedPer	91.1	94.3	91.5
Deformer	DL	Local	98.5	99.4	99.3
Deformer	DL	Centralized	98.9	97.0	99.3
Deformer	DL	FedAvg	91.5	98.9	98.9
Deformer	DL	FedPer	94.8	94.4	95.6
CBraMod	FM	Local	85.4	81.5	86.7
CBraMod	FM	Centralized	97.0	96.7	97.6
CBraMod	FM	FedAvg	97.0	96.9	97.4
CBraMod	FM	FedPer	94.3	89.1	91.9
EEGMamba	FM	Local	21.3	23.3	22.8
EEGMamba	FM	Centralized	90.6	88.5	89.6
EEGMamba	FM	FedAvg	63.9	63.7	69.1
EEGMamba	FM	FedPer	42.8	37.6	47.2
EEGPT	FM	Local	73.7	71.5	65.2
EEGPT	FM	Centralized	98.0	98.1	97.8
EEGPT	FM	FedAvg	97.0	94.3	89.3
EEGPT	FM	FedPer	86.7	82.4	67.0
LaBraM	FM	Local	85.0	84.3	83.9
LaBraM	FM	Centralized	95.6	95.9	94.1
LaBraM	FM	FedAvg	95.9	95.2	95.6
LaBraM	FM	FedPer	90.9	91.5	92.4
Neuro-GPT	FM	Local	44.3	41.1	27.2
Neuro-GPT	FM	Centralized	91.9	93.5	92.2
Neuro-GPT	FM	FedAvg	92.0	92.2	94.1
Neuro-GPT	FM	FedPer	59.1	64.8	49.6

Table 9: **BNCI2014_009 (P300), per-seed AUPRC** (%). The traditional baseline (xDAWN+LDA) is deterministic and run once. Its values are listed in the Seed 42 column only for layout alignment and do not involve a random seed.

Model	Family	Method	Seed 42	Seed 43	Seed 44
xDAWN+LDA	ML	Local	24.0	–	–
xDAWN+LDA	ML	Centralized	57.0	–	–
EEGNet	DL	Local	69.3	64.5	67.0
EEGNet	DL	Centralized	63.0	58.6	59.4
EEGNet	DL	FedAvg	74.6	75.1	74.4
EEGNet	DL	FedPer	78.8	82.0	79.8
LMDA	DL	Local	74.6	74.1	73.6
LMDA	DL	Centralized	63.8	59.5	60.8
LMDA	DL	FedAvg	72.9	73.8	72.9
LMDA	DL	FedPer	79.5	77.7	77.2
ShallowConvNet	DL	Local	71.1	69.3	70.8
ShallowConvNet	DL	Centralized	68.0	68.7	67.3
ShallowConvNet	DL	FedAvg	70.9	71.4	71.9
ShallowConvNet	DL	FedPer	71.1	71.2	71.6
EEG-Conformer	DL	Local	72.0	69.0	68.7
EEG-Conformer	DL	Centralized	65.6	68.4	69.3
EEG-Conformer	DL	FedAvg	72.6	69.9	70.2
EEG-Conformer	DL	FedPer	73.1	69.3	70.6
Deformer	DL	Local	72.0	71.3	72.6
Deformer	DL	Centralized	67.0	64.7	66.1
Deformer	DL	FedAvg	69.9	69.0	69.3
Deformer	DL	FedPer	78.1	78.4	77.7
CBraMod	FM	Local	65.7	65.5	64.4
CBraMod	FM	Centralized	65.0	62.4	63.9
CBraMod	FM	FedAvg	65.0	64.2	68.8
CBraMod	FM	FedPer	76.4	76.5	78.4
EEGMamba	FM	Local	42.9	42.1	40.2
EEGMamba	FM	Centralized	45.5	46.7	50.9
EEGMamba	FM	FedAvg	41.0	45.1	43.6
EEGMamba	FM	FedPer	41.5	42.9	41.9
EEGPT	FM	Local	55.7	44.1	49.1
EEGPT	FM	Centralized	61.4	59.5	57.8
EEGPT	FM	FedAvg	63.6	56.4	63.4
EEGPT	FM	FedPer	63.1	54.8	53.9
LaBraM	FM	Local	66.4	65.4	64.4
LaBraM	FM	Centralized	61.9	62.1	65.2
LaBraM	FM	FedAvg	68.0	69.2	65.7
LaBraM	FM	FedPer	74.7	78.0	73.3
Neuro-GPT	FM	Local	74.9	72.3	75.2
Neuro-GPT	FM	Centralized	73.6	72.5	70.9
Neuro-GPT	FM	FedAvg	71.8	69.4	72.4
Neuro-GPT	FM	FedPer	75.2	75.9	75.8

Table 10: **Full main results**, cross-seed mean ± standard deviation (%) for the 13 models of Table 2, over the same dataset × protocol cells and the same four protocols. "–" marks federated cells where traditional machine learning does not apply, the ML row is a deterministic single run, and bold marks the highest full-precision mean per dataset and protocol, as in Table 2. In the rank column, models are ranked within each cell on full-precision means, exactly tied means share their average rank, and the per-cell ranks are then averaged over the 12 cells; the traditional baselines are excluded from the ranking.

