LANTERN: TASK-ANCHORED NETWORK RESERVE FOR POST-STROKE APHASIA RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Regional vascular responses and measurement noise complicate task-fMRI prediction of language recovery after stroke. We introduce LANTERN, which measures task-conditioned language-network reserve and carries its measurement uncertainty into prediction. The reserve measure captures the diversity of regional response patterns across task conditions and is invariant to region-specific hemodynamic scaling. We relate its underlying response rank to independent task-to-region pathways in networks with hidden populations and feedback. To make measurement reliability part of learning, we jointly bound reserve over overlapping brain-region sets and train a monotone predictor against the worst-case error within those bounds. Independent calibration yields outcome prediction intervals. We evaluate current language impairment in the public Aphasia Recovery Cohort and six-month outcome in a two-center longitudinal cohort of 237 participants. On 84 participants from the held-out center, six-month prediction attains 5.64 WAB-AQ mean absolute error, compared with 7.60 for a brain network transformer. Nominal 90% intervals achieve 92.9% coverage with a mean width of 24.54 AQ points. Uncertainty-based retention reduces MAE to 4.28 in the more certain half of the held-out cohort. Controlled acquisition tests distinguish task diversity, measurement precision, and integration of delayed responses. Clinical adjustment, group-level repeatability, and 83.3% cross-run identification support the reserve's clinical relevance and individual specificity.

1 INTRODUCTION

Stroke-related aphasia can disrupt speech production, comprehension, reading, and participation in everyday life. Intensive language therapy can improve communication even in the chronic phase (Breitenstein et al., 2017), but patients with similar lesion burdens often follow different trajectories (Figure 1, left). Residual network organization, disconnected but surviving tissue, and the resolution of diaschisis contribute information beyond the infarct itself (Marebwa et al., 2017; Wawrzyniak et al., 2022). Recovery also differs across fluency, phonology, and semantic functions (Stefaniak et al., 2022). Language prognosis therefore calls for a measure of how the remaining network responds to different task demands, together with an assessment of how reliably that capacity is measured.

Task fMRI probes regional responses to known language demands, but its measurements combine neural activity with vascular effects (Logothetis et al., 2001). After stroke, regional differences in response amplitude and delay complicate comparisons of network function (Siegel et al., 2016; Braban et al., 2023). A change in vascular gain can scale a region's responses across task conditions even when its neural responses are preserved. Hidden neural populations and feedback connections create a further ambiguity: multiple network explanations can be compatible with the observed activity. The middle panel summarizes these vascular and network ambiguities, motivating a target calculated from task responses that retains a structural interpretation without selecting a unique causal graph.

The first idea (Figure 1, right) compares regional response patterns across task conditions and measures their diversity across regions. For example, vascular scaling that changes two integrated task responses from $(1, 2)$ to $(3, 6)$ leaves their relative pattern unchanged. For a fixed set of measured regions, complementary patterns provide more response diversity than repeated copies of the same

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

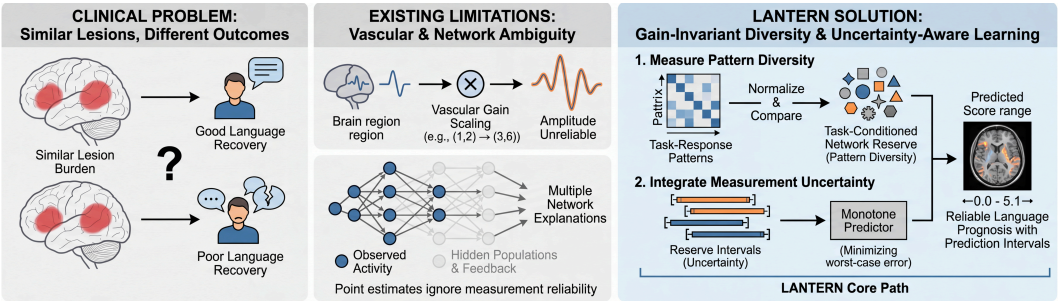


Figure 1: **Clinical motivation and the two core ideas of LANTERN.** *Left:* clinical motivation. *Middle:* vascular and network ambiguities in task-fMRI measurement. *Right:* gain-invariant reserve construction (step 1) and reserve-interval prediction learning with outcome calibration (step 2).

pattern. We call this continuous measure *task-conditioned network reserve*. LANTERN integrates finite impulse response (FIR) estimates over time, removes each region's overall response scale, and summarizes the resulting diversity over anatomical output sets. This yields a measure of response diversity and nonredundancy that remains informative when a small task battery makes integer rank too coarse to distinguish individuals.

The second idea brings measurement reliability into learning. The same reserve estimate can arise from a clearly measured task response or a weak, noisy one. We therefore retain an interval for each reserve feature. Relationships among overlapping brain-region sets constrain these intervals and sharpen their bounds. A monotone predictor learns from the resulting ranges by minimizing the larger error at its two prediction endpoints. These reserve intervals describe uncertainty in the measured features and directly constrain fitting. Independent calibration additionally accounts for outcome variability, yielding the language-score prediction intervals shown at the far right of Figure 1.

