NEUROACT: ADVANCING EEG ANALYSIS AND DECODING VIA DUAL-LOOP AUTONOMOUS RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG analysis and decoding underpin brain-computer interfaces, cognitive assessment, and sleep monitoring. Fixed pipelines and reward-driven search do not explicitly determine how signal structure, representations, predictor behavior, and training dynamics contribute to performance or generalization gaps, leaving intervention choice underdetermined. We introduce NeuroAct, an autonomous research agent framework for EEG analysis and decoding. NeuroAct implements **dual-loop autonomous research** across EEG foundation models, task-specific predictors, and model combinations. In an inner loop, research, development, and review roles use evidence on signal structure, representations, predictor behavior, and training dynamics to test hypotheses, select analyses or interventions, and refine candidates using performance, stability, and cost feedback. An outer loop distills outcomes into conditional, testable recommendations with task and model conditions, then validates and revises them on subsequent tasks. We evaluate four settings: (A) EEG General Analysis, (B) Cross-task Model Optimization, (C) Single-task Open Research, and (D) Collaborative Prediction. In Setting A, NeuroAct's 80.88 macro-average exceeds BrainAgent's 63.13 by 17.75 points. In Setting B, NeuroAct raises EEGPT's mean balanced accuracy (BA) by 7.34 points and HBN Pearson's r from 0.6894 to 0.7362; in Setting C, it improves ADFTD BA by 16.33 points; and in Setting D, it raises Workload BA from 79.69% to 85.02%. Experience reuse raises BCIC-IV-2a BA from 34.66% to 50.81% while reducing LLM tokens by 64.02%, indicating recommendations improve cross-task performance. NeuroAct enables evidence-guided EEG research improved by cross-task experience.

1 INTRODUCTION

EEG analysis and decoding support cognitive-state assessment, sleep monitoring, and brain-computer interfaces. Pretrained EEG foundation models (EEG-FMs) provide reusable representations across diverse tasks (Jiang et al., 2024; Wang et al., 2024a; 2025a; El Ouahidi et al., 2025; Döner et al., 2026). Effective task adaptation requires assessing how signal structure, representations, predictor behavior, and training dynamics contribute to performance or generalization gaps, then selecting evidence that distinguishes competing explanations. Fixed pipelines and reward-driven optimization do not make this diagnostic decision explicit, motivating an ***autonomous research agent that improves within-task EEG decoding and refines research strategies across tasks.***

Task-specific EEG analysis must connect signal structure to how a model represents, predicts, and learns. Benchmark studies show model rankings and transfer depend on task and readout design, encoded features can differ from those used for prediction, and token preservation affects frozen-encoder decoding (Tang et al., 2026; Širca et al., 2026). When refinement of a single EEG model yields diminishing gains, this diagnosis can test whether other models correct its errors on samples. Error overlap determines whether complementary prediction merits a model-combination experiment.

Autonomous research systems combine hypothesis generation, experimentation, and interpretation in continuous workflows (Lu et al., 2026). Model-development systems further use experimental feedback or accumulated experience to revise their approaches (alphaXiv, 2026). In the EEG domain, BrainAgent orchestrates analysis tools (Zhou et al., 2026b), BrainPilot combines domain knowledge and research skills (Li et al., 2026a), and NeuroWeaver searches task pipelines (Wang et al., 2026). For EEG research, open questions concern how diagnostics guide interventions across foundation-

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.

Table 1: EEG analysis, decoding, and research guidance across agent frameworks (Wang et al., 2024c; Zhou et al., 2026b; Li et al., 2026a; Wang et al., 2026). ✓: supported; ◐: partial; —: unspecified.

Scope / mechanism	CodeAct	BrainAgent	BrainPilot	NeuroWeaver	**NeuroAct**
Primary scope	General coding	Signal analysis	Brain research	Pipeline search	Analysis & decoding
Research actions					
Code generation & execution	✓	—	✓	✓	✓
EEG analysis workflows	—	✓	✓	◐	✓
Model editing & training	✓	—	✓	✓	✓
EEG-FM adaptation & fine-tuning	—	—	—	—	✓
Joint multi-task training	—	—	—	—	✓
Model combination	—	—	—	—	✓
Feedback and experience					
Representation & gradient analysis	—	—	✓	—	✓
Feedback-based iteration	✓	◐	✓	✓	✓
Progressive capability disclosure	—	—	✓	—	✓
Cross-task experience revision	—	—	—	—	✓

model adaptation, task-specific modeling, joint training, and model combination, and how cross-task validation and revision make conditional recommendations useful on subsequent tasks.

We introduce **NeuroAct**, an autonomous research agent framework for EEG analysis and decoding. NeuroAct implements a **dual-loop autonomous research** framework: an inner loop selects analyses and interventions from evidence on signal structure, representations, predictor behavior, and training dynamics, while an outer loop distills, validates, and revises conditional recommendations or experiences across research tasks. Within the inner loop, the Coordinator directs the Researcher, Developer, and Reviewer to test competing diagnostic hypotheses through executable analyses or model changes, with predictive performance, stability, and cost informing the next action for each task. Each experience item is a conditional recommendation that records applicable EEG task and model conditions, a proposed action, supporting and counterevidence, and reassessment conditions. Subsequent tasks test its utility, while counterevidence narrows, suspends, or replaces the recommendation. Tab. 1 positions this scope relative to existing agent frameworks.

We organize NeuroAct's evaluation around **four research settings**: *(A) EEG General Analysis*, *(B) Cross-task Model Optimization*, *(C) Single-task Open Research*, and *(D) Collaborative Prediction*. The evaluation covers **six datasets** (HBN (Alexander et al., 2017; Shirazi et al., 2024), Workload (Zyma et al., 2019), PhysioMI (Schalk et al., 2004), BCIC-IV-2a (Tangermann et al., 2012), ADFTD (Miltiadous et al., 2023), and HMC (Alvarez-Estevez & Rijsman, 2021)), **three LLM backend families** (DeepSeek-V4 Flash (DeepSeek-AI et al., 2026), GPT-5.6 Luna (OpenAI, 2026), and Qwen3.8 Flash Next (Qwen Team, 2026)), and the available external baselines CodeAct (Wang et al., 2024c), BrainPilot (Li et al., 2026a), and BrainAgent (Zhou et al., 2026b). With DeepSeek-V4 Flash, NeuroAct reaches an 80.88 macro-average score in Setting A and exceeds 75 with Luna and Qwen on BrainBench (Zhou et al., 2026a). In Setting B, NeuroAct raises mean balanced accuracy (BA) from 47.50 to 54.84 and HBN Pearson's r from 0.6894 to 0.7362. NeuroAct gains 16.33 BA points on ADFTD in Setting C and raises Workload BA from the best single model's 79.69% to 85.02% in Setting D. Experience reuse transfers a PhysioMI-derived recommendation to BCIC-IV-2a, raising test BA from 34.66% to 50.81% while reducing LLM tokens by 64.02%. Our contributions are:

- **Evidence-guided inner-loop EEG research.** NeuroAct tests alternatives and selects analyses or interventions using signal, representation, predictor, and training evidence.
- **Outer-loop conditional strategy improvement.** NeuroAct tests conditional recommendations across tasks and revises them using matched performance and cost comparisons when warranted.
- **Executable EEG analysis and decoding.** NeuroAct operationalizes evidence through EEG General Analysis, Cross-task Model Optimization, Single-task Open Research, and Collaborative Prediction.

2 Related Work

2.1 EEG Decoding Models

Task-specific architectures including EEGNet and EEG Conformer encode local and temporal EEG structure (Lawhern et al., 2018; Song et al., 2023). Pretrained models including BENDR and EEGPT learn reusable representations, while other EEG-FMs differ in treating channels, signal structure, and recording heterogeneity (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024a; 2025a; El Ouahidi et al., 2025; Döner et al., 2026). These approaches expand candidate space from specialized decoders to transferable EEG representations.

Benchmark studies show model rankings, transfer, and generalization vary with task, dataset, and readout design (Banville et al., 2026; Xiong et al., 2026; Liu et al., 2026). Adaptation studies identify cohort shift, data availability, channel adaptation, and training regime as consequential (Nguyen et al., 2026; Yang et al., 2026; Kokate et al., 2026; Kommineni et al., 2026). NeuroAct turns evidence on signal structure, representations, predictor behavior, and training dynamics, including cross-model error overlap, into diagnostic decisions selecting analyses or interventions across adaptation, shared-model training, and model combination rather than a fixed recipe.

2.2 Autonomous Research Agents

Autonomous research systems coordinate hypothesis generation, experimentation, and interpretation in sustained workflows (Swanson et al., 2025; Gottweis et al., 2026; Ghareeb et al., 2026). Computational agents extend this workflow through executable tool use, candidate search, targeted refinement, repository-grounded development, diagnostic code updates, and training feedback, using execution outcomes to focus candidates (Jiang et al., 2025; Yamada et al., 2025; Nam et al., 2026; Yang et al., 2025; Li et al., 2026d; Zhang et al., 2026; Li et al., 2026b). Long-horizon systems support asynchronous experiments and interactive debugging (Hambardzumyan et al., 2026) and allocate effort across competing plans using feedback and remaining resources (Yu et al., 2026a).

Within EEG and brain-signal research, EEGAgent and BrainAgent orchestrate signal-processing tools and domain agents, while CogEEGAgent emphasizes deterministic cognitive-analysis execution and independent confirmation (Zhao et al., 2026; Zhou et al., 2026b; Hou et al., 2026). BrainPilot combines knowledge, skills, and specialists for architecture development under fixed preprocessing and training, whereas NeuroWeaver searches domain-constrained pipelines through tree selection, code refinement, and multi-objective rewards (Li et al., 2026a; Wang et al., 2026). NeuroAct extends the workflow with an EEG evidence loop that diagnoses gaps using signal, representation, predictor, and training evidence before selecting executable analyses or interventions.

2.3 Experience Reuse

Experience-transfer methods reuse outcomes across attempts or tasks. Reflexion and ExpeL reuse verbal feedback or contrasting trajectories, while ReasoningBank and EvoScientist distill and refine research strategies from outcomes (Shinn et al., 2023; Zhao et al., 2024; Ouyang et al., 2026; Lyu et al., 2026). Voyager and Recuris reuse executable skills or revise the memory governing their use, while SkillsBench and controlled studies examine when prescribed skills, tools, or guidance granularity hinder stronger strategies (Wang et al., 2024b; Yu et al., 2026b; Li et al., 2026c; Xu & Wu, 2026). NeuroAct treats experience as a conditional, testable recommendation under explicit EEG task and model conditions, validating its use on new tasks through outcomes and cost before retaining, revising, or narrowing applicability.

3 Method

NeuroAct organizes EEG analysis and decoding through two coupled research loops. The inner loop (Sec. 3.1) decomposes a shared objective into parallel questions, with the Coordinator directing specialists in separate contexts and selecting subsequent actions from reviewed evidence, experimental feedback, and remaining budget. The outer loop (Sec. 3.2) distills outcomes into conditional recommendations and revises them across tasks. Fig. 1 summarizes the framework; Alg. 1 specifies the interaction. Sec. 3.3 presents four settings that exercise different capabilities.

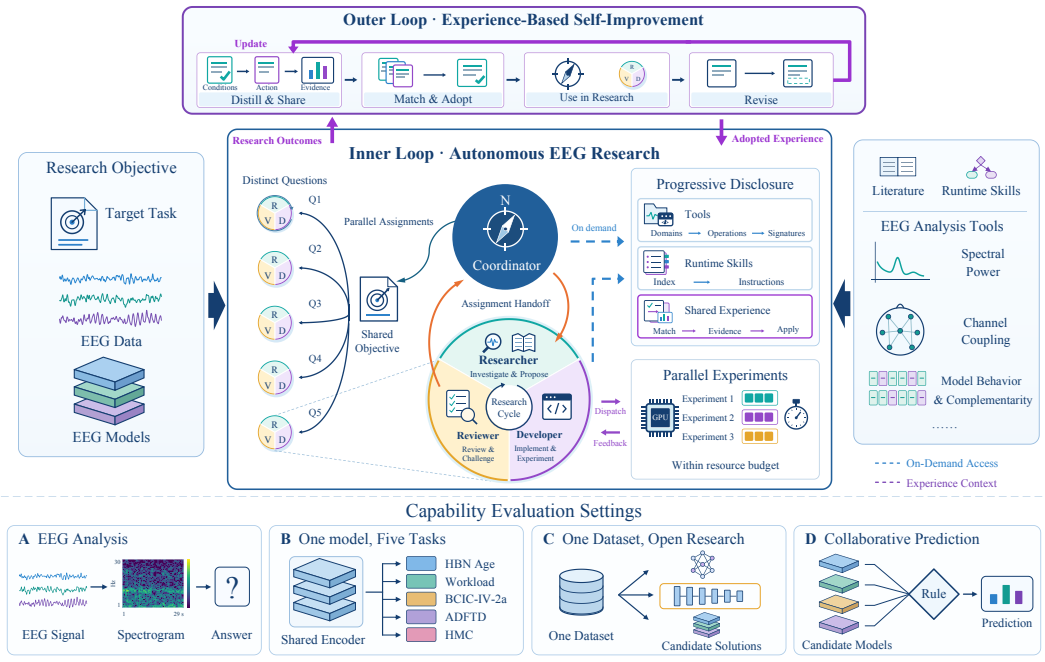

Figure 1: NeuroAct's dual-loop autonomous research framework. The Coordinator decomposes an EEG objective into parallel questions and directs separate cycles involving the Researcher, Developer, and Reviewer. The inner loop selects actions from reviewed evidence and experimental feedback within budget. The outer loop distills, matches, applies, and revises conditional experience across tasks. Capabilities are disclosed on demand; the lower panels show four evaluation settings.