		BNCI2014_001 (MI)				**Nakanishi2015 (SSVEP)**				**BNCI2014_009 (P300)**			
		BA (%)				**Accuracy (%)**				**AUPRC (%)**			
Model	**Rank**	**Local**	**Cent.**	**FedAvg**	**FedPer**	**Local**	**Cent.**	**FedAvg**	**FedPer**	**Local**	**Cent.**	**FedAvg**	**FedPer**
ML													
Paradigm-specific	–	**62.5**	43.5	–	–	**99.1**	98.1	–	–	24.0	57.0	–	–
DL													
EEG-Conformer	4.00	53.1 ± 1.4	42.7 ± 3.0	62.9 ± 0.5	59.5 ± 2.3	68.6 ± 1.9	95.2 ± 0.5	96.8 ± 0.7	92.3 ± 1.7	69.9 ± 1.8	67.7 ± 1.9	70.9 ± 1.5	71.0 ± 1.9
Deformer	4.17	46.3 ± 1.6	40.4 ± 1.7	41.1 ± 0.9	52.5 ± 1.2	**99.1 ± 0.5**	**98.4 ± 1.2**	96.4 ± 4.3	**94.9 ± 0.6**	72.0 ± 0.7	66.0 ± 1.2	69.4 ± 0.5	78.1 ± 0.4
EEGNet	4.58	49.8 ± 1.7	40.1 ± 1.5	54.2 ± 1.3	57.3 ± 2.3	77.1 ± 5.2	94.9 ± 2.5	97.0 ± 0.6	85.1 ± 1.9	66.9 ± 2.4	60.3 ± 2.4	**74.7 ± 0.3**	**80.2 ± 1.6**
LMDA	4.58	51.2 ± 1.5	47.8 ± 0.9	57.2 ± 0.6	**59.8 ± 1.4**	62.8 ± 5.0	91.1 ± 1.4	88.9 ± 0.9	70.6 ± 5.5	74.1 ± 0.5	61.4 ± 2.2	73.2 ± 0.5	78.1 ± 1.2
ShallowConvNet	5.42	60.4 ± 1.4	50.9 ± 1.1	57.1 ± 2.0	56.4 ± 1.7	50.8 ± 1.7	69.0 ± 0.3	71.4 ± 1.6	48.0 ± 0.8	70.4 ± 1.0	68.0 ± 0.7	71.4 ± 0.5	71.3 ± 0.3
FM													
CBraMod	3.83	60.4 ± 0.8	**55.9 ± 2.0**	52.2 ± 0.7	58.5 ± 1.4	84.5 ± 2.7	97.1 ± 0.5	**97.1 ± 0.3**	91.7 ± 2.6	65.2 ± 0.7	63.7 ± 1.3	66.0 ± 2.5	77.1 ± 1.1
Neuro-GPT	4.92	43.8 ± 1.1	55.4 ± 2.5	**67.4 ± 1.4**	50.5 ± 2.4	37.5 ± 9.1	92.5 ± 0.9	92.8 ± 1.1	57.8 ± 7.7	**74.1 ± 1.6**	**72.3 ± 1.4**	71.2 ± 1.6	75.6 ± 0.4
LaBraM	6.08	35.0 ± 1.7	42.4 ± 2.7	47.7 ± 1.5	43.2 ± 4.0	84.4 ± 0.6	95.2 ± 1.0	95.6 ± 0.4	91.6 ± 0.7	65.4 ± 1.0	63.1 ± 1.9	67.6 ± 1.7	75.3 ± 2.4
EEGPT	7.75	31.2 ± 1.2	37.7 ± 3.5	33.6 ± 3.7	31.9 ± 5.2	70.1 ± 4.4	98.0 ± 0.2	93.5 ± 3.9	78.7 ± 10.3	49.7 ± 5.8	59.6 ± 1.8	61.1 ± 4.1	57.3 ± 5.1
EEGMamba	9.67	36.1 ± 2.6	35.0 ± 2.0	31.7 ± 1.3	34.4 ± 2.1	22.5 ± 1.1	89.6 ± 1.0	65.6 ± 3.0	42.5 ± 4.8	41.7 ± 1.4	47.7 ± 2.9	43.2 ± 2.1	42.1 ± 0.7

B.3 Secondary Metrics

To check that the primary-metric conclusions are robust to the choice of metric, Tables 11, 12, and 13 report, for each dataset, secondary metrics complementary to its primary metric. For BNCI2014_001, whose primary metric BA already corrects for class imbalance, the additional metric is macro-F1 (the BA column equals the primary metric and is kept for completeness). For the balanced Nakanishi2015, whose primary metric is plain accuracy, BA and macro-F1 serve as threshold-based cross-checks under a different convention. For the imbalanced two-class BNCI2014_009, BA and macro-F1 are both robust to imbalance, and ROC-AUC is a ranking-based metric from the same family as the primary AUPRC. Taking BNCI2014_009 as an example, the `FedPer` advantage of EEGNet under AUPRC is echoed by the secondary metrics (BA 79.4 versus `Local` 74.7, macro-F1 82.2 versus 75.6). The protocol-ordering patterns of all three datasets largely follow the primary metrics under the secondary metrics, with individual exceptions where two metrics trade off against each other: on BNCI2014_009, LMDA's `FedPer` balanced accuracy is slightly below its `Local` value (77.2 against 77.8) while its AUPRC rises (78.1 against 74.1), and EEG-Conformer shows the same pattern. These are metric-level trade-offs within a model, not reversals of the protocol ordering.

Table 11: **BNCI2014_001 secondary metrics** (balanced accuracy / macro-F1, %).

Model	Method	Balanced Acc.	macro-F1
EEGNet	Local	49.8	49.2
EEGNet	Centralized	40.1	39.6
EEGNet	FedAvg	54.2	52.9
EEGNet	FedPer	57.3	56.6
LMDA	Local	51.2	50.7
LMDA	Centralized	47.8	47.1
LMDA	FedAvg	57.2	56.4
LMDA	FedPer	59.8	59.1
ShallowConvNet	Local	60.4	59.7
ShallowConvNet	Centralized	50.9	50.1
ShallowConvNet	FedAvg	57.1	56.4
ShallowConvNet	FedPer	56.4	55.5
EEG-Conformer	Local	53.1	52.4
EEG-Conformer	Centralized	42.7	42.3
EEG-Conformer	FedAvg	62.9	61.8
EEG-Conformer	FedPer	59.5	58.5
Deformer	Local	46.3	45.4
Deformer	Centralized	40.4	39.9
Deformer	FedAvg	41.1	36.9
Deformer	FedPer	52.5	52.0
CBraMod	Local	60.4	60.1
CBraMod	Centralized	55.9	55.3
CBraMod	FedAvg	52.2	51.5
CBraMod	FedPer	58.5	58.1
EEGMamba	Local	36.1	34.2
EEGMamba	Centralized	35.0	33.9
EEGMamba	FedAvg	31.7	30.1
EEGMamba	FedPer	34.4	33.1
EEGPT	Local	31.2	30.3
EEGPT	Centralized	37.7	36.7
EEGPT	FedAvg	33.6	32.5
EEGPT	FedPer	31.9	31.4
LaBraM	Local	35.0	34.4
LaBraM	Centralized	42.4	41.9
LaBraM	FedAvg	47.7	46.5
LaBraM	FedPer	43.2	42.7
Neuro-GPT	Local	43.8	43.4
Neuro-GPT	Centralized	55.4	54.8
Neuro-GPT	FedAvg	67.4	66.8
Neuro-GPT	FedPer	50.5	49.9

Table 12: **Nakanishi2015 secondary metrics** (balanced accuracy / macro-F1, %).