The paper makes two contributions. **(1) Measure pattern diversity with regional-gain invariance.** We capture the diversity of regional task-response patterns, prove invariance to regional hemodynamic scaling, and connect the underlying response rank to independent task-to-region pathways in networks with hidden populations and feedback. **(2) Integrate measurement uncertainty into prediction learning.** We jointly bound reserve over overlapping output sets and train against the worst-case prediction error within those bounds. Our code is available at https://anonymous.4open.science/r/LANTERN-EC8A/README.md.

2 Related Work

Aphasia network biomarkers. Residual white-matter organization, network controllability, and model-based disconnection measures predict language impairment or treatment gains (Marebwa et al., 2017; Wilmskoetter et al., 2022; Hildesheim et al., 2024; Low et al., 2024). REWIRED further uses lesion-informed cortical dynamics to obtain network features for chronic aphasia severity (Falconer et al., 2026). Task-fMRI prognosis additionally depends on reliable feature definition and protection against performance confounds (Song et al., 2023). These studies motivate a network-level target; our contribution is a gain-invariant measure of task-response diversity and a learning rule that carries its measurement uncertainty into prediction.

Latent causal structure and brain modeling. Linear non-Gaussian models use independent components to orient causal structure (Shimizu et al., 2006; Lacerda et al., 2012). Recent work studies path ranks, edge matchings, and equivalence in latent cyclic systems (Dai et al., 2026); the relevant duality is linked to gammoids and transversal matroids (Ingleton & Piff, 1973). We use these combinatorial objects to interpret a task-response operator. Dynamic causal modeling instead estimates explicit neural and hemodynamic mechanisms (Friston et al., 2003), and brain transformers learn flexible predictive graph representations (Kan et al., 2022). Our measured equivalence relation acts on task-response matrices modulo regional gains, making its estimand explicit even when the full graph remains unresolved.

Set-valued prediction. Log-determinant criteria and submodular information functions are established tools in experimental design (Krause et al., 2008; Bach, 2013). We use their set structure to couple confidence intervals for a gain-invariant response representation. Split conformal prediction calibrates residual uncertainty under exchangeability (Lei et al., 2018; Romano et al., 2019). Our methodological distinction lies in defining the gain-invariant reserve target and using its jointly constrained intervals during model fitting, before the output-calibration step.

3 TASK RESPONSES IN A LATENT RECURRENT BRAIN NETWORK

Let $V(t) \in \mathbb{R}^n$ contain observed neural populations X and unobserved populations L. For one participant and visit, consider

$$\dot{V}(t) = -(I - B)V(t) + CU(t) + E(t), \qquad Y_r(t) = (h_r * V_r)(t) + \xi_r(t), \quad r \in X. \tag{1}$$

The matrix B permits directed cycles and arbitrary interactions involving L; all eigenvalues of $I - B$ have positive real parts. The k columns of C encode the administered task conditions U, not assumed intervention sites inside the brain. Neural innovations may be independent and non-Gaussian as in a LiNG model. For task-response estimation, independence between the known design and the innovations—and, after nuisance adjustment, the observation noise ξ—is sufficient.

Let $F_{rj}(t)$ be the population BOLD impulse response from condition j to region r, and define its signed area $M_{rj} = \int_0^\infty F_{rj}(t)\, dt$. Stability and an integrable hemodynamic response yield

$$M = D_a T, \qquad T = [(I - B)^{-1}C]_{X,:}, \qquad D_a = \mathrm{diag}(a_r), \quad a_r = \int_0^\infty h_r(t)\, dt. \tag{2}$$

We assume $a_r \neq 0$ on the measured tissue. This permits unknown amplitude and latency, and does not equate a vascular delay with loss of a neural pathway. In practice, a finite FIR horizon approximates the integral; the remaining tail enters the response-error budget. The design matrix, including the lagged task regressors, must have sufficient rank to separate the presented conditions. A lesion-derived parcel mask records unavailable tissue explicitly, so it is never confused with a zero-valued healthy observation.

Clinical estimands. For the public cohort, y is the visit-level Western Aphasia Battery Aphasia Quotient (WAB-AQ) (Shewan & Kertesz, 1980). For the internal cohort, y is six-month WAB-AQ and the clinical context z includes baseline AQ, age, time after stroke, lesion burden, and recorded rehabilitation exposure. A higher predicted score expresses a language prognosis conditional on these inputs. The modeled reserve is anchored to the administered task battery and the specified regions, giving the estimand a concrete interpretation.

4 LANTERN

Figure 2 expands the two ideas in Figure 1 into four stages: estimate integrated task responses, construct gain-invariant reserve features, propagate their uncertainty into learning, and calibrate outcome predictions. The two contributions occupy stages 2 and 3; clinical context enters the decoder, and calibration uses participants disjoint from model fitting.