3.1 Inner Loop: Autonomous EEG Research

Specialized collaboration. NeuroAct separates persona contexts and permissions so that scientific assessment, implementation, and review remain distinct. The ***Coordinator*** decomposes the objective into parallel questions and assigns each to a separate research cycle. The ***Researcher*** investigates an assigned question using literature, dataset semantics, and EEG evidence, then develops a feasible proposal. The ***Developer*** implements the selected proposal and executes the required analyses or training. The ***Reviewer*** examines the solution, implementation, and evidence, challenges unsupported conclusions, and returns the reviewed outcome to the Coordinator. Each *handoff* records the requested deliverable, supporting evidence, unresolved issues, and an actionable recommendation; the Coordinator integrates these handoffs, resolves disagreements, and selects subsequent actions.
Distinct investigations, coordinated decisions. For objective g, EEG data $\mathcal{D}$, and available model and code assets $\mathcal{A}$, the Coordinator selects a research decision u_t and specialist assignments $\mathcal{J}_t$,

$$(u_t, \mathcal{J}_t) = \pi_{\mathrm{C}}\big(g, \mathcal{D}, \mathcal{A}; \mathcal{H}_t, \mathcal{P}_t, \mathcal{K}_t, B_t\big), \tag{1}$$

where $\mathcal{H}_t$ records current-study evidence and decisions, $\mathcal{P}_t$ contains implemented solutions and results, $\mathcal{K}_t$ supplies applicable experience, and B_t is the remaining budget. The Coordinator can refine a candidate, redirect exploration, return to an earlier solution, or stop. An assignment $(r_i, g_{t,i}, \xi_{t,i})$ specifies a specialist role r_i and question $g_{t,i}$, with task context $\xi_{t,i}$ containing relevant evidence, adopted experience, and the requested deliverable. The specialist returns

$$h^r_{t,i} = \pi_r(g_{t,i}, \xi_{t,i}; \mathcal{T}_r), \qquad r \in \{\mathrm{R}, \mathrm{D}, \mathrm{V}\}, \tag{2}$$

where $h^r_{t,i}$ is the handoff returned to the Coordinator by role r for assignment i at iteration t, $\mathcal{T}_r$ denotes that role's tools, and V denotes the Reviewer. The policies are prompted rather than trained; conditioning on accumulated evidence and adopted experience lets research behavior evolve without LLM parameter updates. The Coordinator updates $\mathcal{H}_{t+1}$ with handoffs, $\mathcal{P}_{t+1}$ with completed solutions, and B_{t+1} from actual consumption. Research and review can proceed alongside training, while separate experiments run concurrently when resources permit. Distinct questions define parallel branches; the Coordinator compares reviewed outcomes and retains the best-supported solutions.

Algorithm 1 NeuroAct's inner and outer research loops

Require: Objective g, data $\mathcal{D}$, assets $\mathcal{A}$, budget B_0, shared experience $\mathcal{E}_j$
Set $t \leftarrow 0$; Initialize $\mathcal{H}_0, \mathcal{P}_0$
Retrieve and synthesize $\mathcal{E}_j$ for Coordinator adoption into $\mathcal{K}_0$ (Sec. 3.2).
while further research is warranted within budget **do**
  Coordinator selects $(u_t, \mathcal{J}_t)$ using Eq. (1) and records its rationale and expected effect.
  Dispatch $\mathcal{J}_t$ in parallel where resources permit.
  for each completed assignment $(r_i, g_{t,i}, \xi_{t,i}) \in \mathcal{J}_t$ **do**
    if $r_i = \mathrm{R}$, collect Researcher handoff $h^{\mathrm{R}}_{t,i}$.
    if $r_i = \mathrm{D}$, collect Developer handoff $h^{\mathrm{D}}_{t,i}$.
    if $r_i = \mathrm{V}$, collect Reviewer handoff $h^{\mathrm{V}}_{t,i}$.
  end for
  Integrate completed handoffs into $\mathcal{H}_{t+1}$; add completed solutions to $\mathcal{P}_{t+1}$.
  Reassess adopted recommendations against new outcomes; refresh $\mathcal{K}_{t+1}$ and remaining budget B_{t+1}; advance t.
end while
Distill eligible outcomes and recommendations in $\mathcal{E}_j$ using Eq. (4), yielding $\mathcal{E}_{j+1}$ for later studies.
Ensure: Coordinator-selected analytical answer or prediction solution with supporting evidence

Progressive disclosure. Role contexts initially receive compact capability indexes and expand selected details on demand. Tools expose domains before operations and input signatures; runtime skills expose indexes before instructions; matched experience exposes summaries before supporting evidence. Only selected capabilities and adopted advice enter specialist assignments, concentrating context on the current question while preserving access to literature, operational details, and evidence. App. B details the tools, skills, and disclosure levels.

Investigation becomes an experiment. The Researcher proposes an analysis whose possible outcomes distinguish competing explanations and imply different actions. Literature and evidence on signal structure, representations, predictor behavior, and training dynamics identify the uncertainty, a feasible intervention, and its expected effect. When computation is required, the Researcher specifies it for the Coordinator to assign to the Developer. The Developer executes it or implements scoped changes to input processing, pretrained-model adaptation, predictors, or training. Analytical tasks return computed results with evidence-supported interpretations.

Feedback redirects research. The Coordinator compares observed results and resource consumption with expected effects before selecting the next investigation. Reviewer challenges to an implementation or explanation can trigger further investigation. Auxiliary metrics and training behavior distinguish improved discrimination from decision-rule changes or unstable fitting. This evidence updates the solution, research state, and next decision.

3.2 Outer Loop: Experience-Based Self-Improvement

Distillation from contrasting outcomes. The outer loop distills experience about which experimental actions help under particular task and model conditions. The Coordinator compares an intervention's expected and observed effects with considered alternatives, producing a conditional recommendation,

$$k = (c, a, e^+, e^-, b), \tag{3}$$

where c specifies applicable task and model conditions, a the recommended action, e^+ and e^- supporting and conflicting evidence, and b the conditions for reassessment.

Matching and adoption. Retrieval filters ineligible or incompatible entries and ranks candidates by task and model compatibility, research-goal overlap, source evidence, EEG paradigm, and class structure. Similar recommendations are grouped into bounded summaries with source links. The Coordinator compares this synthesis with the target model and training conditions, choosing adoption, a bounded applicability check, or rejection. Candidate advice is shown first to the Coordinator; specialists receive only adopted content relevant to their assignments. Active recommendations enter role contexts within a character budget, with full evidence accessible on demand.

Revision through subsequent research. Target outcomes can narrow a recommendation, revise its action, or suspend its use. Matched trials assess transfer through performance and research cost.

Study j both uses and updates shared experience,

$$\mathcal{K}_{j,t} = \mathrm{Context}_{\mathrm{C}}(\mathcal{E}_j; g_j, \mathcal{D}_j, \mathcal{A}_j, \mathcal{H}_{j,t}), \qquad \mathcal{E}_{j+1} = \mathrm{Revise}_{\mathrm{C}}(\mathcal{E}_j; \mathcal{H}_j^{\mathrm{dev}}), \tag{4}$$

where $\mathrm{Context}_{\mathrm{C}}$ combines retrieval with the Coordinator's applicability decision, while $\mathrm{Revise}_{\mathrm{C}}$ distills or revises recommendations from eligible development outcomes. Revised experience changes decisions in Eq. (1), generating evidence for further revision. Unlike interaction memory, this recursion improves the agent's research strategy by revising which experiments it recommends and when, based on subsequent outcomes (Yu et al., 2026b).

3.3 Four Settings for EEG Research

EEG General Analysis (A). Given a question and EEG recordings, NeuroAct selects computations and sources to produce an evidence-supported answer. This setting exercises analytical reasoning and the investigative capabilities used in subsequent decoding research.
Cross-task Model Optimization (B). Each program adapts one pretrained encoder across heterogeneous datasets using task-specific adapters and output modules. NeuroAct can revise task sampling and loss contributions, then inspect the effects of shared updates on each target. Joint classification and regression create coupled objectives requiring balancing predictive utility across tasks.
Single-task Open Research (C). Given a dataset and prediction objective, NeuroAct can refine a strong candidate or explore another model family, including pretrained and task-specific predictors. This setting exercises open development guided by EEG evidence.
Collaborative Prediction (D). NeuroAct examines aligned errors or regression residuals for complementarity and develops a combination or routing program, $\hat{y} = F_\omega(\{f_m(x)\}_{m\in\mathcal{M}}, z(x))$, where $\mathcal{M}$ denotes members and $z(x)$ contains inference-time EEG features or outputs. Research can modify F_ω and, when permitted, train additional members to outperform the strongest predictor.

4 Experiments

Experimental setup. We evaluate how effectively NeuroAct links EEG analysis to predictive research and improves later studies through experience. Four settings test the breadth of these capabilities; mechanism ablations and experience studies measure effects on prediction and research cost. BrainBench (Setting A) evaluates instruction-driven EEG analysis across foundational, sleep, neurocognitive, and physiological domains. Settings B, C, and D cover Workload, BCIC-IV-2a, ADFTD, HMC, and HBN; PhysioMI supports experience and supplementary studies. DeepSeek-V4 Flash is default LLM; GPT-5.6 Luna and Qwen3.8 Flash Next assess provider consistency, with CodeAct, BrainPilot, and BrainAgent as baseline agents. For Settings B–D, we adapt the baselines' task and execution interfaces. Tab. 1 describes originally reported capabilities; Tab. 3(B-D) evaluates the adapted implementations. We include BENDR, BIOT, LaBraM, EEGPT, CBraMod, CSBrain, and REVE in model pools for Setting D, with ShallowFBCSPNet added for ADFTD. Classification uses balanced accuracy and AUROC (BA, and AUC in tables); HBN regression uses Pearson r and prediction errors. Protected test labels remain unavailable during dual-loop research. The research budget for each study is capped at 50 million LLM tokens or 5 GPU hours. Research costs include all attempted experiments. **Apps. A and E detail tasks, execution, and protocols; App. D reports member pools and complete metrics.**

4.1 Main Results Across the Four Research Settings

We evaluate diverse dataset, EEG model, and LLM configurations. Tab. 2 follows NeuroAct's outcomes and reference solutions; Tab. 3 compares external agent frameworks.
EEG General Analysis across domains. In Tab. 2(A), NeuroAct achieves an 80.88 macro score with DeepSeek and exceeds 75 with both Luna and Qwen, showing consistent analytical capability across providers. DeepSeek attains 87.26 in foundational analysis and 92.16 in physiological integration, demonstrating effective use of EEG computations across distinct analytical objectives. Tab. 3(A) compares these outcomes with external agents, with NeuroAct leading every reported domain.
Cross-task Model Optimization balances heterogeneous target tasks. Tab. 2(B) examines joint classification and regression with a shared encoder. With EEGPT, NeuroAct raises four-target mean BA from 47.50% to 54.84% and HBN r from 0.6894 to 0.7362, while reducing MSE from 6.080 to 5.447. These gains within one shared checkpoint show its ability to balance heterogeneous targets.

Table 2: NeuroAct research outcomes. Setting D reports the best single member, equal-weight reference, and NeuroAct-developed solution from its member pool.