Model	Method	Balanced Acc.	macro-F1
EEGNet	Local	77.1	76.7
EEGNet	Centralized	94.9	94.8
EEGNet	FedAvg	97.0	96.9
EEGNet	FedPer	85.1	84.7
LMDA	Local	62.8	62.4
LMDA	Centralized	91.1	90.9
LMDA	FedAvg	88.9	88.7
LMDA	FedPer	70.6	70.1
ShallowConvNet	Local	50.8	49.2
ShallowConvNet	Centralized	69.0	68.2
ShallowConvNet	FedAvg	71.4	71.1
ShallowConvNet	FedPer	48.0	46.7
EEG-Conformer	Local	68.6	67.6
EEG-Conformer	Centralized	95.2	95.2
EEG-Conformer	FedAvg	96.8	96.8
EEG-Conformer	FedPer	92.3	92.0
Deformer	Local	99.1	99.0
Deformer	Centralized	98.4	98.3
Deformer	FedAvg	96.4	96.1
Deformer	FedPer	94.9	94.8
CBraMod	Local	84.5	83.9
CBraMod	Centralized	97.1	97.0
CBraMod	FedAvg	97.1	97.0
CBraMod	FedPer	91.7	91.5
EEGMamba	Local	22.5	21.2
EEGMamba	Centralized	89.6	89.4
EEGMamba	FedAvg	65.6	64.2
EEGMamba	FedPer	42.5	40.9
EEGPT	Local	70.1	69.3
EEGPT	Centralized	98.0	97.9
EEGPT	FedAvg	93.5	93.3
EEGPT	FedPer	78.7	78.3
LaBraM	Local	84.4	84.1
LaBraM	Centralized	95.2	95.1
LaBraM	FedAvg	95.6	95.5
LaBraM	FedPer	91.6	91.3
Neuro-GPT	Local	37.5	36.7
Neuro-GPT	Centralized	92.5	92.3
Neuro-GPT	FedAvg	92.8	92.6
Neuro-GPT	FedPer	57.8	56.6

Table 13: **BNCI2014_009 secondary metrics** (balanced accuracy / ROC-AUC / macro-F1, %).

Model	Method	Balanced Acc.	ROC-AUC	macro-F1
EEGNet	Local	74.7	87.0	75.6
EEGNet	Centralized	72.9	85.1	73.1
EEGNet	FedAvg	75.8	90.7	78.1
EEGNet	FedPer	79.4	92.5	82.2
LMDA	Local	77.8	90.3	79.3
LMDA	Centralized	71.9	84.8	72.5
LMDA	FedAvg	72.9	90.1	76.0
LMDA	FedPer	77.2	92.1	80.6
ShallowConvNet	Local	72.4	87.7	76.2
ShallowConvNet	Centralized	73.9	88.2	75.3
ShallowConvNet	FedAvg	70.8	89.9	74.3
ShallowConvNet	FedPer	72.7	89.3	76.2
EEG-Conformer	Local	77.0	88.5	77.8
EEG-Conformer	Centralized	74.3	87.5	76.6
EEG-Conformer	FedAvg	76.0	88.8	77.3
EEG-Conformer	FedPer	75.7	88.9	77.7
Deformer	Local	74.9	89.5	77.3
Deformer	Centralized	71.7	85.9	74.0
Deformer	FedAvg	74.3	87.8	76.0
Deformer	FedPer	79.6	92.1	81.5
CBraMod	Local	74.9	85.9	76.3
CBraMod	Centralized	73.9	86.4	74.7
CBraMod	FedAvg	74.0	86.0	75.1
CBraMod	FedPer	79.8	91.3	81.6
EEGMamba	Local	59.8	71.8	61.3
EEGMamba	Centralized	64.0	77.4	65.9
EEGMamba	FedAvg	61.3	73.9	62.4
EEGMamba	FedPer	61.7	73.7	63.0
EEGPT	Local	64.1	78.2	65.9
EEGPT	Centralized	70.5	84.2	71.4
EEGPT	FedAvg	68.9	84.5	71.0
EEGPT	FedPer	66.6	82.4	69.2
LaBraM	Local	72.4	87.8	74.7
LaBraM	Centralized	70.5	86.2	72.7
LaBraM	FedAvg	74.1	88.1	76.2
LaBraM	FedPer	79.3	91.1	81.3
Neuro-GPT	Local	78.7	90.4	79.6
Neuro-GPT	Centralized	76.4	90.2	77.9
Neuro-GPT	FedAvg	79.2	89.7	77.9
Neuro-GPT	FedPer	80.8	91.8	81.8

B.4 PER-MODEL $\Delta(\texttt{FedPer} - \texttt{FedAvg})$

Table 14 gives the exact values behind Figure 4(a), in pp. Each cell is the mean difference across clients and then across seeds. The largest absolute change per column is bolded.

Table 14: **Per-model personalization effect** $\Delta(\texttt{FedPer} - \texttt{FedAvg})$ (pp). †: difference magnitude ≤ 0.2 pp, treated as a practical tie.

Model	BNCI2014_001	Nakanishi2015	BNCI2014_009
Deformer	$+11.4$	-1.5	$+8.7$
CBraMod	$+6.3$	-5.4	$\mathbf{+11.1}$
EEGNet	$+3.1$	-11.9	$+5.5$
EEGMamba	$+2.7$	-23.0	-1.1
LMDA	$+2.6$	-18.3	$+4.9$
ShallowConvNet	-0.7	-23.4	$-0.1^{\dagger}$
EEGPT	-1.8	-14.8	-3.9
EEG-Conformer	-3.4	-4.5	$+0.1^{\dagger}$
LaBraM	-4.5	-4.0	$+7.7$
Neuro-GPT	$\mathbf{-16.9}$	$\mathbf{-34.9}$	$+4.5$

C DETAILS OF STATISTICAL TESTS

C.1 CONVENTIONS OF THE TWO-LEVEL PAIRED STATISTICS

This subsection records the precise conventions summarized in Section 2.4 of the main text. *Model-level descriptive statistics*: for each neural model we compute the protocol differences $\Delta(\cdot - \texttt{Local})$ and $\Delta(\texttt{FedPer} - \texttt{FedAvg})$ as mean differences first across clients and then across seeds, and summarize directional consistency by the proportion of benefiting models ($\Delta > 0$). *Subject-level inferential statistics*: each subject's paired difference is first averaged across the ten models and three seeds, and a two-sided Wilcoxon signed-rank test is applied to the 9–10 independent subjects, the only independent sampling unit in our design. *Worst-client statistic*: per-client scores are first averaged across seeds, the worst client of each model is taken, and the resulting values are summarized across models as mean $\pm$ standard deviation. *Tercile analysis*: on per-client data we test the association between method differences and client baseline strength. Clients are sorted by their baseline score and split into weakest, medium, and strongest terciles, and the mean method difference is compared across groups. The baseline definitions, split ratios, and group membership are given in Appendix C.4. All statistics compare columns pairwise within exactly the same model pool and never merge values across datasets.

Multiple comparisons. The paper reports twelve protocol contrasts defined by the design, four per dataset (three against `Local` plus `FedPer` vs. `FedAvg`). We report uncorrected subject-level p values as the primary analysis and add a Holm–Bonferroni sensitivity over the full twelve-contrast family. Under that correction, SSVEP `Centralized`/`FedAvg`/`FedPer` vs. `Local` and MI `FedPer` vs. `Local` (all $p = 0.004$) move to 0.044, and BNCI2014_009 `FedPer` vs. `Local` ($p = 0.002$) to 0.024, so the SSVEP and `FedPer` conclusions are unchanged; MI `FedAvg` vs. `Local` moves from 0.020 to 0.120 (to 0.060 under a within-MI correction of four contrasts), so the MI `FedAvg` improvement is suggestive rather than established, as stated in Section 3.1. BNCI2014_009 `FedPer` vs. `FedAvg` ($p = 0.084$) was not significant before correction either. Model-pool tests remain references only.