4.1 MEASURING RESPONSE DIVERSITY WITH REGIONAL-GAIN INVARIANCE

To compare regional task-response patterns independently of their overall scale, write the rth row of M as $m_r^\top$. Define a row projector and its output-set sum by

$$P_r(M) = \begin{cases} m_r m_r^\top / \|m_r\|_2^2, & m_r \neq 0, \\ 0, & m_r = 0, \end{cases} \qquad K_Z = \sum_{r \in Z} P_r(M), \quad Z \subseteq X. \tag{3}$$

Let $\bar{M}$ denote M with nonzero rows normalized to unit length. For a resolution parameter $\tau > 0$, the reserve polynomial and its log-volume are

$$D_\tau(Z) = \det(I_k + \tau^{-1} K_Z), \qquad c_\tau(Z) = \log D_\tau(Z) = \sum_{j=1}^{\min(|Z|,k)} \log(1 + \sigma_j^2(\bar{M}_Z)/\tau). \tag{4}$$

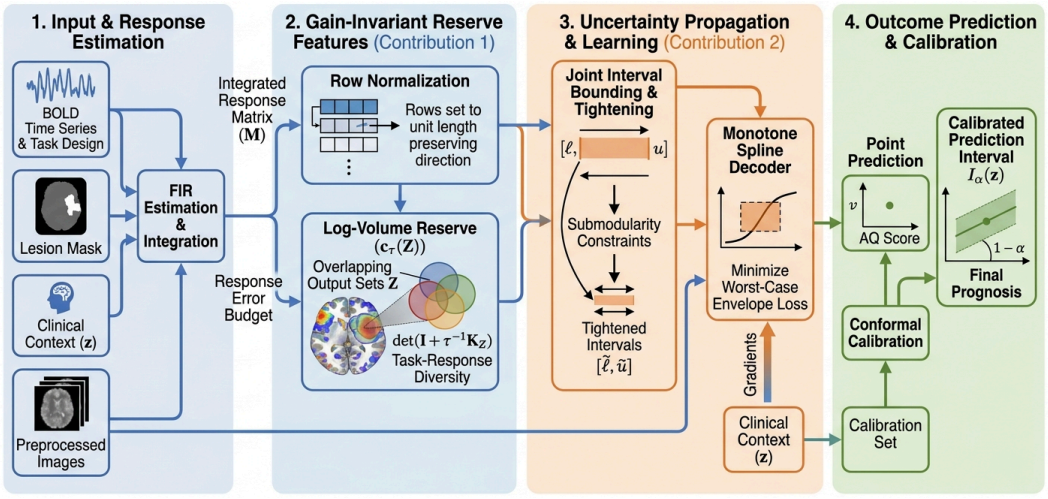


Figure 2: **LANTERN method framework.** Stages 1–4 implement the pipeline described in this section. The displayed determinant is $D_\tau(Z)$, with log-volume $c_\tau(Z) = \log D_\tau(Z)$; the final output $I_\alpha(z)$ is on the WAB-AQ scale.

The projectors discard overall row scale, so at fixed output availability reserve compares only the diversity and nonredundancy of response directions. The finite-scale spectrum distinguishes nearly collinear responses from well-separated responses even when both have the same algebraic rank. We fix $\tau = 0.25$ in the main analysis and assess sensitivity in Appendix D.

Theorem 1 (Task-response quotient and structural degree)**.** *Under Equation* (1)*, the FIR-area operator satisfies Equation* (2)*. Moreover, $P_r(M) = P_r(N)$ for every row if and only if $N = DM$ for an invertible diagonal matrix D. The polynomial $D_\tau(Z)$ has degree $\mathrm{rank}(T_Z)$ in τ^{-1}. For generic nonzero coefficients consistent with the augmented neural graph, this degree equals the maximum number of vertex-disjoint task-to-output paths, including paths through hidden populations and feedback. Equivalently, it is determined by a complementary support-matching rank. The weighted-minor expansion and exact matching identity are given in Appendix A.*

The projector collection is thus a maximal invariant of the measured response under regional gain changes. Its polynomial retains the rank structure as its leading degree and supplies continuous weights through squared minors. Coefficients of D_τ can change between networks having the same generic ranks: these weights describe the measured physiological response, while the degree describes its structural capacity. The theorem requires no assignment of an ICA component to a named brain region. Proofs appear in Appendix A.

4.2 Bounding reserve from uncertain task responses

We carry response-estimation error into a reserve interval for each output set. An FIR estimate (Lindquist, 2008) gives $\hat{M}$ and a simultaneous response-error budget. After propagating row-normalization error, let ϵ_Z satisfy $\|\hat{\bar{M}}_Z - \bar{M}_Z\|_2 \le \epsilon_Z$ on the designated coverage event. With estimated singular values $\hat{\sigma}_j$, initialize

$$\ell_Z = \sum_j \log\left(1 + \frac{(\hat{\sigma}_j - \epsilon_Z)_+^2}{\tau}\right), \qquad u_Z = \sum_j \log\left(1 + \frac{(\hat{\sigma}_j + \epsilon_Z)^2}{\tau}\right). \tag{5}$$

The zero-padding convention includes all $\min(|Z|, k)$ singular values. Weak responses increase the interval rather than forcing a binary edge or rank decision.

Proposition 1 (Coverage and submodularity)**.** *On the simultaneous operator-error event, $\ell_Z \le c_\tau(Z) \le u_Z$ for every queried Z. Also, c_τ is normalized, monotone, and submodular:*

$$c_\tau(\varnothing) = 0, \quad c_\tau(A) \le c_\tau(B)\ (A \subseteq B), \quad c_\tau(A \cup B) + c_\tau(A \cap B) \le c_\tau(A) + c_\tau(B). \tag{6}$$

Intersecting the initial intervals with these inequalities preserves coverage and cannot enlarge any marginal interval. No separation between zero and nonzero singular values is required.