(A) EEG General Analysis

LLM	Found. ↑	Sleep ↑	Neurocog. ↑	Physiol. ↑	Macro ↑
DeepSeek	**87.26**	**75.36**	**68.73**	**92.16**	**80.88**
Luna	84.96	70.27	67.54	79.55	75.58
Qwen	78.67	72.20	65.89	84.27	75.26

(B) Cross-task Model Optimization

		Workload		BCIC-IV-2a		ADFTD		HMC		HBN	
Model	Stage	BA ↑	AUC ↑	BA ↑	AUC ↑	BA ↑	AUC ↑	BA ↑	AUC ↑	r ↑	MAE ↓
LaBraM	Initial	50.00	0.598	28.65	0.565	35.28	0.555	69.95	0.917	0.653	2.009
	NeuroAct	**65.29**	**0.770**	**33.94**	**0.625**	**56.37**	**0.726**	**70.74**	**0.928**	**0.656**	**1.998**
REVE	Initial	49.55	0.815	25.43	0.581	**53.09**	**0.753**	52.80	0.857	0.616	2.149
	NeuroAct	**76.55**	**0.850**	**50.26**	**0.832**	51.98	0.749	**66.02**	**0.918**	**0.719**	**1.810**
EEGPT	Initial	49.55	0.679	31.60	0.616	**46.51**	0.647	62.33	0.901	0.689	1.908
	NeuroAct	**68.00**	**0.735**	**34.03**	**0.710**	45.98	**0.659**	**71.35**	**0.932**	**0.736**	**1.749**
CBraMod	Initial	50.00	0.624	28.82	0.624	46.05	**0.732**	60.29	0.887	0.624	2.058
	NeuroAct	**53.93**	**0.763**	**50.69**	**0.778**	**52.54**	0.720	**69.95**	**0.918**	**0.633**	**1.962**

(C) Single-task Open Research

	Workload		BCIC-IV-2a		ADFTD		HMC		HBN	
Stage	BA ↑	AUC ↑	BA ↑	AUC ↑	BA ↑	AUC ↑	BA ↑	AUC ↑	r ↑	MAE ↓
Compared	81.92	0.884	52.60	0.789	54.93	**0.751**	73.03	0.934	0.488	2.425
NeuroAct	**83.42**	**0.904**	**60.59**	**0.839**	**71.26**	0.748	**73.36**	**0.938**	**0.715**	**1.791**

(D) Collaborative Prediction

	Workload		BCIC-IV-2a		ADFTD		HMC		HBN	
Predictor	BA ↑	AUC ↑	BA ↑	AUC ↑	BA ↑	AUC ↑	BA ↑	AUC ↑	r ↑	MAE ↓
Best single	79.69	0.888	60.85	**0.844**	71.26	0.748	73.12	0.934	0.758	1.661
Equal weights	77.95	0.895	58.07	0.837	64.44	0.799	75.02	0.943	0.738	1.757
NeuroAct	**85.02**	**0.897**	**61.02**	0.841	**74.89**	**0.843**	**75.14**	**0.945**	**0.780**	**1.585**

The REVE and CBraMod studies extend shared-model research to different encoder families. In the LaBraM comparison in Tab. 3(B), NeuroAct leads the external agents on every target, including 56.37% ADFTD BA against the strongest external result of 33.33%. App. D gives full results.

Single-task Open Research develops task-specific solutions. Given a dataset and objective, Setting C lets the agent consult literature, implement methods, and redirect research. Compared in Tab. 2(C) restricts research to existing program interfaces without authoring new scientific code. On ADFTD, NeuroAct implements class-balanced training and explores a different architecture, improving BA from 54.93% with EEGNet to 71.26% with ShallowFBCSPNet. On BCIC-IV-2a, the REVE solution reaches 60.59% BA and 0.839 AUROC, improving on the restricted reference's 52.60% BA. Tab. 3(C) shows a 15.36-point BCIC-IV-2a gain over the strongest completed external result. These results show the value of revising the predictive approach as evidence accumulates. Metrics appear in App. D.

Collaborative Prediction turns complementarity into usable gain. Tab. 2(D) compares NeuroAct with individual predictors and averaging. On Workload, averaging ranks samples well; shifting pooled scores using training class proportions raises BA to 85.02%, exceeding the best member's 79.69%. On HBN, convex weights shrunk toward equal weighting improve Pearson r from 0.758 to 0.780 and reduce MAE from 1.661 to 1.585 over the best member. NeuroAct leads external agents on Workload in Tab. 3(D). On ADFTD, sample-dependent combination improves BA from the best member's 71.26% to 74.89% and AUROC from 0.748 to 0.843. App. F details these methods.

4.2 Ablations and Experience-Based Improvement

Ablating the research mechanisms. Tab. 4 fixes CBraMod and DeepSeek on three datasets to test the framework. The ablations replace specialist collaboration with one persona, restrict EEG analysis to basic statistics, remove runtime skills, or limit code editing to configurable registered recipes rather

Table 3: Agent comparison with DeepSeek. Setting A includes source-reported CodeAct and BrainAgent scores (Zhou et al., 2026a). Setting B uses LaBraM. Settings B, C, and D report BA (%) and HBN Pearson r. Metrics are detailed in App. D.

(A) EEG General Analysis

Agent	Found.	Sleep	Neuro.	Phys.	Macro
CodeAct	68.16	61.01	60.38	54.52	61.02
BrainAgent	67.38	67.97	67.21	49.98	63.13
BrainPilot	64.82	58.43	50.83	61.52	58.90
NeuroAct	**87.26**	**75.36**	**68.73**	**92.16**	**80.88**

(B) Cross-task Model Optimization: LaBraM

Agent	Work.	BCIC-IV-2a	ADFTD	HMC	HBN
CodeAct	50.00	25.00	32.39	39.30	0.228
BrainAgent	50.00	27.78	33.33	36.34	0.334
BrainPilot	49.55	28.12	24.82	60.30	0.500
NeuroAct	**65.29**	**33.94**	**56.37**	**70.74**	**0.656**

(C) Single-task Open Research

Agent	Work.	BCIC-IV-2a	ADFTD	HMC	HBN
CodeAct	60.11	45.23	56.60	72.90	0.687
BrainAgent	83.04	41.75	70.51	48.09	0.559
BrainPilot	64.54	40.45	62.25	71.46	0.547
NeuroAct	**83.42**	**60.59**	**71.26**	**73.36**	**0.715**

(D) Collaborative Prediction

Agent	Work.	BCIC-IV-2a	ADFTD	HMC	HBN
CodeAct	77.58	35.50	58.28	74.46	0.760
BrainAgent	76.44	60.68	52.61	74.50	0.758
BrainPilot	76.44	60.59	62.04	73.91	0.767
NeuroAct	**85.02**	**61.02**	**74.89**	**75.14**	**0.780**

Table 4: Mechanism ablations with DeepSeek and a fixed CBraMod family. Full reports the best complete-framework BA within this model constraint. App. E.2 gives complete metrics.

	Workload		BCIC-IV-2a		ADFTD	
Ablation	BA ↑	AUC ↑	BA ↑	AUC ↑	BA ↑	AUC ↑
Full NeuroAct	**77.73**	**0.886**	**56.34**	**0.805**	**62.94**	**0.831**
w/o specialist personas	77.11	0.853	51.65	0.778	61.29	0.723
w/o EEG diagnostics	72.63	0.752	52.17	0.798	41.80	0.579
w/o code editing	66.22	0.722	52.60	0.789	48.70	0.576
w/o runtime skills	70.60	0.783	52.60	0.788	48.95	0.633
w/o auxiliary metrics	77.29	0.862	49.74	0.786	61.62	0.829

than workspace implementations. The auxiliary-metric ablation receives only the primary score; Full additionally receives macro-F1, AUROC, training loss, and gradient information. Each ablation removes one capability while preserving the surrounding setup; App. E.2 provides full metrics and costs. Full NeuroAct improves Workload BA by 11.51 points over registered recipes, demonstrating the value of interventions beyond parameter selection. Full exceeds the basic-statistics condition by 21.14 BA points on ADFTD and the primary-metric-only condition by 6.60 points on BCIC-IV-2a. Runtime skills improve BA across all three datasets. Fig. 2(a) reports paired runtime-skill studies on Workload, ADFTD, and PhysioMI in App. E.2, showing better prediction with lower token use on Workload and PhysioMI and an ADFTD gain of 8.57 percentage points at similar token cost.

Experience transfer and revision. Tab. 5 defaults to CBraMod and tests experience in later research. Cross-dataset transfer carries patch-information advice from PhysioMI to BCIC-IV-2a. Cross-model transfer fixes PhysioMI and provides CBraMod-derived experience to BENDR or BIOT. Each begins with or without the source recommendation. The revision study uses new source evidence to update advice on interpreting training dynamics and planning experiments, then tests it in new research. Updated advice raises PhysioMI BA from 62.36% to 63.13% and reduces research time from 149.8 to 139.1 minutes. Cross-dataset reuse raises BCIC-IV-2a BA from 34.66% to 50.81% and cuts tokens from 29.21M to 10.51M. BENDR has similar BA with 24.28% fewer research tokens and higher AUROC. BIOT gains 2.34 BA points with 12.34% fewer tokens and 22.08% less research time. Fig. 2(b) distinguishes performance and cost effects. App. E.3 provides complete metrics.

4.3 Research Decision Traces

EEG analysis shapes an executable predictive method. In an ADFTD study, the Researcher examines the recording montage and investigates spectral features suited to the task. The absence of a declared reference makes absolute power difficult to compare, motivating within-window relative power and frequency-band ratios. The Developer implements these features with regional aggregation and a gradient-boosting classifier. Error analysis finds low recall for frontotemporal dementia (FTD), with nonzero recall for every class. The agent treats class separation as the next research question

Table 5: Experience transfer and revision with DeepSeek. All rows report full research costs.

Study	Experience	BA ↑	AUC ↑	Tokens (M) ↓	Min ↓
PhysioMI → BCIC-IV-2a	w/o	34.66	0.592	29.21	236.7
	with	**50.81**	**0.775**	**10.51**	**165.3**
CBraMod → BENDR	w/o	**58.65**	0.806	29.00	264.1
	with	58.28	**0.811**	**21.96**	**219.2**
CBraMod → BIOT	w/o	47.92	0.730	10.78	209.7
	with	**50.26**	**0.749**	**9.45**	**163.4**
PhysioMI advice revision	initial	62.36	0.840	**13.90**	149.8
	revised	**63.13**	**0.847**	15.43	**139.1**

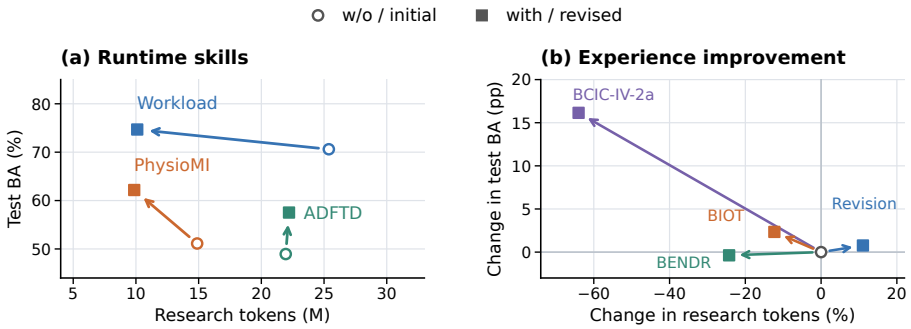

Figure 2: Performance-cost tradeoffs. (a) Absolute outcomes by runtime-skill condition. (b) Changes from each study's w/o or initial-advice reference at the origin. Token changes are percentages. Upward-leftward movement indicates better prediction at lower token cost.

and tests complementary connectivity and spatial-covariance features. The spectral solution remains the strongest evaluated approach in this investigation, reaching 67.27% BA and 0.811 AUROC. EEG-specific analysis thus informs both implementation and the next experiment.

Model feedback changes the intervention. Workload research illustrates two ways to act on model behavior. In one study, early overfitting and variable validation performance motivate training-time perturbations scaled to the EEG amplitude. The resulting CBraMod comparison increases BA from 65.98% to 71.62% and AUROC from 0.768 to 0.805. In another, lowering the encoder learning rate leaves validation BA unchanged at 70.44%. Investigation of class imbalance then leads to a class-weighted objective, reaching 76.89% validation BA and 75.55% test BA. The traces illustrate how feedback can redirect an intervention toward a better-supported explanation.

Experience focuses the next experiment. The PhysioMI recommendation directs BCIC-IV-2a research toward retaining CBraMod patch information. The first implementation also changes the training recipe, so the Coordinator requests a narrower comparison that preserves the established recipe. The resulting solution reaches 50.81% BA with 64.02% fewer research tokens than the study without experience. Reuse thus provides a concrete research direction, while target-side inspection refines how it is tested. App. G details the evidence and implementations behind these decisions.

5 Conclusion

NeuroAct advances autonomous EEG analysis and decoding by making domain evidence part of experimental decision-making rather than treating adaptation as undirected search. Its inner loop turns investigation into executable interventions, while the outer loop carries useful experience into later research and revises it with new evidence. The four settings demonstrate EEG General Analysis, Cross-task Model Optimization, Single-task Open Research, and Collaborative Prediction. Fixed-model ablations support the contribution of specialized research capabilities, and experience studies across datasets and model families show both predictive gains and reduced research cost. These results establish a path toward EEG research agents that improve current solutions, accumulate validated recommendations across tasks, and conduct subsequent studies more effectively.

AI USE STATEMENT

AI was used primarily to run the research agents evaluated in this work and to assist with manuscript writing and language polishing. The authors reviewed and revised the manuscript and take full responsibility for its content.

ETHICS STATEMENT

This study reuses existing EEG datasets and does not recruit new participants. EEG recordings and associated demographic or clinical labels require appropriate privacy protections, and dataset use remains subject to the original access and licensing terms. We report aggregate benchmark results. NeuroAct is evaluated as a research framework; clinical use would require independent validation and assessment of demographic and recording-condition shifts.

REPRODUCIBILITY STATEMENT

Sec. 3 describes NeuroAct's research procedure. App. A documents dataset views, prediction targets, and evaluation units; Apps. B and C specify the runtime, tools, skills, and role instructions. Complete predictive metrics appear in App. D, while App. E details baseline adaptations, ablations, and experience studies. Apps. F and G describe implemented interventions and the evidence behind representative research decisions.