C.2 WILCOXON SIGNED-RANK TESTS

The inferential statistics of Table 15 use subjects as the unit (client-pooled): each subject's paired difference is first averaged across the ten models and three seeds, and a two-sided Wilcoxon signed-rank test is applied to the 9–10 independent subjects, the only independent sampling unit in our design.

Table 15: **Two-level paired statistics of protocol effects**. Δ mean: equal-weight average of the ten models' differences versus their own `Local`. Benefiting: proportion with $\Delta > 0$. Subject-level p: two-sided Wilcoxon signed-rank test on the 9–10 independent subjects.

Dataset	**Protocol**	**Δ (pp)**	**Benef. (/10)**	p
BNCI2014_001 (MI)	Centralized	-1.9	3/10	0.250
	FedAvg	$+3.8$	6/10	0.020
	FedPer	$+3.7$	7/10	0.004
Nakanishi2015 (SSVEP)	Centralized	$+26.4$	9/10	0.004
	FedAvg	$+23.8$	9/10	0.004
	FedPer	$+9.6$	8/10	0.004
BNCI2014_009 (P300)	Centralized	-2.0	2/10	0.084
	FedAvg	$+1.9$	7/10	0.770
	FedPer	$+5.7$	10/10	0.002

As a supplement, Tables 16, 17, and 18 report the same comparisons under a *model-pool* unit, in which the 30 *model*$\times$*seed* combinations are treated as samples. Its sample size is larger, but the units are not independent, since the three seeds of the same model are correlated, so it serves only as a reference. Δ is the mean paired difference (pp).

Table 16: **BNCI2014_001 (MI)**.

Comparison	**Unit**	n	**Δ mean**	p
Centralized vs. Local	client-pooled	9	-1.9	0.250
Centralized vs. Local	model-pool	30	-1.9	0.205
FedAvg vs. Local	client-pooled	9	$+3.8$	0.020
FedAvg vs. Local	model-pool	30	$+3.8$	0.080
FedPer vs. Local	client-pooled	9	$+3.7$	0.004
FedPer vs. Local	model-pool	30	$+3.7$	0.0005
FedPer vs. FedAvg	client-pooled	9	-0.1	0.910
FedPer vs. FedAvg	model-pool	30	-0.1	0.655

Table 17: **Nakanishi2015 (SSVEP)**.

Comparison	**Unit**	n	**Δ mean**	p
Centralized vs. Local	client-pooled	9	$+26.4$	0.004
Centralized vs. Local	model-pool	30	$+26.4$	< 0.0001
FedAvg vs. Local	client-pooled	9	$+23.8$	0.004
FedAvg vs. Local	model-pool	30	$+23.8$	< 0.0001
FedPer vs. Local	client-pooled	9	$+9.6$	0.004
FedPer vs. Local	model-pool	30	$+9.6$	< 0.0001
FedPer vs. FedAvg	client-pooled	9	-14.2	0.004
FedPer vs. FedAvg	model-pool	30	-14.2	< 0.0001

Table 18: **BNCI2014_009 (P300)**.

Comparison	Unit	n	Δ mean	p
Centralized vs. Local	client-pooled	10	-2.0	0.084
Centralized vs. Local	model-pool	30	-2.0	0.047
FedAvg vs. Local	client-pooled	10	$+1.9$	0.770
FedAvg vs. Local	model-pool	30	$+1.9$	0.096
FedPer vs. Local	client-pooled	10	$+5.7$	0.002
FedPer vs. Local	model-pool	30	$+5.7$	< 0.0001
FedPer vs. FedAvg	client-pooled	10	$+3.7$	0.084
FedPer vs. FedAvg	model-pool	30	$+3.7$	0.0006

C.3 Client-Level Association: Remaining Datasets

Figure 4(b) shows only the client scatter of Nakanishi2015. The corresponding scatters for BNCI2014_001 and BNCI2014_009 are shown in Figure 7: the association is not significant on MI ($r = -0.23$), and there is no systematic association on P300 ($r = +0.04$). Per-client values can be reconstructed from the per-client final-round files of the repository. Because the horizontal axis and the plotted difference share the same baseline term, this association carries a built-in *negative* bias ($\mathrm{Cov}(B, T - B) = \mathrm{Cov}(B, T) - \mathrm{Var}(B)$), so the positive $r = +0.94$ is not an artifact of that coupling; it remains a descriptive association over nine clients rather than a mechanistic estimate.

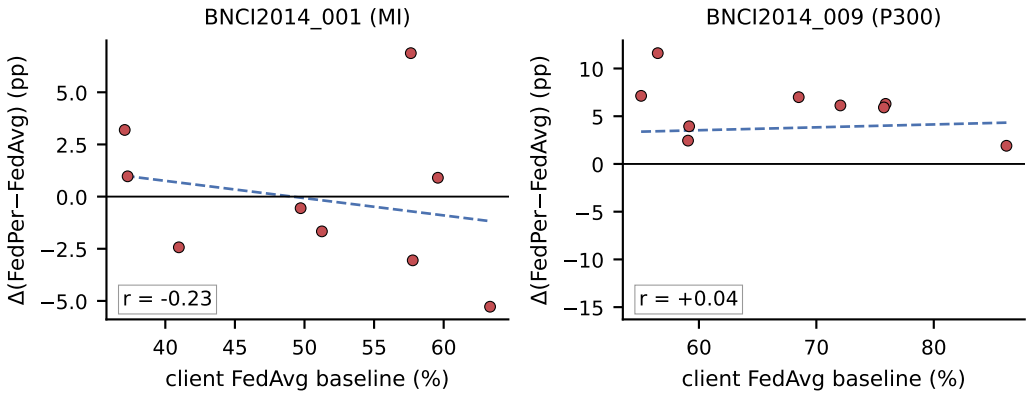

Figure 7: **Client-level association**. Per-client $\Delta(\texttt{FedPer} - \texttt{FedAvg})$ against the client's `FedAvg` baseline on BNCI2014_001 (left) and BNCI2014_009 (right). Unlike Nakanishi2015 (Figure 4(b), $r = +0.94$), no systematic tilt of private-head gains or losses with client-baseline strength is observed on these two datasets.