Algorithm 1 LANTERN for one participant and one fitted outcome head

Require: Event design, BOLD time series, lesion mask, clinical context z.

1: Estimate regional FIR responses; integrate over the specified horizon to obtain $\hat{M}$.
2: Propagate response uncertainty through row normalization; compute ϵ_Z.
3: Form output-set spectra and the intervals in Equation (5).
4: Tighten intervals with Equation (6) and increment bounds.
5: Normalize coordinates; evaluate the two monotone decoder endpoints.
6: **return** Point prediction at the envelope midpoint and calibrated interval (9).

We query all 63 nonempty unions of six prespecified anatomical output groups. Their union/intersection closure supports inexpensive bound propagation. For example, $u_{A\cup B} \leftarrow \min\{u_{A\cup B}, u_A + u_B - \ell_{A\cap B}\}$. Monotonicity and one-region increment bounds provide additional updates. The resulting envelope is a coverage-preserving outer approximation; each query is evaluated from the response matrix. Only after these updates do we divide each coordinate by its known maximum $c_{\max}(Z) = q\log(1 + |Z|/(q\tau))$, $q = \min(|Z|, k)$, for input to the decoder. This final normalization avoids imposing submodularity on differently rescaled coordinates.

4.3 Learning from reserve intervals and calibrating predictions

The refined intervals enter model fitting in stage 3 of Figure 2. Let $[\tilde{\ell}, \tilde{u}]$ be the normalized reserve envelope. We use a monotone spline decoder (Ramsay, 1988)

$$g_\theta(z, c) = \beta_0 + \beta^\top z + \sum_{q=1}^{63}\sum_{b=1}^{4} w_{qb}(c_q - t_b)_+, \qquad w_{qb} \geq 0, \quad t_b \in \{0, 0.25, 0.5, 0.75\}. \tag{7}$$

Monotonicity encodes a specific modeling choice: at fixed clinical context, greater task-conditioned reserve should not reduce predicted language capacity. We optimize

$$\min_\theta \frac{1}{N}\sum_i \max\Big\{[y_i - g_\theta(z_i, \tilde{\ell}_i)]^2, [y_i - g_\theta(z_i, \tilde{u}_i)]^2\Big\} + \lambda\|\theta_{-0}\|_2^2. \tag{8}$$

This convex objective trains against the larger endpoint error, which is the exact worst-case loss over the reserve interval box. A wider interval matters whenever the predictor is sensitive to that reserve coordinate. Stage 4 then calibrates residual outcome error using a disjoint set of participants; it expands the endpoint range while leaving the fitted decoder and its midpoint-based point prediction fixed.

Theorem 2 (Propagation and predictive calibration)**.** *If the true normalized reserve lies in* $[\tilde{\ell}, \tilde{u}]$*, then its fitted prediction lies between* $g_\theta(z, \tilde{\ell})$ *and* $g_\theta(z, \tilde{u})$*. For an independently fitted model, use calibration scores* $s_i = \max\{g_\theta(z_i, \tilde{\ell}_i) - y_i,\ y_i - g_\theta(z_i, \tilde{u}_i),\ 0\}$*. Let* q_α *be the* $\lceil (n_{\text{cal}}+1)(1-\alpha)\rceil$*th ordered score, with the usual* $+\infty$ *convention. Under exchangeability of the calibration and test examples,*

$$I_\alpha(z) = [g_\theta(z, \tilde{\ell}) - q_\alpha, g_\theta(z, \tilde{u}) + q_\alpha] \cap [0, 100] \tag{9}$$

has marginal coverage at least $1 - \alpha$ *for an outcome in* $[0, 100]$.

The two coverage statements answer different questions: the conformal interval covers predictive residuals, and cross-center transfer is an empirical finding.

5 Evaluation design

Cohorts and endpoints. Public ARC Gibson et al., 2024 evaluates current WAB-AQ in 152 participants with naming-task fMRI, resting fMRI, T1 MRI, and usable clinical assessments. Its released design contains two conditions, `image` and `abstract`. The internal cohort evaluates six-month WAB-AQ in 237 participants with first left-hemispheric ischemic stroke and aphasia. Baseline imaging at 2–6 weeks includes a six-condition language battery. Center A supplies 116 fitting and 37 independent calibration participants; all 84 center-B participants form the locked evaluation cohort. Appendix B gives eligibility, acquisition, and visit alignment.

Table 1: **Language-outcome prediction.** Lower MAE and higher R^2 are better. ARC evaluates current AQ in 152 participants; center B evaluates six-month AQ in 84 participants. All models receive the same available clinical and lesion covariates.

	ARC: current AQ		Center B: six-month AQ	
Method	MAE ↓	R^2 ↑	MAE ↓	R^2 ↑
Clinical + lesion	11.52	0.36	10.16	0.31
FC elastic net	11.02	0.40	9.74	0.29
Task-FIR ridge	9.45	0.54	9.20	0.46
Brain network transformer	9.51	0.54	7.60	0.57
Point-reserve decoder	7.88	0.70	6.47	0.71
LANTERN	**6.75**	**0.77**	**5.64**	**0.77**

Representation and comparators. Lesion-aware preprocessing follows a BIDS/fMRIPrep workflow (Gorgolewski et al., 2016; Esteban et al., 2019). We summarize 18 prespecified language-related parcels in six groups, with fixed Schaefer2018 atlas correspondences (Schaefer et al., 2018). All 63 nonempty group unions are queried. FIR integration uses a 40-s horizon; run/block bootstrap repeats fitting and normalization to estimate response-error budgets. We compare clinical-plus-lesion ridge, resting-FC elastic net, task-FIR ridge, a brain network transformer (Kan et al., 2022), and a point-reserve decoder. The point-reserve model retains the spline architecture and isolates the effect of envelope learning. Appendix B.4 specifies fitting and tuning.