REFERENCES

Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, Shannon Litke, Bridget O'Hagan, Jennifer Andersen, Batya Bronstein, Anastasia Bui, Marijayne Bushey, Henry Butler, Victoria Castagna, Nicolas Camacho, Elisha Chan, Danielle Citera, Jon Clucas, Samantha Cohen, Sarah Dufek, Megan Eaves, Brian Fradera, Judith Gardner, Natalie Grant-Villegas, Gabriella Green, Camille Gregory, Emily Hart, Shana Harris, Megan Horton, Danielle Kahn, Katherine Kabotyanski, Bernard Karmel, Simon P. Kelly, Kayla Kleinman, Bonhwang Koo, Eliza Kramer, Elizabeth Lennon, Catherine Lord, Ginny Mantello, Amy Margolis, Kathleen R. Merikangas, Judith Milham, Giuseppe Minniti, Rebecca Neuhaus, Alexandra Levine, Yael Osman, Lucas C. Parra, Ken R. Pugh, Amy Racanello, Anita Restrepo, Tian Saltzman, Batya Septimus, Russell Tobe, Rachel Waltz, Anna Williams, Anna Yeo, Francisco X. Castellanos, Arno Klein, Tomas Paus, Bennett L. Leventhal, R. Cameron Craddock, Harold S. Koplewicz, and Michael P. Milham. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4(1):170181, 2017. doi: 10.1038/sdata.2017.181.

alphaXiv. OpenResearch. Computer software, 2026. URL `https://github.com/alphaXiv/OpenResearch`. Accessed September 18, 2026.

Diego Alvarez-Estevez and Roselyne M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):e0256111, 2021. doi: 10.1371/journal.pone.0256111.

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Mingfang (Lucy) Zhang, Elisa Cascardi, Katelyn Begany, Teon Brooks, and Jean-Rémi King. NeuralBench: A unifying framework to benchmark NeuroAI models. arXiv preprint arXiv:2605.08495, 2026.

DeepSeek-AI et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. Technical report, DeepSeek-AI, 2026. URL `https://arxiv.org/abs/2606.19348`.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: efficient and topology-agnostic foundation model for eeg signal analysis. volume 38, pp. 70682–70708, 2026.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG - adapting to any setup

with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, pp. 22541–22577. Curran Associates, Inc., 2025. doi: 10.52202/085713-0760.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655(8122):497–505, 2026. doi: 10.1038/s41586-026-10652-y.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Karen Hambardzumyan, Nicolas Baldwin, Edan Toledo, Rishi Hazra, Michael Kuchnik, Bassel Al Omari, Thomas Simon Foster, Anton Protopopov, Jean-Christophe Gagnon-Audet, Ishita Mediratta, Kelvin Niu, Michael Shvartsman, Alisia Lupidi, Alexis Audran-Reiss, Parth Pathak, Tatiana Shavrina, Despoina Magka, Hela Momand, Derek Dunfield, Nicola Cancedda, Pontus Stenetorp, Carole-Jean Wu, Jakob Nicolaus Foerster, Yoram Bachrach, and Martin Josifoski. AIRA_2: Overcoming bottlenecks in AI research agents. arXiv preprint arXiv:2603.26499, 2026.

Dengzhe Hou, Lingyu Jiang, Fangzhou Lin, and Kazunori D Yamada. CogEEGAgent: Toward autonomous cognitive EEG analysis with grounded execution and selection-aware verification. arXiv preprint arXiv:2607.25045, 2026.

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*, pp. 16405–16426, 2024.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. AIDE: AI-driven exploration in the space of code. arXiv preprint arXiv:2502.13138, 2025.

Kuntal Kokate, Bruno Aristimunha, Dung Truong, and Arnaud Delorme. Channel adaptation for EEG foundation models: A systematic benchmark across architectures, tasks, and training regimes. arXiv preprint arXiv:2604.23091, 2026.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Tiantian Feng, and Shrikanth Narayanan. A multi-dimensional framework for evaluating generalization in EEG foundation models. In *Proceedings of the 7th Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pp. 221–244. PMLR, 2026.

Demetres Kostas, Stephane 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.

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.

Haoxuan Li, Tianci Gao, Jianhe Li, Yang Fan, Runze Shi, Weiran Wang, Tianxiang Zhao, Zezhao Wu, Xiaoyang Jiang, Qihui Zhang, Jia Li, Xiao Xiao, Kai Du, Xiaoxuan Jia, Chao Xie, and Lu Mi. BrainPilot: Automating brain discovery with agentic research. arXiv preprint arXiv:2607.15079, 2026a.

Qizheng Li, Yifei Zhang, Xiao Yang, Xu Yang, Zhuo Wang, Weiqing Liu, and Jiang Bian. FT-Dojo: Towards autonomous LLM fine-tuning with language agents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, Zonglin Di, Yifeng He, Shenghan Zheng, Kyoung Whan Choe, Jiankai Sun, Shuyi Wang, Chujun Tao, Binxu Li, Xuandong Zhao, Hejia Geng, Xiaojun Wu, Junwei Zhou, Xiaokun Chen, Hanwen Xing, Yubo Li, Qunhong Zeng,

Di Wang, Yuanli Wang, Roey Ben Chaim, Penghao Jiang, Haotian Shen, Luyang Kong, Xinyi Liu, Runhui Wang, Xuanqing Liu, Jiachen Li, Xin Lan, Yueqian Lin, Wengao Ye, Junwei He, Songlin Li, Yue Zhang, Yipeng Gao, Yijiang Li, Ze Ma, Liqiang Jing, Tianyu Wang, Kaixin Li, Yiqi Xue, Haoran Lyu, Yizhuo He, Yuchen Tian, Shutong Wu, Bowei Wang, Yixuan Gao, Bo Chen, Litong Liu, Sikai Cheng, Jiajun Bao, Shuaicheng Tong, Shuwen Xu, Terry Yue Zhuo, Tinghan Ye, Qi Qi, Miao Li, Longtai Liao, Zelin Tan, Chang Shi, Xilin Tang, Srinath Tankasala, Boqin Yuan, Yaoyao Qian, Jianhong Tu, Chenguang Wang, Yizhou Sun, Wei Wang, Aaron Taylor, Ziyue Yang, Changkun Guan, Zhikang Dong, Xinyu Zhang, Steven Dillmann, Han-chung Lee, and Dawn Song. SkillsBench: Benchmarking how well agent skills work across diverse tasks. arXiv preprint arXiv:2602.12670, 2026c.

Yu Li, Chenyang Shao, Xinyang Liu, Ruotong Zhao, Peijie Liu, Hongyuan Su, Zhibin Chen, Qinglong Yang, Anjie Xu, Yi Fang, Qingbin Zeng, Tianxing Li, Jingbo Xu, Fengli Xu, Yong Li, and Tie-Yan Liu. AutoSOTA: An end-to-end automated research system for state-of-the-art AI model discovery. arXiv preprint arXiv:2604.05550, 2026d.

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*, pp. nwag466, 2026. doi: 10.1093/nsr/nwag466.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Yougang Lyu, Xi Zhang, Xinhao Yi, Yuyue Zhao, Shuyu Guo, Wenxiang Hu, Jan Piotrowski, Jakub Kaliski, Jacopo Urbani, Zaiqiao Meng, Lun Zhou, and Xiaohui Yan. EvoScientist: Towards multi-agent evolving AI scientists for end-to-end scientific discovery. arXiv preprint arXiv:2603.08127, 2026.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6): 95, 2023. doi: 10.3390/data8060095.

Jaehyun Nam, Jinsung Yoon, Jiefeng Chen, Jinwoo Shin, Sercan Arik, and Tomas Pfister. Mle-star: Machine learning engineering agent via search and targeted refinement. volume 38, pp. 116692–116712, 2026.

Anh T. Nguyen, Zihua Sun, and Michelle J. Johnson. Lightweight adaptation of EEG foundation models for stroke motor imagery decoding: Domain shift and subject-level robustness. arXiv preprint arXiv:2609.00282, 2026.

OpenAI. GPT-5.6 system card. System card, OpenAI, July 2026. URL `https://deploymentsafety.openai.com/gpt-5-6`.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026.

Qwen Team. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. Technical report, Alibaba Group, August 2026. URL `https://arxiv.org/abs/2608.30320`.

G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/tbme.2004.827072.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B Esper, Dung Truong, Arnaud Delorme, Michael P Milham, and Scott Makeig. Hbn-eeg: The fair implementation of the healthy brain network (hbn) electroencephalography dataset. *bioRxiv*, pp. 2024–10, 2024.

Urban Širca, Maryam Alimardani, Stefanos Zafeiriou, and Konstantinos Barmpas. Beyond accuracy: Robustness, interpretability and expressiveness of EEG foundation models. arXiv preprint arXiv:2605.17562, 2026.

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.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Ling Tang, Qian Chen, Jilin Mei, Houshi Xu, Quanshi Zhang, Jing Shao, Na Zou, Xia Hu, and Dongrui Liu. What do EEG foundation models capture from human brain signals? arXiv preprint arXiv:2605.11410, 2026.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai Joshua Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI Competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

Guagnyu 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., 2024a. doi: 10.52202/079017-1239.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856.

Guoan Wang, Shihao Yang, and Feng Liu. NeuroWeaver: An autonomous evolutionary agent for exploring the programmatic space of EEG analysis pipelines. arXiv preprint arXiv:2602.13473, 2026.

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 *International Conference on Learning Representations*, pp. 75310–75346, 2025a.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50208–50232. PMLR, 2024c.

Xingyao Wang, Simon Rosenberg, Juan Michelini, Calvin Smith, Hoang Tran, Engel Nyst, Rohit Malhotra, Xuhui Zhou, Valerie Chen, Robert Brennan, and Graham Neubig. The OpenHands Software Agent SDK: A composable and extensible foundation for production agents, 2025b. URL `https://arxiv.org/abs/2511.03690`.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Xiaonan Xu and Wenjing Wu. Skill availability and presentation granularity in large-language-model agents: A controlled SkillsBench study. arXiv preprint arXiv:2605.31408, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv preprint arXiv:2504.08066, 2025.

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*, volume 36, pp. 78240–78260. Curran Associates, Inc., 2023. doi: 10.52202/075280-3420.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M Van Hulle. Are eeg foundation models worth it? comparative evaluation with traditional decoders in diverse bci tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xu Yang, Xiao Yang, Shikai Fang, Yifei Zhang, Jian Wang, Bowen Xian, Qizheng Li, Jingyuan Li, Minrui Xu, Yuante Li, Haoran Pan, Yuge Zhang, Weiqing Liu, Yelong Shen, Weizhu Chen, and Jiang Bian. R&D-Agent: An LLM-agent framework towards autonomous data science. arXiv preprint arXiv:2505.14738, 2025.

Xinle Yu, Fan Bai, Kaiser Sun, Hengshuo Miao, Abhay Anand, Zhongyan Luo, Kun Zhou, and Zhen Wang. PrimeScientist: Strategic allocation of research effort in autonomous research. arXiv preprint arXiv:2609.17846, 2026a.

Zhaochen Yu, Yingcheng Wu, Zhenfei Yin, Kaiyuan Chen, Zhe Zhao, Mengdi Wang, Shuicheng Yan, and Ling Yang. Recursive experiential–working memory evolution for long-horizon agent harnesses. arXiv preprint arXiv:2608.24876, 2026b.

Yifei Zhang, Xu Yang, Xiao Yang, Bowen Xian, Qizheng Li, Shikai Fang, Jingyuan Li, Jian Wang, Minrui Xu, Yuge Zhang, Weiqing Liu, and Jiang Bian. Reasoning as gradient: Scaling MLE agents beyond tree search. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9013–9038. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.438.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936.

Sha Zhao, Mingyi Peng, Haiteng Jiang, Tao Li, and Shijian Li. EEG Agent: A unified framework for automated EEG analysis using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18063–18071, 2026. doi: 10.1609/aaai.v40i21.38867.

Yangxuan Zhou, Yuning Chen, Chen Wu, Jiquan Wang, Shijian Li, Gang Pan, and Sha Zhao. BrainBench: Benchmarking large language models for comprehensive EEG understanding. arXiv preprint arXiv:2608.04156, 2026a.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Shijian Li, and Gang Pan. BrainAgent: A large language model-driven multi-agent framework for autonomous brain signal understanding. arXiv preprint arXiv:2606.25400, 2026b.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019. doi: 10.3390/data4010014.

A Data and Prediction Tasks

A.1 Prediction targets and evaluation units

HBN provides pediatric age regression from resting EEG (Alexander et al., 2017; Shirazi et al., 2024); Workload distinguishes mental-arithmetic conditions (Zyma et al., 2019); PhysioMI (Schalk et al., 2004) and BCIC-IV-2a (Tangermann et al., 2012) provide four-class motor imagery; ADFTD distinguishes Alzheimer's disease, frontotemporal dementia, and healthy controls (Miltiadous et al., 2023); HMC provides five-stage sleep classification (Alvarez-Estevez & Rijsman, 2021). Recordings are partitioned at the required subject or recording level before constructing prediction units.

Table 6: Dataset characteristics and prediction-unit counts.

Dataset	Target	Channels	Window (s)	Training	Validation	Test
HBN	Age regression	129	2	121,740	28,620	19,740
Workload	Binary classification	19	4	1,537	300	297
PhysioMI	4-class classification	64	4	6,210	1,734	1,893
ADFTD	3-class classification	19	10	4,743	1,115	1,155
HMC	5-class classification	4	30	91,681	22,804	22,440
BCIC-IV-2a	4-class classification	22	4	2,784	1,152	1,152

Training, validation, and test splits follow EEG-FM-Bench (Xiong et al., 2026) for the classification datasets and NeuralBench (Banville et al., 2026) for HBN. Tab. 6 lists their sizes. HBN contributes 60 windows per subject, using R10 for validation and R5 for testing.