C.4 Tercile Group Composition

The tercile statistics in Sections 3.1 and 3.3 are computed under the following convention, with grouping performed within each dataset only. A subject's baseline score is its fixed-final-round primary metric under independent training (`Local`), averaged over the ten neural models and three seeds. For the `FedPer−FedAvg` difference (Section 3.3), the baseline and ranking instead use the corresponding `FedAvg` scores. Subjects are sorted by baseline score and split into weakest, medium, and strongest groups: datasets with 9 subjects are split 3/3/3, and BNCI2014_009 with 10 subjects is split 3/3/4. Each subject's method difference is likewise averaged over models and seeds, and a group mean is the average of its members' differences. Group membership (subjects numbered 0–9 in data-file order) and the per-group mean differences (pp) are as follows. `FedAvg−Local`: BNCI2014_001 $+2.9/+6.2/+2.2$ (weakest $\{4, 1, 5\}$, medium $\{6, 3, 7\}$, strongest $\{0, 2, 8\}$); Nakanishi2015 $+34.1/+23.4/+13.8$ ($\{1, 0, 2\}/\{6, 4, 3\}/\{8, 7, 5\}$); BNCI2014_009 $+8.3/+0.02/-1.4$ ($\{2, 5, 6\}/\{0, 7, 1\}/\{3, 4, 9, 8\}$).

`FedPer−FedAvg`: BNCI2014_001 $+0.6/+1.6/-2.5$ $(\{1,4,5\}/\{3,6,8\}/\{7,2,0\})$; Nakanishi2015 $-20.7/-12.5/-9.3$ $(\{1,0,2\}/\{4,3,8\}/\{6,5,7\})$; BNCI2014_009 $+7.1/-1.3/+5.1$ $(\{5,0,7\}/\{6,2,3\}/\{1,9,4,8\})$. The raw data are the per-client final-round primary-metric files of each dataset, and all values above can be reproduced by the analysis scripts in the repository.

D Training Dynamics

D.1 Round-Wise Test-Metric Convergence

Figure 8 traces how the cross-client mean primary metric of each protocol evolves over the 200 training rounds.

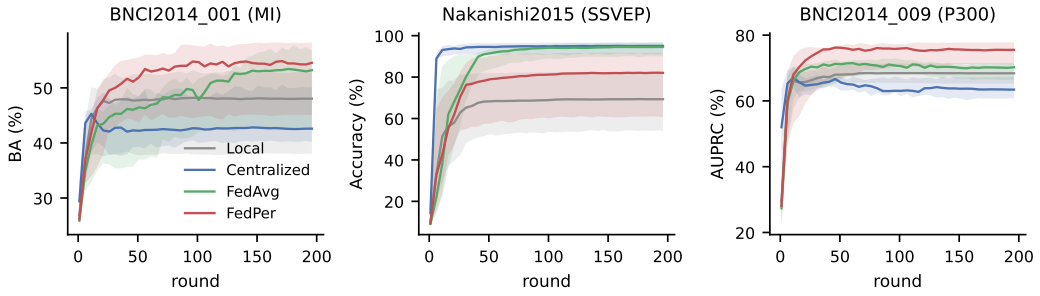

Figure 8: **Round-wise training dynamics**. For each protocol, the cross-client mean primary metric is first computed per model and round, giving ten model-level curves. The solid line is their median across the ten neural models and the shaded band their interquartile range (sampled every 5 rounds), so no single model or client is plotted separately. All three datasets reach a plateau within the 200-round budget. The federated and `Centralized` curves of Nakanishi2015 are nearly saturated after 50 rounds. Some runs on BNCI2014_009 show a late mild decline (consistent with the best-round degradation statistics of Section 4.1), and between-model dispersion narrows visibly under federated protocols.

D.2 Round-Wise Training Loss

Figure 9 shows the corresponding round-wise training loss.

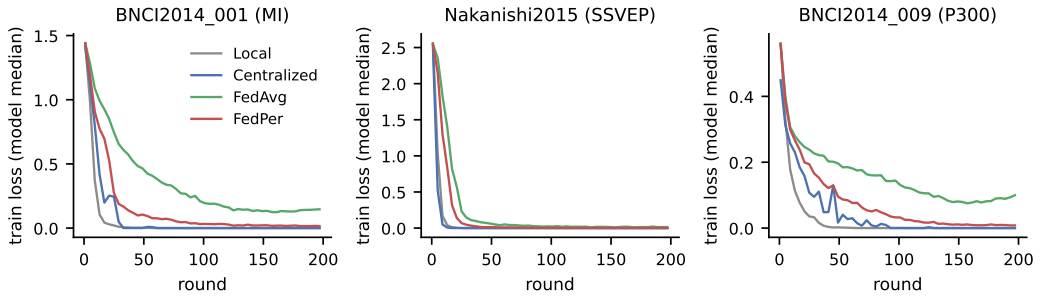

Figure 9: **Round-wise training loss** (for each model, the mean loss over clients per round, then the median across the ten neural models, seed 42). The curves of `Local` and `Centralized` fall close to zero. `FedAvg` plateaus at a clearly higher level, whereas `FedPer` settles just above the `Local` and `Centralized` floor.

The curves of Figure 9 separate for a known reason. `Local` and `Centralized` optimize directly on their training data, so their loss keeps falling as the models memorize the training set. Under

FedAvg the global model is a compromise for every local distribution, so after each round of aggregation the local loss cannot decrease further. FedPer aggregates only the representation and leaves the classification head on the client, so part of the model still fits each user and its loss settles between the two extremes. This loss separation is consistent with client drift (Karimireddy et al., 2020) — we do not measure update directions, update norms, or pre-/post-aggregation losses directly — and it offers a mechanistic explanation for why collaboration helps unevenly, the finding of Section 3.1.

D.3 Best-Round Degradation Statistics

The mean difference between the best round and the fixed final round is 2.21 pp on BNCI2014_001, 0.59 pp on Nakanishi2015, and 3.58 pp on BNCI2014_009. On BNCI2014_009, 11 of the 40 (*model*×*protocol*) combinations drop by more than 5 pp, the largest drop being 13.78 pp. Because the best round is selected from test curves, this statistic serves only as a training-stability diagnostic (see the protocol discussion in Section 4.1).