Evaluation and contribution tests. ARC uses participant-grouped five-fold outer cross-validation with inner hyperparameter selection. Internal tuning uses only the 116 fitting participants, and calibration never fits decoder weights. Primary metrics are MAE and R^2; uncertainty estimates use 2,000 paired participant-bootstrap draws. Controlled response experiments test continuous diversity, regional gains, additive noise, task-event budgets, integration horizons, and valid reserve bounds. A descriptive analysis of all 237 internal response profiles evaluates all task subsets without outcome-based selection. Component ablations use 20 center-A development resplits. Held-out retention and prediction-interval analyses evaluate uncertainty utility, while clinical adjustment and the two existing baseline runs test association, group-level repeatability, and cross-run patient identification.

6 Results

6.1 Language impairment and six-month outcome

Table 1 reports the two complementary endpoints. On ARC, LANTERN attains MAE 6.75 (95% CI 5.95–7.55) and $R^2 = 0.77$. At held-out center B, MAE is 5.64 (4.57–6.85) and $R^2 = 0.77$, compared with 7.60 (6.09–9.21) and $R^2 = 0.57$ for the brain network transformer. The paired reduction is 1.96 AQ points (95% CI 0.72–3.25), or 25.8% relative to that comparator. The point-reserve decoder improves on task-amplitude and FC baselines, and envelope learning provides a further improvement. Participant-level predictions and the complete bootstrap comparisons appear in Appendix C.

6.2 Response-pattern diversity and regional-gain robustness

The first contribution separates task-response diversity from regional amplitude. In Figure 3a, two-condition responses have the same algebraic rank over almost the entire range, while continuous reserve resolves their angular separation. Direct checks verify the path/edge-rank identities in Theorem 1; gain heterogeneity leaves quotient reserve invariant to numerical precision (Appendix A).

Figures 3b,c compare raw-response and quotient reserve on the same gain–noise grid and color scale. Pure gain changes leave quotient reserve invariant. Additive noise creates estimation error, especially when low gains weaken the measured response. Thus exact gain invariance and noisy-response robustness are distinct properties. Appendix D reports the gain-only and independently gain-invariant Pearson controls.

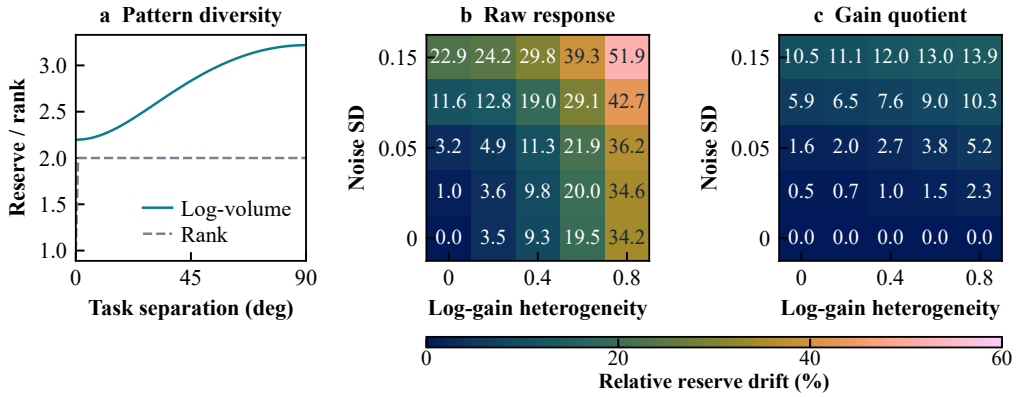


Figure 3: **Pattern diversity and gain robustness.** (a) Continuous log-volume distinguishes two-condition response patterns with the same rank. (b,c) Relative drift of raw-response log-volume and gain-quotient reserve under crossed regional gain and additive noise, averaged over 24 independent profiles.

6.3 Task information and delayed-response integration

Task diversity versus measurement precision. We evaluate all 63 nonempty subsets of the six-condition battery. Across the 237 internal response profiles, every one-condition subset shows near-zero between-patient reserve variation (Figure 4a). With fixed output availability, a single nonzero response coordinate has $P_r = 1$, so repeating that condition improves precision without generating new feature directions. All two-condition subsets yield positive variation (SD 0.033–0.049); the six-condition battery gives SD 0.068. This explains why even ARC's two-condition design can carry individual information beyond integer rank.

The event-budget analysis subsamples fitting-cohort response profiles at five information budgets (Figure 4b). For six conditions, increasing the budget from 60 to 240 events reduces mean normalized reserve error from 0.0293 to 0.0125. The one-condition comparison shows limited feature variation regardless of estimation precision. Thus task contrast creates the feature's individual variation, while additional events improve its precision. The analysis varies retained event density while holding the same underlying response structure.