HBN metrics are computed over windows, with MAE and RMSE in years and MSE in squared years. For classsification tasks, BA is mean class recall. Macro-F1 averages class F1 scores. AUROC uses the recorded binary or multiclass one-versus-rest calculation. Evaluation units and metric implementations are held fixed within each comparison.

A.2 Model inputs and pretraining provenance

Predictor inputs use 200 Hz, except EEGPT and LUNA at 256 Hz. EEGPT converts microvolt (μV) signals to millivolts (mV) by multiplying by 0.001; CBraMod, LaBraM, and CSBrain multiply μV values by 0.01. The ADFTD ShallowFBCSPNet uses μV signals directly.

BENDR maps each window to $[-1, 1]$ using its minimum and maximum over all channels and time points. BIOT divides each channel by the 95th percentile of its absolute values within the window. EEGNet and LUNA apply channel-wise window z-scores. REVE also uses channel-wise window z-scores and clips values to $[-15, 15]$. Its HBN data undergo channel-wise standardization over complete recordings before windowing (Banville et al., 2026); candidate settings either preserve these values or apply an additional window-level z-score.

LaBraM and EEGPT include PhysioMI in their pretraining corpora (Jiang et al., 2024; Wang et al., 2024a). Main comparisons for Settings B, C, and D and the main CBraMod ablations therefore use BCIC-IV-2a for motor imagery. REVE's pretraining corpus also includes HBN releases, creating dataset overlap for its HBN evaluations and collaborative predictors containing REVE (El Ouahidi et al., 2025). Following NeuralBench (Banville et al., 2026), we retain HBN to evaluate agent-driven improvements under pretrained initializations with known data overlap.

A.3 BrainBench comparison conditions

The official BrainBench release contains 950 Foundational Analysis, 1,025 Sleep Assessment, 1,035 Neurocognitive Assessment, and 1,095 Physiological Integration cases (Zhou et al., 2026a). Difficulty weighting is applied within subsets, followed by an equal average of their scores. Execution errors remain in the denominator.

The CodeAct and BrainAgent scores in Tab. 3(A) are source-reported BrainBench results (Zhou et al., 2026a); BrainPilot and NeuroAct scores come from our evaluations. App. E describes the execution configurations.

B NeuroAct Harness and Implementation

B.1 Runtime, roles, and execution

NeuroAct implements the dual-loop framework using the OpenHands Software Agent SDK (Wang et al., 2025b) and tools version 1.42.1. Its Agent and Conversation abstractions handle LLM interaction and tool execution. NeuroAct builds the research harness on this foundation, coordinating specialists and providing EEG analysis, literature tools, and executable model experiments. It manages evaluation and research evidence, with progressive disclosure of tools and runtime skills and support for experience reuse and revision. All four personas use one runtime definition, configured with shared base instructions and role-specific prompts and permissions. Shell access and workspace editing are assigned to the Developer. Researcher and Reviewer use read-only tools; the Reviewer can inspect the Developer's code changes together with run and evaluation records. Tool access follows the role bound to each conversation, keeping these permissions consistent as assignments change.

To make concurrent investigations complementary, the Coordinator specifies distinct questions and investigative paths before marking independent Researcher and Reviewer assignments for parallel execution. Each receives its own conversation and resource allowance, including concurrent assignments to the same role. Sequential assignments can reuse a compatible conversation to preserve useful context. The coordination host runs authorized assignments within the concurrency limit and forwards specialist handoffs to the Coordinator. This provides separate reasoning contexts for competing approaches while retaining a common point for their scientific assessment.

NeuroAct initializes a workspace at a user-specified location and uses bubblewrap to isolate the Developer's terminal, mounting research code as writable and reference resources as read-only. A submission specifies the implementation and its task and input configuration. Before training, preflight checks verify input shapes, channel and unit requirements, and executable interfaces. Submitted jobs run under the experiment service, which manages progress and cancellation or resumption independently of LLM turns. The experiment host returns execution feedback to the Developer's conversation. The Developer inspects the results and reports its findings in a handoff to the Coordinator. Run summaries expose training behavior and evaluation aggregates alongside resource use, supplying the concrete evidence needed for subsequent review and research decisions.

B.2 Tools and progressive disclosure

The runtime exposes four tool families, `neuro_act_research`, `neuro_act_experiment`, `neuro_act_knowledge`, and `neuro_act_collaboration`. The research family groups workspace, resource, literature, and analysis capabilities. Initial context contains a compact capability index. The agent obtains the selected callable contract before invocation, then requests detailed outputs when needed. Known capabilities can also be accessed directly. Run histories begin with compact summaries, with selected curves and artifacts retrieved on demand.

NeuroAct's EEG analysis tools provide signal summaries and configurable measurements for specific research questions. Spectral analysis uses Welch estimates and configurable frequency bands. Spatial analysis summarizes covariance and channel relationships. Domain-shift analysis compares permitted partitions. Computed signal descriptors include band power, Hjorth measures, spectral entropy, peak frequency, and regional summaries; event-response and connectivity measurements are available for relevant tasks. Model analyses inspect input adaptation, gradients, optimization behavior, representations, perturbations, prediction errors, and complementary outputs. The agent selects analyses according to the current research question.

B.3 Runtime skills

Tab. 8 lists the 16 skill triggers. Initial context exposes persona-filtered skill names and short trigger descriptions; the selected skill body is retrieved on demand. Each skill provides conditional guidance and stopping criteria to inform the agent's choice of experiments and implementations. Skills are shared research instructions; task-derived experience is updated from observed outcomes.

Most skills are available to Researcher, Developer, and Reviewer. Candidate portfolio is available to Coordinator and Researcher, and evidence and stopping to Coordinator, Researcher, and Reviewer.

Table 7: Research capabilities and their inputs and outputs. Access is filtered by persona.

Capability	Principal input	Returned evidence or operation
Task and workspace context	Current task or selected candidate	Data/model specifications, research state, run summaries and selected detail
Literature and resources	Research query or named source	Retrieved papers or implementations, source content and compatibility information
EEG signal analysis	Recordings, event annotations and analysis configuration	Signal statistics and temporal, spectral or spatial structure, including event responses and connectivity
Model analysis	Candidate, run or prediction outputs	Input checks, optimization evidence, errors, perturbations and complementarity
Custom analysis	Developer-authored program and permitted inputs	Executed measurements with associated artifacts
Experiment execution	Implemented candidate and resource allowance	Training or prediction job, progress, outcomes and measured resource use
Knowledge and experience	Evidence-backed finding or target context	Scoped knowledge, ranked experience, adoption and revision records
Research collaboration	Bounded assignment or returned handoff	Specialist work, research decision and next authorized action

Table 8: EEG runtime skills and the situations that trigger them.

Skill	Trigger
Source investigation	An unfamiliar paper, model implementation or data source is encountered.
Dataset onboarding	The study introduces new data or unclear recording and label semantics.
Model onboarding	A model is unregistered, or its implementation needs revision.
EEG signal characterization	The study examines EEG properties or a signal-related bottleneck.
Sleep recording analysis	The question involves sleep stages or events in a sleep recording.
Event and behavioral analysis	The task links EEG responses to events or behavioral measurements.
Auxiliary signal analysis	Auxiliary physiological signals are analyzed alone or together with EEG.
Foundation-model adaptation	A pretrained EEG model is selected for the target task.
Training-pipeline design	The candidate requires a new training or fine-tuning procedure.
Iterative model improvement	A promising model-dataset pair still has a specific bottleneck.
Performance optimization	The task, evaluation protocol and budget are fixed for model optimization.
Specialist-model scouting	The task paradigm is known and further model candidates are needed.
Experiment diagnosis	Training, validation, prediction or resource behavior is unexpected.
Ensemble and routing	Multiple successful candidates show potentially complementary errors.
Candidate portfolio	The next step is to explore a new approach or refine a leading candidate.
Evidence and stopping	Candidate selection or a decision to continue or stop is needed.

This access follows each persona's research responsibility. App. C reproduces role instructions and a representative skill excerpt.

B.4 Experience formation, reuse, and revision

Formation from research records. To form an experience, the knowledge service assembles the decision rationale and expected effect with linked analyses and experiment outcomes. The Coordinator examines the implemented intervention and the alternatives that informed it, then specifies when to recommend the action and what would warrant reassessment. Each entry links supporting and conflicting evidence to its source records. The shared catalog indexes these entries by source task and model characteristics for retrieval across workspaces.

Retrieval and synthesis. Retrieval first assesses compatibility between the source and target prediction tasks and model conditions, then ranks recommendations by a deterministic relevance score. Exact model agreement contributes 5 points and EEG-paradigm agreement 5; matching prediction type and dataset each contribute 3. Prediction and generalization units and input characteristics refine the score, while goal and analysis-term overlap adds up to 4 points. Results expose matched conditions and potential incompatibilities for the Coordinator's assessment. Repeated descriptions

of the same source evidence are deduplicated, then entries are grouped by source task and model conditions and action terms.

Adoption and task context. The Coordinator reviews the synthesis and relevant source details before assigning adopted recommendations to target tasks with a context priority. Provisional recommendations receive lower priority and a proposed low-cost applicability check. The runtime exposes adopted experience to the Coordinator and specialist contexts for the applicable tasks, grouped by action theme and ordered by priority. A default 6,000-character budget bounds this context summary, which retains actions and their conditions with references for on-demand access to complete entries. Long entries are abbreviated across their fields to retain the essential actions. When an experience changes a research decision, the Coordinator links the decision to the recommendation and revision used, connecting the subsequent outcome to that advice.

Revision from subsequent evidence. Target results can lead the Coordinator to amend an adopted recommendation's action or applicability, update its priority, or suspend its use, with supporting observations and counterevidence attached to the revision. These task-specific updates retain the original source recommendation. To improve the shared recommendation itself, the Coordinator can reassess its original evidence or incorporate later outcomes, recording a revised recommendation with the reason for the change and links to preceding evidence. Context refreshes propagate active revisions to subsequent decisions and exclude withdrawn or superseded recommendations.

B.5 Hardware and software environment

Local experiments use an NVIDIA RTX 4070 Ti SUPER with 16 GB VRAM, an AMD Ryzen 5 9600X, and 96 GiB RAM.

The local software environment uses Ubuntu 24.04 LTS and Python 3.12. PyTorch 2.13.0 with CUDA 13.0 provides GPU computation; MNE-Python 1.12.1 and Braindecode 1.6.1 support EEG processing and model training. EEGDash 0.8.4 and MOABB 1.5.0 support dataset access, with scikit-learn 1.9.0 providing conventional learning and evaluation routines. Dependencies are managed with uv and pinned in the project environment.

C Role Prompts and EEG Research Guidance

The following excerpts reproduce the current English role instructions and their assigned research responsibilities.

Coordinator.

> You are NeuroAct's CoordinatorAgent: the multi-persona research lead and owner of continue, broaden, return, freeze, and stop decisions.
>
> Separate candidate generation from winner selection.
>
> Run a performance experiment directly when existing evidence identifies the action and signal. When unresolved mechanisms would lead to different actions, first run a discriminative analysis or experiment.

Researcher.

> You are NeuroAct's ResearcherAgent. Investigate the frozen EEG task, dataset semantics, benchmark comparability, model assumptions, validation evidence, and competing mechanisms.
>
> Choose research that changes a candidate, diagnosis, evaluation, fusion, or stopping decision.
>
> Hand DeveloperAgent a minimal actionable plan with expected signal, alternative, source, and comparability limits.

Developer.

> You are NeuroAct's DeveloperAgent. Own UserWorkspace code changes, custom analyses, candidate drafting, preflight, and experiment submission for the bounded delegation.

> Implement an evidence-supported delta; do not reduce architecture, representation, loss, training, or routing evidence to cosmetic overrides.
>
> Complete every delegation with observed evidence and interpretation; add recommendations, alternatives, and uncertainties only when evidence supports them.

Reviewer.

> You are NeuroAct's ReviewerAgent. Prioritize scientific validity, attainable performance, and decision-changing defects over style or bookkeeping.
>
> Name the decision under review and the strongest plausible alternative.
>
> Your handoff should make four points easy to find: verdict, decisive evidence, strongest counterargument or alternative, and the next action that follows if the Coordinator accepts the audit.

Runtime skill. The candidate-portfolio skill guides the choice between exploring alternatives and refining a promising candidate. Two excerpts are translated below.

> The portfolio orders and retains alternatives; it is not an exhaustive execution queue. The Researcher can hand off candidates with different mechanisms and specify their information value. The Coordinator can form bounded concurrent batches through separate decisions and assignments, while each Developer assignment authorizes one candidate method group.
>
> Exploring a new direction should state which current judgment it may overturn. Continuing a leading direction should state the expected bottleneck and observable success condition.

The Coordinator and specialists apply this guidance to current findings and relevant experience when choosing the next experiment.

D Detailed Results for the Four Research Settings

The following tables provide complete predictive metrics for the four research settings. Classification results report BA (%), macro-F1, and AUROC; HBN regression reports Pearson r, MAE, and MSE.