E Cross-Client Representation Similarity

Method. For each neural-model run, we take the representation vectors that feed the classifier at the final round, one set per client, and compute the pairwise linear CKA (Kornblith et al., 2019) between all clients of the same model and dataset, then average. Each client's feature matrix comes from that client's own held-out test trials, evaluated on the final-round model, i.e. exactly the forward pass that produces the per-client scores of Table 2; the trials of two different clients are *not* paired. For unpaired feature sets the quantity we compute is, identically, the cosine similarity between the two feature-space covariance operators, $\|X_c Y_c^\top\|_F^2 / \sqrt{\|X_c X_c^\top\|_F^2 \, \|Y_c Y_c^\top\|_F^2} = \langle S_X, S_Y \rangle_F / (\|S_X\|_F \|S_Y\|_F)$ with $S_X = X_c^\top X_c$, so it measures how similar the two clients' second-order feature structure is rather than trial-level alignment, and it is invariant to the ordering of rows. It is not the Gram/HSIC expression given by Kornblith et al. (2019), which presumes the same n paired samples on both sides, and the two forms are not equal in general; all values reported here are those of the covariance form. Two consequences follow. First, since the shared representation is re-broadcast before evaluation, all clients run identical encoder parameters under FedAvg, and an identical shared representation under FedPer; a FedAvg CKA below 1 therefore reflects differences between clients' input distributions and their client-local BatchNorm running statistics (buffers are never aggregated), not different encoders. Second, a row-permutation control is not meaningful here, because no cross-client pairing exists to permute. CKA is still only compared **across protocols within the same model**: the restriction is that coarse agreement cannot be separated from architecture-level differences, not that feature dimensions differ. The analysis uses seed 42. Figure 10 plots the per-model curves.

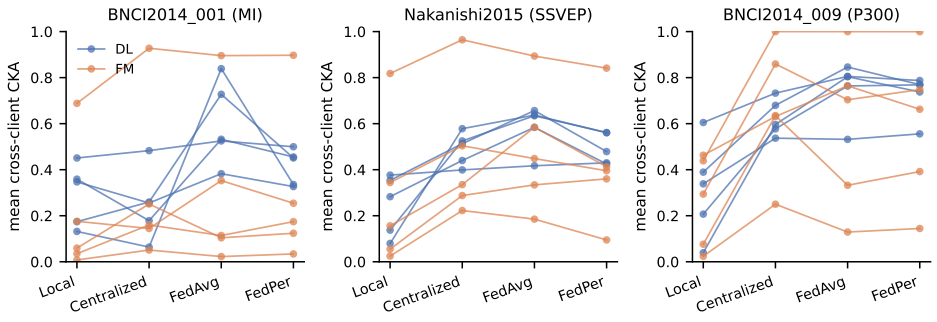

Figure 10: **Mean cross-client CKA** of each model under the four protocols (one line per model, blue for DL and orange for FM). FedAvg exceeds Local in all 30 *model*×*dataset* combinations.

Full values. Table 19 lists the underlying values for every model, protocol, and dataset.

Table 19: **Mean cross-client linear CKA** per model, protocol, and dataset.

	BNCI2014_001				Nakanishi2015				BNCI2014_009			
Model	**Local**	**Cent.**	**FedAvg**	**FedPer**	**Local**	**Cent.**	**FedAvg**	**FedPer**	**Local**	**Cent.**	**FedAvg**	**FedPer**
EEGNet	0.36	0.18	0.53	0.46	0.35	0.51	0.66	0.48	0.39	0.68	0.85	0.77
LMDA	0.35	0.26	0.38	0.33	0.28	0.44	0.58	0.42	0.34	0.54	0.53	0.56
ShallowConvNet	0.45	0.48	0.53	0.50	0.38	0.40	0.42	0.43	0.60	0.73	0.81	0.79
EEG-Conformer	0.17	0.26	0.73	0.45	0.08	0.58	0.64	0.56	0.21	0.58	0.76	0.77
Deformer	0.13	0.06	0.84	0.34	0.14	0.52	0.63	0.56	0.04	0.60	0.80	0.74
CBraMod	0.04	0.16	0.11	0.17	0.05	0.29	0.33	0.36	0.29	0.86	0.70	0.75
EEGMamba	0.69	0.93	0.90	0.90	0.82	0.96	0.89	0.84	0.44	1.00	1.00	1.00
EEGPT	0.01	0.05	0.02	0.03	0.02	0.22	0.19	0.09	0.02	0.25	0.13	0.14
LaBraM	0.06	0.25	0.10	0.12	0.16	0.34	0.58	0.41	0.08	0.63	0.33	0.39
Neuro-GPT	0.18	0.14	0.35	0.25	0.34	0.50	0.45	0.40	0.46	0.63	0.76	0.66

Four observations. First, **federated aggregation aligns representations**: in all 30 (*model*×*dataset*) combinations, the cross-client CKA under `FedAvg` exceeds that under `Local`. The median rises from 0.17/0.22/0.32 under `Local` (BNCI2014_001/Nakanishi2015/BNCI2014_009) to 0.45/0.58/0.76 under `FedAvg`. Second, **alignment strength follows a task gradient**: MI $<$ SSVEP $<$ P300. This gradient does not by itself track where collaboration pays off, since P300, the most alignable regime, has the smallest `FedAvg` gain. Section 3.3 reads that inversion through a second, unmeasured factor. Third, **on P300, "representations are alignable while boundaries need privacy"**: representations on this dataset are highly alignable under federation, and the best protocol is precisely `FedPer` with a shared representation and private classification heads, which is representation-level support for the RQ3 conclusion. Fourth, **private heads re-diversify representations**: the alignment under `FedPer` is below that under `FedAvg` in most combinations (e.g., EEG-Conformer 0.73→0.45 and Deformer 0.84→0.34 on BNCI2014_001), because local training of the classification head pulls the shared representation toward each client through backpropagation, so personalization is not without representation-level cost. Two degenerate cases require caution: the CKA of EEGMamba is close to 1 under many protocols (suspected representation collapse, consistent with its low performance on BNCI2014_001 and Nakanishi2015), and the CKA of EEGPT stays near 0 throughout (sparse features in 32,768 dimensions). These two rows are excluded from pattern interpretation, and dropping them leaves the six `Local` and `FedAvg` medians unchanged.

A second factor we infer rather than measure. The account in Section 3.3 treats representation alignment as one factor and the cross-subject consistency of the decision boundary as another. Only the first is measured here. The second cannot be read off this benchmark directly, because each client is evaluated on its own temporally held-out trials and the participating subjects share no common evaluation set on which their decision behaviour could be compared. We instead read the `Centralized`–`FedPer` contrast of Table 3 as circumstantial evidence, since `Centralized` pools the classification head while `FedPer` keeps it private. That contrast is not boundary-only, however: `Centralized` also optimizes on pooled data and does not accumulate client drift, so the two protocols differ in optimization as well as in where the boundary is placed. The contrast is therefore consistent with that reading rather than a test of it.

F Systems and Reproducibility

F.1 Communication Validation and Per-Dataset Parameter Counts

The theoretical communication formula (Section 2.4) was validated against per-run measurements: the measured total communication of all 180 federated runs of the paper's ten neural models matches "shared-parameter bytes $\times$ ($2\times$ rounds $+1$) $\times$ number of clients" byte for byte (relative deviation $< 10^{-9}$). Here "measured" means an instrumented count of the logical parameter payload actually exchanged (FP32 named shared parameters); it excludes client-local buffers such as BatchNorm running statistics, tensor serialization overhead, and transport-protocol headers, so it is a logical payload accounting rather than a measurement of network traffic. Parameter counts vary with the channel and class counts of each dataset. The actual values per dataset are listed in Table 20 (`FedAvg`, full-parameter convention).