Integration recovers delayed response area. Holding the full-area target fixed, we vary hemodynamic latency up to 16 s and integrate over 16–48 s (Figure 4c). At the largest delay, mean normalized reserve error falls to 0.109% with a 40-s horizon and 0.073% with 48 s. Capturing the response tail makes the finite FIR estimate approach the area operator used by the theory. Appendix D gives the delay protocol and outcome sensitivity to the horizon and τ.

6.4 Measurement uncertainty in learning and prediction

Table 2 isolates the components of the two contributions. Removing gain normalization causes the largest deterioration. Removing joint tightening increases reserve error, while removing envelope loss worsens outcome prediction even with the same point representation. Lesion context supplies complementary information about tissue loss.

Joint constraints narrow reserve bounds across the output sets (Figure 5a). Systematic operator-error verification confirms that the bounds contain the true continuous reserve, and set-function diagnostics confirm that tightening preserves feasibility. In the clinical cohort, the mean interval contraction is 0.93%.

Ranking patients by uncalibrated, unclipped endpoint width identifies cases with lower prediction error (Figure 5b). Retaining the more certain half of center B, or 42 patients, reduces MAE from 5.64 to 4.28 AQ points: a 1.36-point reduction (95% CI 0.39–2.61), or 24.1%. The uncertainty score correlates with absolute error at $\rho = 0.40$. The retention criterion uses only measurement information, following the selective-prediction perspective of Geifman & El-Yaniv (2017).

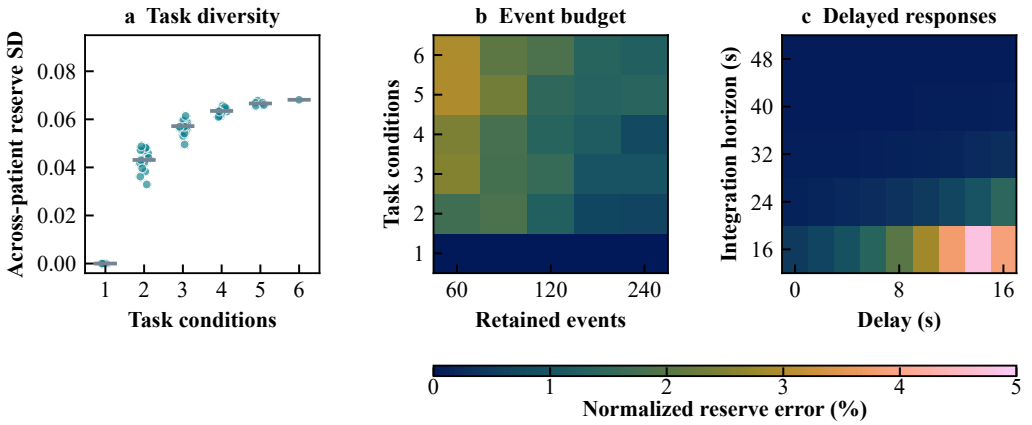


Figure 4: **Measurement conditions for informative reserve.** (a) Between-patient normalized-reserve SD in 237 participants, with one dot per task subset and a segment at the mean for each task dimension. (b) Estimation error across task dimensions and retained event budgets. (c) Error across hemodynamic delays and integration horizons. Panels (b,c) summarize 24 independent response profiles per cell and share a scale for 100 times the absolute normalized-reserve error.

Table 2: **Contribution-specific ablations.** Outcome MAE averages 20 resplits of the 116 center-A fitting participants. Reserve error is normalized absolute error in the controlled evaluation. Each row changes one component; downstream-only changes retain the same reserve estimate.

Variant	MAE ↓	Reserve error (controlled)	MAE change
Full LANTERN	6.46	0.077	+0.00
No gain quotient	8.51	0.209	+2.05
No submodular envelope	7.42	0.134	+0.96
No envelope loss	7.49	0.077	+1.03
No lesion context	7.85	0.077	+1.39

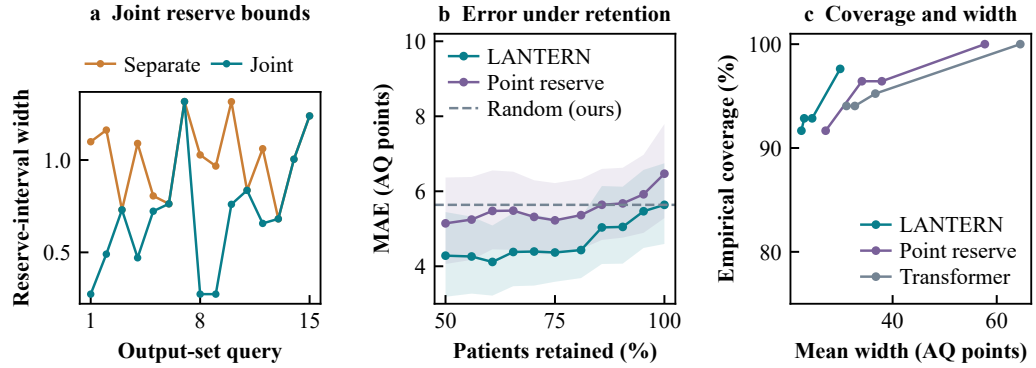


Figure 5: **Measurement uncertainty from reserve bounds to predictions.** (a) Separate and jointly tightened widths for overlapping output queries in a direct response-matrix example. (b) Center-B MAE versus retained fraction, with 95% participant-bootstrap bands and a common patient ordering for both decoders. (c) Empirical coverage versus mean AQ-interval width at nominal levels 80%, 85%, 90%, and 95%; every method is calibrated on the same disjoint center-A cohort.