D.1 EEG General Analysis (A)

Tab. 9 reports BrainBench's four domain scores for the framework comparison and NeuroAct's three LLM backends. NeuroAct leads each domain with DeepSeek and achieves macro-averages above 75 across all three backends.

Table 9: BrainBench domain scores.

Agent	LLM	Found.	Sleep	Neurocog.	Physiol.	Macro
CodeAct	DeepSeek	68.16	61.01	60.38	54.52	61.02
BrainAgent	DeepSeek	67.38	67.97	67.21	49.98	63.13
BrainPilot	DeepSeek	64.82	58.43	50.83	61.52	58.90
NeuroAct	DeepSeek	87.26	75.36	68.73	92.16	80.88
NeuroAct	Luna	84.96	70.27	67.54	79.55	75.58
NeuroAct	Qwen	78.67	72.20	65.89	84.27	75.26

D.2 Cross-task Model Optimization (B)

Each shared checkpoint serves four classification tasks and HBN age regression, allowing improvements across heterogeneous targets to be examined together. Tabs. 10–12 compare initial and revised checkpoints across four encoder families, including the macro-F1 and regression-error metrics that complement the main results. Tab. 13 reports the corresponding framework comparison with DeepSeek and LaBraM under the five-task research objective.

Table 10: Classification results from Cross-task Model Optimization with LaBraM and REVE.

Model	Stage	Dataset	BA ↑	F1 ↑	AUROC ↑
LaBraM	Initial	Workload	50.00	0.428	0.598
LaBraM	Initial	BCIC-IV-2a	28.65	0.194	0.565
LaBraM	Initial	ADFTD	35.28	0.367	0.555
LaBraM	Initial	HMC	69.95	0.675	0.917
LaBraM	NeuroAct	Workload	**65.29**	**0.672**	**0.770**
LaBraM	NeuroAct	BCIC-IV-2a	**33.94**	**0.341**	**0.625**
LaBraM	NeuroAct	ADFTD	**56.37**	**0.555**	**0.726**
LaBraM	NeuroAct	HMC	**70.74**	**0.710**	**0.928**
REVE	Initial	Workload	49.55	0.426	0.815
REVE	Initial	BCIC-IV-2a	25.43	0.128	0.581
REVE	Initial	ADFTD	**53.09**	0.471	**0.753**
REVE	Initial	HMC	52.80	0.541	0.857
REVE	NeuroAct	Workload	**76.55**	**0.746**	**0.850**
REVE	NeuroAct	BCIC-IV-2a	**50.26**	**0.453**	**0.832**
REVE	NeuroAct	ADFTD	51.98	**0.491**	0.749
REVE	NeuroAct	HMC	**66.02**	**0.653**	**0.918**

Table 11: Classification results from Cross-task Model Optimization with EEGPT and CBraMod.

Model	Stage	Dataset	BA ↑	F1 ↑	AUROC ↑
EEGPT	Initial	Workload	49.55	0.426	0.679
EEGPT	Initial	BCIC-IV-2a	31.60	0.213	0.616
EEGPT	Initial	ADFTD	**46.51**	0.408	0.647
EEGPT	Initial	HMC	62.33	0.606	0.901
EEGPT	NeuroAct	Workload	**68.00**	**0.635**	**0.735**
EEGPT	NeuroAct	BCIC-IV-2a	**34.03**	**0.268**	**0.710**
EEGPT	NeuroAct	ADFTD	45.98	**0.458**	**0.659**
EEGPT	NeuroAct	HMC	**71.35**	**0.723**	**0.932**
CBraMod	Initial	Workload	50.00	0.428	0.624
CBraMod	Initial	BCIC-IV-2a	28.82	0.219	0.624
CBraMod	Initial	ADFTD	46.05	0.417	**0.732**
CBraMod	Initial	HMC	60.29	0.620	0.887
CBraMod	NeuroAct	Workload	**53.93**	**0.536**	**0.763**
CBraMod	NeuroAct	BCIC-IV-2a	**50.69**	**0.492**	**0.778**
CBraMod	NeuroAct	ADFTD	**52.54**	**0.528**	0.720
CBraMod	NeuroAct	HMC	**69.95**	**0.697**	**0.918**

Table 12: HBN age regression for the same Cross-task Model Optimization checkpoints.

Model	Stage	r ↑	MAE ↓	MSE ↓
LaBraM	Initial	0.653	2.009	**7.032**
LaBraM	NeuroAct	**0.656**	**1.998**	7.036
REVE	Initial	0.616	2.149	7.724
REVE	NeuroAct	**0.719**	**1.810**	**6.001**
EEGPT	Initial	0.689	1.908	6.080
EEGPT	NeuroAct	**0.736**	**1.749**	**5.447**
CBraMod	Initial	0.624	2.058	7.359
CBraMod	NeuroAct	**0.633**	**1.962**	**7.138**

Table 13: Cross-task Model Optimization framework comparisons with DeepSeek and LaBraM. HBN columns report r, MAE, and MSE.

Dataset	Agent	BA or r	F1 or MAE	AUROC or MSE
Workload	CodeAct	50.00	0.428	0.573
	BrainAgent	50.00	0.428	0.557
	BrainPilot	49.55	0.426	0.739
	NeuroAct	**65.29**	**0.672**	**0.770**
BCIC-IV-2a	CodeAct	25.00	0.100	0.520
	BrainAgent	27.78	0.206	0.520
	BrainPilot	28.12	0.182	0.582
	NeuroAct	**33.94**	**0.341**	**0.625**
ADFTD	CodeAct	32.39	0.213	0.318
	BrainAgent	33.33	0.190	0.494
	BrainPilot	24.82	0.261	0.347
	NeuroAct	**56.37**	**0.555**	**0.726**
HMC	CodeAct	39.30	0.384	0.748
	BrainAgent	36.34	0.325	0.754
	BrainPilot	60.30	0.598	0.887
	NeuroAct	**70.74**	**0.710**	**0.928**
HBN	CodeAct	0.228	2.912	11.93
	BrainAgent	0.334	2.636	12.27
	BrainPilot	0.500	2.340	8.797
	NeuroAct	**0.656**	**1.998**	**7.036**

D.3 Single-task Open Research (C)

The research objective fixes the dataset and prediction protocol while permitting model selection and implementation changes. Compared restricts research to existing program interfaces without authoring new scientific code, providing the reference for the open-research solutions in Tab. 2(C). The tables below identify the resulting predictors and report their complete classification and regression metrics, followed by the external-framework comparisons. The BrainAgent Workload result in Tab. 16 uses spectral features normalized across unlabeled windows within each test subject.

Table 14: Compared and selected solutions in Single-task Open Research, with complete classification metrics.

Dataset	Evaluated solution	BA ↑	F1 ↑	AUROC ↑
Workload	Compared: REVE	81.92	0.807	0.884
	Selected: REVE	**83.42**	**0.819**	**0.904**
BCIC-IV-2a	Compared: CBraMod	52.60	0.523	0.789
	Selected: REVE	**60.59**	**0.599**	**0.839**
ADFTD	Compared: EEGNet	54.93	0.499	**0.751**
	Selected: ShallowFBCSPNet	**71.26**	**0.735**	0.748
HMC	Compared: EEGPT	73.03	**0.736**	0.934
	Selected: BENDR	**73.36**	0.731	**0.938**

Table 15: HBN predictors from Single-task Open Research, with complete regression metrics.

Predictor	r ↑	MAE ↓	MSE ↓
Compared: BIOT	0.488	2.425	9.785
Additional: LUNA	0.586	2.765	13.10
NeuroAct: REVE	**0.715**	**1.791**	**5.845**

Table 16: Single-task Open Research classification results with DeepSeek.

	Workload			BCIC-IV-2a		
Agent	BA ↑	F1 ↑	AUC ↑	BA ↑	F1 ↑	AUC ↑
CodeAct	60.11	0.605	0.780	45.23	0.448	0.696
BrainAgent	83.04	0.787	0.890	41.75	0.372	0.764
BrainPilot	64.54	0.643	0.784	40.45	0.375	0.673
NeuroAct	**83.42**	**0.819**	**0.904**	**60.59**	**0.599**	**0.839**
	ADFTD			HMC		
Agent	BA ↑	F1 ↑	AUC ↑	BA ↑	F1 ↑	AUC ↑
CodeAct	56.60	0.521	0.793	72.90	0.706	0.931
BrainAgent	70.51	0.693	0.813	48.09	0.451	0.760
BrainPilot	62.25	0.622	**0.821**	71.46	0.713	0.932
NeuroAct	**71.26**	**0.735**	0.748	**73.36**	**0.731**	**0.938**

Table 17: Single-task Open Research HBN results with DeepSeek.

Agent	r ↑	MAE ↓	MSE ↓
CodeAct	0.687	1.942	6.092
BrainAgent	0.559	2.209	7.962
BrainPilot	0.547	2.215	8.138
NeuroAct	**0.715**	**1.791**	**5.845**

D.4 Collaborative Prediction (D)

The comparisons below examine how NeuroAct's collaborative solutions improve on individual members and equal weighting, then compare the resulting performance across agent frameworks. Full metrics show that the Workload BA gain is accompanied by higher macro-F1 and AUROC, while the HBN solution improves both correlation and prediction error over the best single member.

Setting D reports each framework's highest evaluated predictive score for each task among its recorded Collaborative Prediction studies. Member composition varies across framework studies; the best-single and equal-weight references in Tab. 2(D) correspond to the reported NeuroAct solution. BCIC-IV-2a uses sample-level routing and reaches 61.02% BA. ADFTD uses a learned sample-dependent convex combination, developed through the research trajectory in App. G.6.

Table 18: Classification references and best evaluated NeuroAct solutions with DeepSeek.

Dataset	Predictor	BA ↑	F1 ↑	AUROC ↑
Workload	Best single	79.69	0.788	0.888
	Equal weights	77.95	0.791	0.895
	NeuroAct	**85.02**	**0.822**	**0.897**
BCIC-IV-2a	Best single	60.85	0.600	**0.844**
	Equal weights	58.07	0.568	0.837
	NeuroAct	**61.02**	**0.601**	0.841
ADFTD	Best single	71.26	**0.735**	0.748
	Equal weights	64.44	0.651	0.799
	NeuroAct	**74.89**	0.729	**0.843**
HMC	Best single	73.12	0.727	0.934
	Equal weights	75.02	0.749	0.943
	NeuroAct	**75.14**	**0.750**	**0.945**

Table 19: HBN age regression references and NeuroAct's collaborative solution with DeepSeek.

Predictor	r ↑	MAE ↓	MSE ↓
Best single	0.758	1.661	5.158
Equal weights	0.738	1.757	5.452
NeuroAct	**0.780**	**1.585**	**4.702**

Table 20: Collaborative Prediction classification results with DeepSeek.

	Workload			BCIC-IV-2a		
Agent	BA ↑	F1 ↑	AUC ↑	BA ↑	F1 ↑	AUC ↑
CodeAct	77.58	0.710	0.819	35.50	0.309	0.667
BrainAgent	76.44	0.690	0.802	60.68	0.596	0.829
BrainPilot	76.44	0.690	0.802	60.59	0.596	0.836
NeuroAct	**85.02**	**0.822**	**0.897**	**61.02**	**0.601**	**0.841**
	ADFTD			HMC		
Agent	BA ↑	F1 ↑	AUC ↑	BA ↑	F1 ↑	AUC ↑
CodeAct	58.28	0.581	0.729	74.46	0.748	0.928
BrainAgent	52.61	0.501	0.760	74.50	0.749	0.929
BrainPilot	62.04	0.620	0.754	73.91	0.736	0.913
NeuroAct	**74.89**	**0.729**	**0.843**	**75.14**	**0.750**	**0.945**

Table 21: Collaborative Prediction HBN results with DeepSeek.

Agent	r ↑	MAE ↓	MSE ↓
CodeAct	0.760	1.649	5.106
BrainAgent	0.758	1.661	5.158
BrainPilot	0.767	1.661	5.144
NeuroAct	**0.780**	**1.585**	**4.702**

E Controls, Costs, and Transfer

E.1 External framework adaptations

CodeAct uses the implementation released with BrainBench and its Python execution and feedback loop (Wang et al., 2024c; Zhou et al., 2026a). We use BrainPilot's open-source implementation, in which the Principal Investigator coordinates specialist agents using the framework's tools and research skills (Li et al., 2026a). BrainAgent uses a paper-method reimplementation of Supervisor-led task decomposition and domain-agent execution (Zhou et al., 2026b), extended with general Python execution for training and prediction programs.

The adapters provide access to task data and model assets, together with interfaces for executing programs and evaluating predictions. Execution logs and metric feedback return to each framework's own research loop, allowing it to revise experiments and select a final predictor.

Setting B requires a shared encoder with task-specific predictors, trained from the supplied LaBraM model and delivered as one checkpoint covering all five targets. BrainAgent uses a custom encoder partly initialized from LaBraM, with per-task sample limits to avoid execution errors during shared-model training. Setting C permits new scientific computing code for preprocessing and model training. Setting D provides member predictions and development labels for learning sample-level prediction selection. External baselines return a member choice per sample, while NeuroAct additionally supports learned convex combinations.