Table 20: **Parameter counts per dataset** (`FedAvg`, full-parameter convention).

Model	BNCI2014_001	Nakanishi2015	BNCI2014_009
EEGNet	3,444	7,388	1,554
LMDA	4,096	6,516	3,320
ShallowConvNet	46,124	44,212	27,322
EEG-Conformer	789,572	767,436	779,906
Deformer	809,172	769,756	780,114
CBraMod	4,954,204	4,960,612	4,890,202
EEGMamba	3,318,247	3,319,855	3,317,845
EEGPT	25,418,750	25,680,468	25,352,978
LaBraM	5,895,940	5,903,948	5,850,738
Neuro-GPT	166,590	185,810	161,576

Table 21 gives the theoretical total communication of complete training (GiB), computed from each dataset's actual parameter counts and client counts.

Table 21: **Theoretical total communication** of complete training (GiB).

Model	Method	BNCI2014_001	Nakanishi2015	BNCI2014_009
EEGNet	FedAvg	0.05	0.10	0.02
EEGNet	FedPer	0.02	0.02	0.02
LMDA	FedAvg	0.06	0.09	0.05
LMDA	FedPer	0.04	0.03	0.04
ShallowConvNet	FedAvg	0.62	0.59	0.41
ShallowConvNet	FedPer	0.49	0.19	0.40
EEG-Conformer	FedAvg	10.62	10.32	11.65
EEG-Conformer	FedPer	2.10	1.80	2.19
Deformer	FedAvg	10.88	10.35	11.65
Deformer	FedPer	10.76	9.99	11.59
CBraMod	FedAvg	66.61	66.69	73.05
CBraMod	FedPer	65.66	65.66	72.96
EEGMamba	FedAvg	44.61	44.63	49.56
EEGMamba	FedPer	44.60	44.60	49.56
EEGPT	FedAvg	341.74	345.26	378.73
EEGPT	FedPer	339.98	339.98	377.75
LaBraM	FedAvg	79.27	79.38	87.40
LaBraM	FedPer	78.31	78.31	87.01
Neuro-GPT	FedAvg	2.24	2.50	2.41
Neuro-GPT	FedPer	2.11	2.10	2.34

F.2 Wall-Clock Time and Peak Memory

Table 22 reports the peak memory and wall-clock time of each model on the three datasets.

Table 22: **Peak memory and wall-clock time** (MB and s) per dataset.

Model	Peak memory (MB)	BNCI2014_001	Nakanishi2015	BNCI2014_009
EEGNet	616	137	74	208
LMDA	1,539	532	136	834
ShallowConvNet	4,122	171	72	199
EEG-Conformer	9,896	662	649	1,226
Deformer	16,068	604	531	1,854
CBraMod	1,249	2,211	1,272	4,560
EEGMamba	893	2,832	1,576	6,224
EEGPT	6,560	4,053	1,538	6,735
LaBraM	1,491	1,347	1,127	3,879
Neuro-GPT	9,885	1,141	616	2,311

Wall-clock times are the third-smallest of the four per-protocol three-seed medians, obtained on a single RTX 5090 with four concurrent jobs (contending for the same GPU), and serve only as a conditional order-of-magnitude reference. Peak memory is the maximum recorded over all runs of the model. The three cost columns are therefore not a matched measurement triplet: memory is a training-time allocated-GPU high-water mark rather than an inference-time or edge-device footprint, wall-clock is a specific order statistic under contention, and neither supports deployment-feasibility claims on its own.

F.3 Performance–Communication Trade-off (RQ4)

Figure 5 presents the performance–cost landscape of the ten neural models under the two federated protocols. The performance–cost relation is task-dependent, but the three datasets send one consistent message: lightweight models are clearly favored in the performance–cost trade-off. The best combination on BNCI2014_009, EEGNet+`FedPer`, has only 1,554 parameters, 0.02 GiB of theoretical communication, and 0.6 GB of peak memory. Scaling up the model does not yield a higher metric on this task. Combined with the all-model consistency of `FedPer` (Table 15), it offers the strongest metric at the lowest cost and is the default deployment option for this paradigm. On BNCI2014_001 the best combination, Neuro-GPT+`FedAvg`, costs 2.24 GiB of communication, two orders of magnitude below the largest foundation model (EEGPT, 341.74 GiB). On Nakanishi2015, TRCA, Deformer, and several combinations that share the full model are near saturation, leaving little marginal room for more parameters. Deployment choices should weight communication and maintenance costs more heavily.

F.4 Additional System Observations

The system measures reveal two facts invisible to parameter count alone. First, **parameter count is not memory**: EEGMamba has 3.3M parameters yet needs only 0.9 GB of peak memory (the long-sequence efficiency of state-space models), while Neuro-GPT, with one-twentieth of its parameters, needs 9.9 GB for self-attention activations and Deformer reaches 16.1 GB (Table 22). On edge BCI devices, memory is often the harder constraint. Second, **the communication saved by personalization depends on the head's share of total parameters**: EEG-Conformer has a large classification head, and `FedPer` cuts its communication from 11.65 to 2.19 GiB. EEGPT, CBraMod, and LaBraM keep almost all parameters in the shared representation, so the two federated protocols cost nearly the same (Table 21). Hence, the performance gains of personalization (RQ3 in the main text) and its communication savings follow different mechanisms and cannot be inferred from each other. In addition, the cross-seed standard deviations of Table 10 in Appendix B.2 serve as a deployment-stability dimension: EEGPT shows one of the highest such spreads in the paper (up to ± 10.3 pp) across protocol columns, whereas EEGNet stays within ± 2.3 pp under both federated protocols. Wall-clock times come from a single RTX 5090 with four concurrent jobs and, subject to GPU contention and queuing, serve only as a conditional magnitude reference (Table 22).

F.5 Full Dataset, Preprocessing, and Training Specifications

Dataset specifications. BNCI2014_001: 0.1–75 Hz band-pass, $[0, 4)$ s time window, original 250 Hz sampling rate. Nakanishi2015: 1–70 Hz band-pass, $[0.15, 4.3)$ s time window, 256 Hz sampling rate. BNCI2014_009: 0.1–75 Hz band-pass, $[0, 0.8)$ s time window, 256 Hz sampling rate. Data are obtained through MOABB and frozen into a unified version. Within the sessions provided by MOABB, we use the first recording session of each subject for BNCI2014_001 (two per subject) and BNCI2014_009 (three per subject). Nakanishi2015 provides a single session, and MOABB exposes nine of the ten subjects of the original release. Splits follow the original temporal order within each client (runs, blocks, or trial order, as summarized in Table 1), and no trial appears in both parts. All reported test results are therefore *same-session, later-data* results for the participating subjects, not cross-session, cross-day, or new-subject results. Because P300 epochs are cut from a continuous recording rather than from independent segments, windows adjacent to the single train/test boundary of BNCI2014_009 may share part of their raw signal; this can affect only the trials immediately next to that boundary, and every transform is fitted on the training split alone.