Figure 5c turns to prediction-interval coverage. At nominal 90%, LANTERN achieves 92.9% coverage with mean width 24.54 AQ points, compared with 95.2% and 36.69 for the calibrated transformer. Widths are computed after clipping to $[0, 100]$. Appendix C gives individual intervals, the full calibration curves, and the retention protocol.

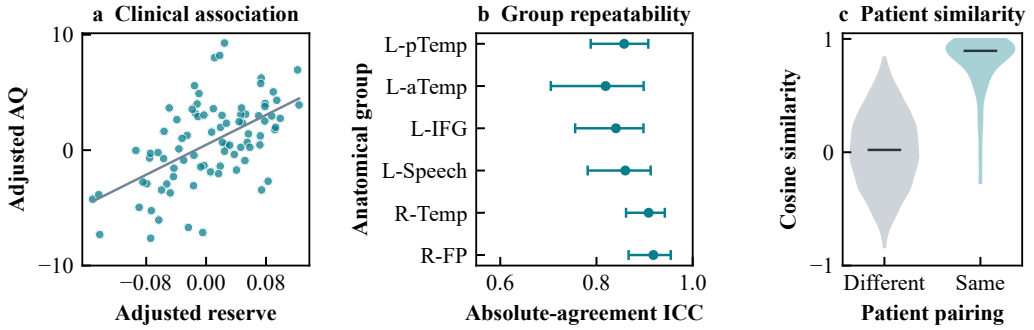


Figure 6: **Clinical relevance, repeatability, and individual distinctiveness.** (a) Full-network reserve and six-month AQ after clinical adjustment fitted at center A and applied to 84 held-out participants. (b) Group-level absolute-agreement ICC across the two baseline task runs, with 95% participant-bootstrap intervals. (c) Cross-run cosine similarity for 84 same-patient and 6,972 different-patient pairs; violins summarize distributions and bars mark medians. Top-1 matching across 84 candidates identifies 83.3% of patients. Anatomical group definitions appear in Table S2.

6.5 Clinical relevance and individual consistency

Full-network reserve remains associated with six-month AQ after development-fitted adjustment for baseline AQ, age, and log lesion volume: held-out Spearman $\rho = 0.55$ (95% CI 0.38–0.69; Figure 6a). Group reserve has cross-run ICC(A,1) values of 0.82–0.92 (Figure 6b), linking the representation to clinical variation and repeatability within one visit.

Cross-run patient identification. Group ICC assesses coordinate reliability (Koo & Li, 2016); we next test whether the full reserve profile preserves enough individual variation to distinguish patients (Finn et al., 2015). For each of the 84 held-out patients, we compare the 63-coordinate profile from run 1 with all 84 run-2 profiles by cosine similarity. Coordinates are standardized using center-A fitting statistics, then each profile is centered across coordinates to remove its average standardized reserve level. The most similar run-2 profile identifies the correct patient for 70/84 queries (83.3%), compared with 1.2% expected under random matching (Figure 6c). Thus the two baseline runs retain a reproducible individual pattern beyond each profile's overall standardized level within the same visit.

Appendix E gives the full matching matrix, clinically matched profiles, cortical contribution maps, and finer-scale reliability; Appendix F evaluates reserve completion when regional measurements are unavailable.

7 Discussion and conclusion

LANTERN combines a gain-invariant measure of task-response diversity with prediction learned from reserve intervals. Independent calibration yields outcome prediction intervals. Its pathway interpretation accommodates hidden populations and feedback. Clinical association and cross-run patient identification support the relevance and distinctiveness of reserve profiles within a visit; uncertainty-based retention links measurement reliability to prediction error. One evaluation targets current impairment, the other six-month outcome.

The acquisition experiments trace informative reserve back to concrete properties of the scanning protocol: multiple task contrasts supply the individual variation, event count sets the estimation precision, and the integration horizon determines how much of a delayed response is recovered. Together with the gain controls, these results tie the reserve definition to quantities an acquisition designer can actually change. Prospective rehabilitation studies can build on these prognostic intervals to evaluate decision thresholds and treatment benefit.

AI USE STATEMENT

Generative AI was used to polish the language of the manuscript and to review the text for clarity, consistency, and internal coherence.

REFERENCES

Francis Bach. *Learning with submodular functions: A convex optimization perspective*. Now Publishers, 2013.

Andra Braban, Robert Leech, Kevin Murphy, and Fatemeh Geranmayeh. Cerebrovascular reactivity has negligible contribution to hemodynamic lag after stroke: implications for functional magnetic resonance imaging studies. *Stroke*, 54(4):1066–1077, 2023.

Caterina Breitenstein, Tanja Grewe, Agnes Flöel, Wolfram Ziegler, Luise Springer, Peter Martus, Walter Huber, Klaus Willmes, E Bernd Ringelstein, Karl Georg Haeusler, et al. Intensive speech and language therapy in patients with chronic aphasia after stroke: a randomised, open-label, blinded-endpoint, controlled trial in a health-care setting. *The Lancet*, 389(10078):1528–1538, 2017.