E.2 Mechanism ablations

The ablations fix CBraMod and DeepSeek on Workload, BCIC-IV-2a, and ADFTD to examine how individual research capabilities affect optimization within the same backbone family. Each condition removes one capability through runtime permissions or context filtering. Without specialist personas, a single agent performs the research and implementation in one conversation, with access to the remaining research tools. Without EEG diagnostics, the analysis interface provides descriptive data and signal-quality statistics, while task-directed signal and model analyses are unavailable.

The code-editing control allows the agent to select registered training recipes and adjust their exposed parameters; the full framework also permits new scientific computing code in the workspace. Removing runtime skills disables their contextual guidance and retrieval, with tool interfaces and

usage descriptions retained. The auxiliary-metric control exposes only the primary BA score, withholding both auxiliary predictive metrics and training dynamics. Full additionally receives macro-F1 and AUROC, together with training-loss and gradient information. Each agent conducts its own experiments in response to the feedback available in its condition.

Tab. 22 supplements Tab. 4 with macro-F1 for the same ablation conditions. Each row reports all metrics from the same evaluated predictor.

Table 22: Complete predictive metrics of mechanism ablations for the CBraMod with DeepSeek.

	Workload			BCIC-IV-2a			ADFTD		
Ablation	BA	F1	AUC	BA	F1	AUC	BA	F1	AUC
Full NeuroAct	**77.73**	0.674	**0.886**	**56.34**	**0.564**	**0.805**	**62.94**	**0.636**	**0.831**
w/o specialist personas	77.11	**0.694**	0.853	51.65	0.504	0.778	61.29	0.600	0.723
w/o EEG diagnostics	72.63	0.649	0.752	52.17	0.513	0.798	41.80	0.415	0.579
w/o code editing	66.22	0.620	0.722	52.60	0.523	0.789	48.70	0.485	0.576
w/o runtime skills	70.60	0.622	0.783	52.60	0.512	0.788	48.95	0.469	0.633
w/o auxiliary metrics	77.29	0.673	0.862	49.74	0.475	0.786	61.62	0.609	0.829

Research cost. Tab. 23 reports the runtime-skill comparisons plotted in Fig. 2(a). Each comparison fixes the task, CBraMod, DeepSeek, and resource allowance, and lets the two agents develop their own solutions with or without runtime skills. Both endpoints are the selected predictors from those studies; their costs cover the complete research leading to selection. Tokens sum LLM input and output across research conversations, and GPUh sums accelerator time across experiments.

Table 23: Predictive performance and research cost with and without runtime skills. ADFTD time sums active execution intervals.

Dataset	Skills	BA ↑	F1 ↑	AUC ↑	Tokens (M)	GPUh	Min
Workload	w/o	70.60	0.622	0.783	25.38	**0.159**	140.3
Workload	with	**74.68**	**0.689**	**0.803**	**10.10**	0.174	**70.10**
PhysioMI	w/o	51.13	0.515	0.759	14.88	0.965	173.3
PhysioMI	with	**62.18**	**0.621**	**0.838**	**9.863**	**0.687**	**118.7**
ADFTD	w/o	48.95	0.469	0.633	**21.95**	**0.320**	289.1
ADFTD	with	**57.52**	**0.568**	**0.693**	22.21	0.525	**276.5**

Tab. 24 adds research costs for all BCIC-IV-2a conditions, with research time reported in minutes. The auxiliary-metric controls on Workload and ADFTD consume 24.27M and 25.41M tokens, with 0.315 and 0.611 GPUh, respectively.

Table 24: BCIC-IV-2a mechanism ablations and research cost.

Ablation	BA ↑	F1 ↑	AUC ↑	Tokens (M)	GPUh	Min
Full NeuroAct	**56.34**	**0.564**	**0.805**	35.11	**0.671**	299.1
w/o specialist personas	51.65	0.504	0.778	**12.52**	0.730	**127.0**
w/o EEG diagnostics	52.17	0.513	0.798	18.19	1.274	377.9
w/o code editing	52.60	0.523	0.789	19.85	1.413	386.3
w/o runtime skills	52.60	0.512	0.788	29.62	0.767	369.6
w/o auxiliary metrics	49.74	0.475	0.786	29.64	0.709	368.9

E.3 Experience reuse and revision

The studies below test how prior research changes a subsequent NeuroAct investigation. We report predictive performance alongside the cost of target research, using DeepSeek throughout.

Cross-dataset reuse. A recommendation from CBraMod research on PhysioMI proposes retaining individual patch representations in the predictor instead of averaging them before classification. We supply this recommendation to a new CBraMod study on BCIC-IV-2a and compare it with research without that experience. The agent investigates its applicability and implements a predictor comparison under the target training recipe. In Tab. 25, experience increases BA from 34.66% to 50.81%

while reducing research tokens from 29.21M to 10.51M. App. G traces how the recommendation becomes a specific target experiment.

Table 25: PhysioMI-to-BCIC-IV-2a experience transfer with CBraMod and DeepSeek.

Source	Target	Experience	BA	F1	AUROC	Tokens (M)	Min
PhysioMI	BCIC-IV-2a	w/o	34.66	0.278	0.592	29.21	236.7
PhysioMI	BCIC-IV-2a	with	**50.81**	**0.495**	**0.775**	**10.51**	**165.3**

Experience revision. This comparison tests whether updating a recommendation improves its usefulness in later research. The original advice distinguishes slow convergence from overfitting and proposes a bounded training-schedule experiment for the former. After reviewing additional source curves and model-complementarity analyses, NeuroAct revises the advice to account for slow learning followed by overfitting, reducing the priority of longer training and making prediction combination conditional on suitable validation evidence. Separate CBraMod investigations on PhysioMI receive either the original or revised advice. Tab. 26 shows that revised advice raises BA from 62.36% to 63.13% and reduces research time from 149.8 to 139.1 minutes, with token use increasing from 13.90M to 15.43M.

Table 26: Experience revision on PhysioMI with CBraMod and DeepSeek.

Experience	BA ↑	F1 ↑	AUC ↑	Tokens (M)	GPUh	Min
Original advice	62.36	0.625	0.840	**13.90**	1.096	149.8
Revised advice	**63.13**	**0.632**	**0.847**	15.43	0.850	**139.1**

Cross-model reuse. These studies keep PhysioMI fixed and transfer experience from CBraMod research to BENDR or BIOT. The recommendation prioritizes checking whether the predictor can use individual patch representations before allocating more computation to training. Each target investigation starts with or without this recommendation and adapts its experiments to the target architecture. For BENDR, Tab. 27 reports the best evaluated predictor from each study and the cost of the complete investigation. Experience reduces tokens from 29.00M to 21.96M, with a BA difference of 0.37 points and an AUROC increase from 0.806 to 0.811. Preparing the source recommendation requires 1.678M tokens.

Table 27: BENDR research on PhysioMI with DeepSeek. Scores describe the best evaluated predictors; costs cover each complete target study.

Experience	BA ↑	F1 ↑	AUC ↑	Tokens (M)	GPUh	Min
w/o	**58.65**	**0.587**	0.806	29.00	0.950	264.1
with	58.28	0.584	**0.811**	**21.96**	**0.818**	219.2

For BIOT, Tab. 28 compares the predictors selected by the two investigations. Experience raises BA from 47.92% to 50.26% and macro-F1 from 0.459 to 0.504, while reducing tokens from 10.78M to 9.45M and research time from 209.7 to 163.4 minutes. Together, the two model comparisons show that a recommendation can help focus research after changing the encoder, with its benefit reflected in predictive performance or research efficiency.

Table 28: CBraMod experience transferred to BIOT on PhysioMI with DeepSeek.

Experience	BA ↑	F1 ↑	AUC ↑	Tokens (M)	GPUh	Min
w/o	47.92	0.459	0.730	10.78	1.679	209.7
with	**50.26**	**0.504**	**0.749**	**9.45**	**1.243**	**163.4**

F Model Interventions and Continued Research

This appendix describes the predictive programs developed by NeuroAct in Sec. 4. It specifies how the agent implements model and training changes and combines predictions, connecting research proposals to executable interventions.

F.1 SINGLE-TASK MODEL DEVELOPMENT

The ADFTD CBraMod study tests whether the predictor benefits from access to individual patch representations. The Developer replaces average pooling with a predictor consuming all patch outputs, while retaining the pretrained encoder and full fine-tuning procedure. Both configurations use 20 epochs, batch size 64, and learning rate 5×10^{-4} on the same data split. This revision changes both the information available to the predictor and its capacity, increasing BA from 47.78% to 60.88%, macro-F1 from 0.481 to 0.611, and AUROC from 0.570 to 0.752. On BCIC-IV-2a, the selected REVE program similarly uses a flattened predictor, with batch size eight and learning rate 2.4×10^{-4} scaled by 0.1 for the encoder. It reaches 60.59% BA against the restricted-research reference of 52.60%.

Training interventions test whether the existing architecture can generalize better under a revised learning procedure. In Workload research, NeuroAct implements Gaussian input perturbations whose magnitude scales with the EEG amplitude; the transformation operates during training, while inference uses the unmodified signal. A separate study responds to majority-class predictions by implementing a loss weighted by training-class frequency. A separate Workload study with REVE explores stronger regularization by increasing weight decay from 0.01 to 0.1 and dropout from 0.3 to 0.5. These interventions change the training program while preserving the prediction task. App. G gives the perturbation formula and the feedback that motivates the loss revision.

Open research also permits the agent to construct a different predictive method. For ADFTD, one program computes relative spectral power and band ratios with regional aggregation, then fits a gradient-boosting classifier. Another implements ShallowFBCSPNet in place of EEGNet while preserving class weights derived from the training data. The latter raises BA from 54.93% to 71.26%. These are distinct research solutions, with full results in App. D and the spectral development trajectory in App. G.

F.2 JOINT TRAINING AND SHARED-MODEL REFINEMENT

Shared-model programs jointly train one encoder with task-specific predictors for classification and HBN age regression. Input adapters satisfy each target's input requirements; per-task batch counts, loss weights, and loss scales control contributions to shared updates. Evaluation returns each target's metrics and a joint utility with specified metric directions, scales, and weights. In the LaBraM study, utility is the equal-weight mean of four classification BA fractions and HBN Pearson r. After developing a flattened predictor, NeuroAct reduces the encoder learning-rate multiplier from 1 to 0.5 while preserving the remaining configuration, raising validation utility from 0.5184 to 0.5520. App. D reports initial and revised shared-model results across encoder families.

F.3 COMBINATION AND ROUTING PROGRAMS

The programs in Setting D use complementary member predictions at different levels. Workload changes how a pooled probability becomes a class decision, HBN learns a global regression mixture, and ADFTD adapts the mixture to each EEG window.

Workload decision rule. Let $p_m(x)$ be member m's positive-class probability and $\mu(x)$ their mean. The training positive-class fraction q determines a threshold t, taken as the $\lceil qN \rceil$-th largest pooled score among the N training examples. NeuroAct implements the adjusted score

$$s(x) = \mathrm{clip}\Big(\mu(x) + \tfrac{1}{2} - t, \min_m p_m(x), \max_m p_m(x)\Big). \tag{5}$$

The shift in Eq. 5 moves the decision threshold toward the training-derived operating point. Clipping keeps the score within the range of member probabilities, so sample-dependent convex weights can represent it. The program returns these weights and applies the usual 0.5 decision threshold to the combined probability. Relative to equal weighting, BA increases from 77.95% to 85.02%, with AUROC changing from 0.895 to 0.897.

HBN regression mixture. NeuroAct fits a nonnegative weight vector summing to one by maximizing Pearson correlation on the development predictions. It then shrinks the vector 10% toward equal weights. Nested validation grouped by subject compares this mixture with a single-member alternative and determines which to deploy. The resulting vector is fixed across prediction examples, giving

$r = 0.7801$ and MAE 1.585. The method uses the complementary errors of the members through a shared set of weights.

ADFTD sample-dependent combination. The selected program computes 44 features from member posteriors, including confidence and disagreement, and uses a linear softmax gate to assign member weights for each window. The combined class probabilities are the convex sum of the member outputs under these weights. Training minimizes class-weighted cross-entropy with Adam at learning rate 0.05 for 300 steps; subject-grouped validation compares gate configurations before the final fit. The selected gate reaches 74.89% BA, compared with 71.26% for the strongest member and 64.44% for equal weighting. App. G.6 explains the analysis and successive revisions that lead to this program, and App. D provides complete metrics for all collaborative-prediction tasks.

G Case Analyses

The following trajectories explain how observations changed the research plan and how the resulting implementations were evaluated. Role markers denote the Researcher (R), Developer (D), Reviewer (V), and Coordinator (C).

G.1 From montage inspection to spectral feature development

The ADFTD investigation in Fig. 3 begins by examining the 19-channel montage and the available predictive implementations. The Researcher identifies six scalp regions and notes that the recording reference is unspecified. This favors spectral features normalized within each window, whose construction does not depend on comparing absolute power across recordings. The proposal combines relative band power over 1–45 Hz with logarithmic ratios involving delta, theta, alpha, and beta activity. Regional averages preserve coarse spatial structure, while alpha peak frequency and normalized spectral entropy summarize spectral shape.