Model-level preprocessing. Each model applies model-level preprocessing consistent with its reference implementation (its own sampling rate, frequency band, and normalization). On

BNCI2014_009, the default 8–32 Hz (motor-imagery) band of the from-scratch models filters out P300 slow waves and is uniformly replaced with the 0.3–75 Hz broadband. The remaining models keep the frequency bands of their reference implementations. Within a dataset, all models share the same frozen data, time windows, and client partition.

Training configuration details. All neural models use AdamW with learning rate 10^{-3}, weight decay 10^{-2}, and batch size 16. The 5-epoch linear warmup and the subsequent cosine annealing to 10^{-6} are counted in rounds for the federated protocols and in epochs for `Local` and `Centralized`. During training, every client's test set is evaluated every round (epoch). The main result is fixed to the actually trained final round without any test-curve-based model selection. In federated protocols, client optimizer states persist across rounds. All experiments are repeated over three random seeds (42/43/44).

Protocol semantics. All four protocols start every client from the same initial parameters. In the federated protocols, at each round every client trains locally, the shared parameters are averaged weighted by local sample counts, and the merged shared parameters are then re-broadcast to every client before the round's evaluation, so all clients evaluate the same shared representation. The private scope of `FedPer` is exactly the classification-head parameters defined by the per-model boundary of Table 23: those parameters, and only those, stay at the client, and the head is replaced by a task-sized output layer for every model. Optimizer state (AdamW moments) is client-local in all protocols. BatchNorm weight and bias are ordinary parameters and follow the same ownership rule as their enclosing module, but BatchNorm *running statistics* are buffers and are never aggregated in either federated protocol; they remain client-local, as they do for every buffer in the implementation. The communication accounting of Appendix F.1 follows the same definition and counts only named shared parameters. Consequently the `FedAvg–FedPer` contrast isolates the sharing of the classification head, but the two protocols already share a further, smaller deviation from full parameter averaging, namely client-local normalization statistics, and the magnitude of the head intervention is model-dependent: it removes about 81% of the parameters for EEG-Conformer but only about 0.5% for Deformer, so the same protocol difference does not correspond to the same amount of private capacity across backbones.

Table 23: **`FedPer` private scope per backbone**: the module whose parameters stay client-local, with every other parameter shared and aggregated. The module path is the name used in the released code, so the boundary can be checked against it directly.

Model	Private module (task head)
EEGNet	`main_model.classifier_block`
ShallowConvNet	`main_model.classifier_block`
EEG-Conformer	`main_model.2`
LMDA	`main_model.classifier`
Deformer	`main_model.mlp_head`
CBraMod	`main_model.task_head`
EEGMamba	`main_model.task_head`
EEGPT	`main_model.task_head`
LaBraM	`main_model.task_head`
Neuro-GPT	`main_model.task_head`

Model-level preprocessing matrix. Table 24 lists the frozen per-model signal contract. Every model resamples to its own target rate, band-pass filters in that band, and applies its own normalization, so within a dataset all models see the same frozen artifact and the same client partition but not the same tensor.

Table 24: **Model-level preprocessing** (frozen per model; applied after the dataset-level window). "–" marks no additional step. CAR: common average reference. EA: Euclidean alignment, fitted on training segments only.

Model	Rate (Hz)	Band (Hz)	Notch (Hz)	Additional
CSP+LDA	250	8–32	–	–
TRCA	250	8–40	–	–
xDAWN+LDA	250	1–24	–	CAR
EEGNet	250	8–32	–	CAR
ShallowConvNet	250	8–32	–	–
EEG-Conformer	250	8–32	–	CAR
LMDA	250	8–32	–	CAR
Deformer	250	8–32	–	CAR
CBraMod	200	4–32	60	CAR, 200-sample tokens
EEGMamba	200	0.5–50	–	EA
EEGPT	256	0.5–50	–	EA
LaBraM	200	0.1–75	50	scale 10^{-2}
Neuro-GPT	250	0.5–100	60	z-score

Input length. Several models carry a fixed input duration in their frozen contract. Where the dataset window is shorter, the signal is repeated from the start to reach that duration, which is the behavior of the reference loaders. On BNCI2014_009's 0.8 s window this produces 4.0 s inputs for EEG-Conformer, LMDA, Deformer, EEGMamba, EEGPT, LaBraM, and Neuro-GPT, a 1.0 s token-aligned input for CBraMod, and the native 0.8 s for EEGNet, ShallowConvNet, and xDAWN+LDA, whose 1 s contract is switched off by the P300 override of Appendix F.5. BNCI2014_001's 4 s window matches the fixed contract exactly; Nakanishi2015's 4.15 s window is trimmed to 4.0 s for those models while EEGNet, ShallowConvNet, and TRCA keep 4.15 s, and CBraMod is snapped to the nearest 200-sample token multiple (4.0 s). The resulting input lengths therefore differ across models within a dataset even though all models see the same frozen artifact.

On BNCI2014_009 only, the paradigm-inappropriate 8–32 Hz motor-imagery default of the from-scratch models and of CBraMod is replaced by a broadband 0.3–75 Hz (CBraMod: 0.1–75 Hz) band without CAR, matching the reference pipeline's run-entry semantics on P300 data; all other models and datasets keep the bands above. Euclidean alignment, where used, is fitted on training segments only and then applied to both splits, so no test statistic enters the transform.

Traditional-baseline specifications. CSP+LDA uses 10 spatial filters with log-variance features, TRCA uses 1 task-related component with template matching, and xDAWN+LDA uses 6 spatial filters with regularization coefficient 0.1. All three are deterministic and run once.

System environment. All experiments ran on a single NVIDIA GeForce RTX 5090 (32 GB) GPU with up to four training jobs scheduled concurrently. Wall-clock time therefore serves only as a conditional order-of-magnitude reference and does not support efficiency conclusions for dedicated hardware. The software stack comprises Python 3.12.3, PyTorch 2.12.1 (CUDA 13.0), source-compiled mamba-ssm 2.3.2 (required by EEGMamba), MNE 1.12.1, MOABB 1.5.0, NumPy 2.4.6, scikit-learn 1.9.0, and SciPy 1.18.0. Each run records the runtime environment, git revision, and per-client preprocessing parameters (band-pass, time window, sampling rate). Pre-trained weights are taken from the official releases of each model, and weight files are verified for integrity before use.