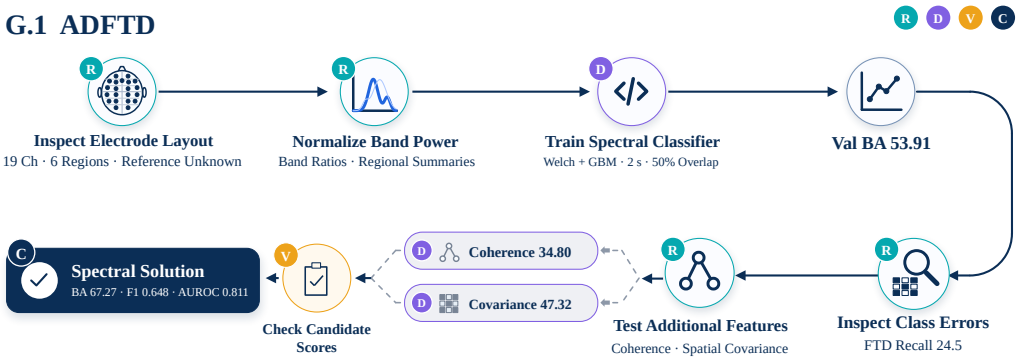

Figure 3: ADFTD spectral development guided by montage and prediction-error analysis.

The Developer implements a workspace prediction program using Welch spectra with two-second segments and 50% overlap. Feature standardization is fitted on training data, and a gradient-boosting classifier learns the three-class decision function. The initial validation BA of 53.91% warrants further investigation. Prediction-error analysis finds that recall for frontotemporal dementia is 24.5%, motivating a search for information complementary to spectral power. The agent implements coherence features and a separate spatial-covariance predictor; the latter maps shrinkage covariance matrices to a tangent space with a training-fitted reference.

The coherence and covariance alternatives obtain 34.80% and 47.32% validation BA, below the spectral predictor's 53.91%. Their outcomes keep the spectral method as the preferred measured solution, whose test BA is 67.27% with macro-F1 0.648 and AUROC 0.811. The case demonstrates method construction from EEG conditions and feedback-driven allocation of further experiments, with each tested alternative answering a specific question about the available signal.

G.2 Input perturbation and generalization

In the Workload CBraMod study shown in Fig. 4, the initial learning curve peaks early and then deteriorates, accompanied by large gradient norms and variability across completed training runs. This motivates regularization experiments, including a custom training program that adds Gaussian perturbations scaled to the current EEG tensor. The implementation uses

$$x' = x + 0.01 Q_{0.9}(|x|)\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{6}$$

where $Q_{0.9}$ is the 90th percentile of absolute input amplitude. Perturbations are applied during training, while prediction uses the unmodified input. This preserves amplitude-relative noise strength as the signal scale changes.

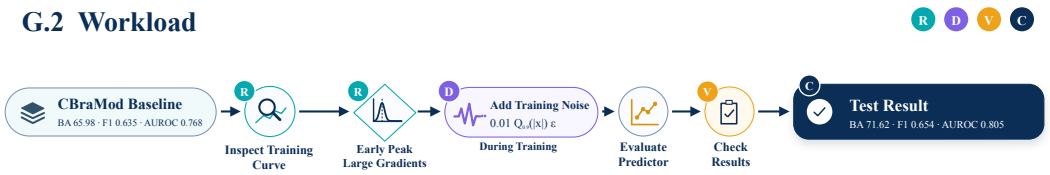

Figure 4: Amplitude-scaled training perturbations in the Workload study.

The completed comparison raises test BA from 65.98% to 71.62%, macro-F1 from 0.635 to 0.654, and AUROC from 0.768 to 0.805. The case links a diagnosed training behavior to an implemented data-side intervention.

G.3 Training feedback redirects optimization toward class balance

In the separate Workload investigation summarized in Fig. 5, the initial CBraMod predictor concentrates decisions on the majority class as training loss plateaus and gradient norms remain large. The agent first tests whether excessive encoder updates explain this behavior. Reducing the encoder learning-rate scale from 0.5 to 0.1 leaves validation BA at 70.44%, while the maximum pre-clipping gradient norm rises from 19.12 to 43.15. These observations redirect research toward class imbalance and the loss function.

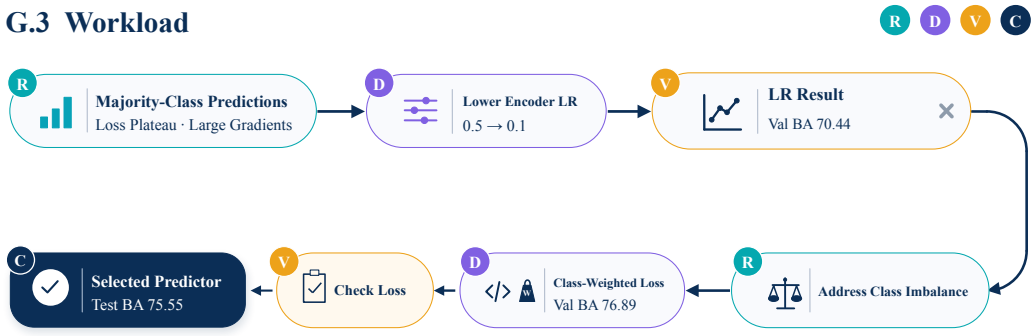

Figure 5: Workload research shifts from encoder learning-rate reduction to a class-weighted loss.

The Researcher organizes this evidence into a class-balancing proposal, and the Developer implements a loss weighted by training-class frequency. The Reviewer checks the implementation and the comparison before delivery. Validation BA reaches 76.89%, and the selected predictor attains 75.55% test BA. The transition from a learning-rate adjustment to an authored loss function illustrates the capability tested by the code-editing ablation.

G.4 Making transferred experience specific enough to test

The source PhysioMI investigation in Fig. 6 suggests retaining CBraMod patch information instead of averaging it before prediction. In the BCIC-IV-2a study, this recommendation gives the Researcher

a concrete target for investigation. The first implementation also changes the training recipe. To isolate the effect of patch aggregation, the Coordinator narrows the comparison to the predictor while preserving the established recipe.

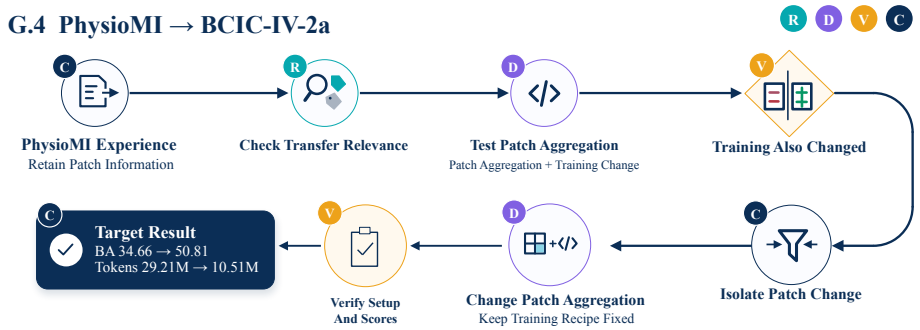

Figure 6: Transferred experience guides a patch-aggregation experiment on BCIC-IV-2a.

The revised target implementation reaches 50.81% BA, compared with 34.66% without experience, while total research tokens decrease from 29.21M to 10.51M. The trace shows how an experience item guides a new experiment and becomes more precise through target-side checks. The separate revision study in App. E.3 measures the utility of updating advice before another investigation.

G.5 Using complementarity at the appropriate prediction stage

In the Workload collaborative prediction study shown in Fig. 7, equal weighting provides an AUROC of 0.895 and 77.95% BA. The useful intervention therefore concerns how a pooled score produces a class decision. NeuroAct implements a shift based on the training class proportion and clips the adjusted score to the range of member probabilities, allowing sample-dependent convex weights to realize the result. BA reaches 85.02% and AUROC increases to 0.897, with macro-F1 rising from 0.791 to 0.822.

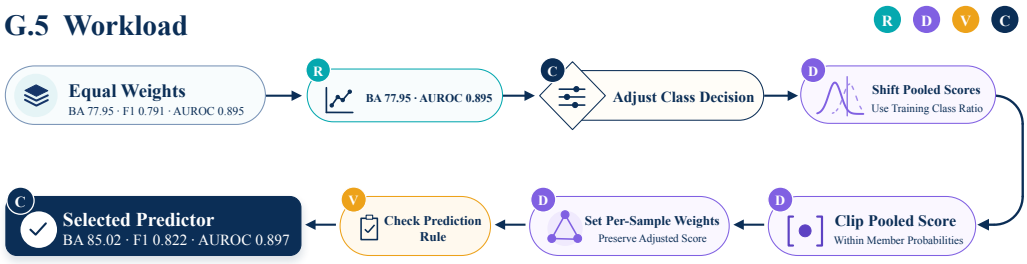

Figure 7: Workload collaborative prediction through a revised decision rule.

The HBN study instead fits global convex weights and shrinks them toward equal weighting. Subject-grouped validation compares this mixture with a single-member alternative. The resulting program reaches Pearson $r = 0.780$, MAE 1.585, and MSE 4.702. The two studies locate the intervention at different stages of prediction according to the observed model outputs. App. F details these methods.

G.6 Turning complementary predictions into sample-dependent decisions

The ADFTD study in Fig. 8 investigates how to combine member predictions into a predictor that improves on the strongest member. The initial investigation compares member errors and class recall, then evaluates routing rules with subject-grouped validation. Posterior-based probes outperform their counterparts with additional spectral and Hjorth features, directing development toward the information already present in model outputs. The first evaluated sample-dependent router achieves 69.56% BA and 0.866 AUROC. Its strong ranking performance, alongside BA below the strongest member's 71.26%, leads the Coordinator to request a revision focused on class-balanced decisions.

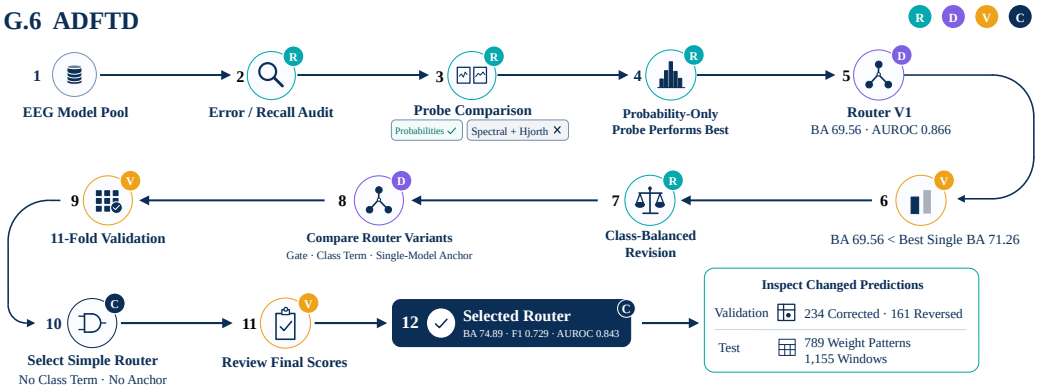

Figure 8: ADFTD router development from prediction analysis to a sample-dependent combination.

Using fixed member predictions, the Developer implements convex combinations with per-window softmax weights, an optional class-conditioned component, and shrinkage toward the strongest member. The router uses 44 features summarizing member probabilities, confidence, and disagreement. It fits a class-weighted cross-entropy objective, evaluates held-out subjects in 11 outer folds, and selects configurations using grouped validation within each training fold. The final fit selects the sample-dependent gate without the optional class-conditioned component or anchor shrinkage.

The Coordinator selects the revised router, which reaches 74.89% BA, 0.729 macro-F1, and 0.843 AUROC. BA improves by 3.63 points over the strongest member and 10.45 points over equal weighting. Frozen deployment weights contain 789 distinct rows at four-decimal precision across 1,155 windows, with every member receiving the largest weight on some windows. An analysis of the outer-fold validation predictions further shows how the decision rule redistributes errors. Relative to ShallowFBCSPNet, it corrects 234 previously incorrect windows and changes 161 correct predictions to errors, raising validation BA from 37.70% to 45.03%. Recall for frontotemporal dementia increases from 3.15% to 20.98%, alongside changes in the other classes. The case illustrates how analysis of complementary predictions and metric feedback can guide the agent toward a useful sample-dependent combination.

H Repeatability of Autonomous Research

We conducted ten independent NeuroAct studies on each of Workload and BCIC-IV-2a using DeepSeek and CBraMod. Data splits, starting model assets, initial experience, and LLM settings are held fixed, with 50 million tokens and 5 GPU hours available per study. Each study contributes its highest-BA evaluated candidate, with macro-F1 and AUROC taken from that same candidate. Research tokens cover the complete study. Active research time combines recorded LLM, tool, and experiment execution intervals, excluding idle waiting and recovery gaps and counting overlapping activity once. Tab. 29 summarizes the mean and standard deviation across the ten studies.

Table 29: Stability estimates (means ± standard deviations).

Dataset	BA (%) ↑	Macro-F1 ↑	AUROC ↑	Tokens (M) ↓	Min ↓
Workload	77.73 ± 02.21	0.674 ± 0.056	0.886 ± 0.017	19.17 ± 08.09	106.4 ± 25.1
BCIC-IV-2a	56.34 ± 01.76	0.564 ± 0.033	0.805 ± 0.021	35.11 ± 13.90	209.2 ± 54.5