Haoyue Dai, Immanuel Albrecht, Peter Spirtes, and Kun Zhang. Distributional equivalence in linear non-gaussian latent-variable cyclic causal models: Characterization and learning. *arXiv preprint arXiv:2603.04780*, 2026.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

Isaac Falconer, Maria Varkanitsa, Emerson Kropp, and Swathi Kiran. Network measures from the rewired simulation framework enhance prediction of post-stroke aphasia severity. *bioRxiv*, pp. 2026–05, 2026.

Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4): 1273–1302, 2003.

DJ Gale, RV de Wael, O Benkarim, and B Bernhardt. Surfplot: Publication-ready brain surface figures (v0. 1.0). zenodo, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Makayla Gibson, Roger Newman-Norlund, Leonardo Bonilha, Julius Fridriksson, Gregory Hickok, Argye E Hillis, Dirk-Bart Den Ouden, and Christopher Rorden. The aphasia recovery cohort, an open-source chronic stroke repository. *Scientific Data*, 11(1):981, 2024.

Evan M Gordon, Timothy O Laumann, Adrian W Gilmore, Dillan J Newbold, Deanna J Greene, Jeffrey J Berg, Mario Ortega, Catherine Hoyt-Drazen, Caterina Gratton, Haoxin Sun, et al. Precision functional mapping of individual human brains. *Neuron*, 95(4):791–807, 2017.

Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):160044, 2016.

Franziska E Hildesheim, Anja Ophey, Anna Zumbansen, Thomas Funck, Tibor Schuster, Keith W Jamison, Amy Kuceyeski, and Alexander Thiel. Predicting language function post-stroke: A model-based structural connectivity approach. *Neurorehabilitation and neural repair*, 38(6):447–459, 2024.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Aubrey W Ingleton and Mike J Piff. Gammoids and transversal matroids. *Journal of Combinatorial Theory, Series B*, 15(1):51–68, 1973.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022.

Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Martin A Lindquist. The statistical analysis of fmri data. 2008.

Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157, 2001.

Trevor A Low, Matthew J Chilvers, Harold Zhu, Helen L Carlson, Ashley D Harris, Bradley G Goodyear, and Sean P Dukelow. Structural network topology associated with naming improvements following intensive aphasia therapy in post-stroke aphasia. *Journal of the Neurological Sciences*, 462:123065, 2024.

Barbara K Marebwa, Julius Fridriksson, Grigori Yourganov, Lynda Feenaughty, Chris Rorden, and Leonardo Bonilha. Chronic post-stroke aphasia severity is determined by fragmentation of residual white matter networks. *Scientific reports*, 7(1):8188, 2017.

James O Ramsay. Monotone regression splines in action. *Statistical science*, pp. 425–441, 1988.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

Lloyd S Shapley et al. A value for n-person games. 1953.

Cynthia M Shewan and Andrew Kertesz. Reliability and validity characteristics of the western aphasia battery (wab). *Journal of Speech and Hearing Disorders*, 45(3):308–324, 1980.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Joshua S Siegel, Abraham Z Snyder, Lenny Ramsey, Gordon L Shulman, and Maurizio Corbetta. The effects of hemodynamic lag on functional connectivity and behavior after stroke. *Journal of Cerebral Blood Flow & Metabolism*, 36(12):2162–2176, 2016.

Serena E Song, Lisa C Krishnamurthy, Amy D Rodriguez, Joo H Han, Bruce A Crosson, and Venkatagiri Krishnamurthy. Methodologies for task-fmri based prognostic biomarkers in response to aphasia treatment. *Behavioural Brain Research*, 452:114575, 2023.

James D Stefaniak, Fatemeh Geranmayeh, and Matthew A Lambon Ralph. The multidimensional nature of aphasia recovery post-stroke. *Brain*, 145(4):1354–1367, 2022.

Reinder Vos de Wael, Oualid Benkarim, Casey Paquola, Sara Lariviere, Jessica Royer, Shahin Tavakol, Ting Xu, Seok-Jun Hong, Georg Langs, Sofie Valk, et al. Brainspace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Communications biology*, 3(1):103, 2020.

Max Wawrzyniak, Hans R Schneider, Julian Klingbeil, Anika Stockert, Gesa Hartwigsen, Cornelius Weiller, and Dorothee Saur. Resolution of diaschisis contributes to early recovery from post-stroke aphasia. *Neuroimage*, 251:119001, 2022.

Janina Wilmskoetter, Xiaosong He, Lorenzo Caciagli, Jens H Jensen, Barbara Marebwa, Kathryn A Davis, Julius Fridriksson, Alexandra Basilakos, Lorelei P Johnson, Chris Rorden, et al. Language recovery after brain injury: A structural network control theory study. *The Journal of Neuroscience*, 42(4):657–669, 2022.

A Proofs and mathematical details

A.1 Integrated responses and regional-gain invariance

Let $A = I - B$ with all eigenvalues in the open right half-plane. The neural impulse response to the task design is $e^{-At}C\mathbb{1}\{t \geq 0\}$. Its integral is

$$\int_0^\infty e^{-At}C\,dt = A^{-1}C, \tag{10}$$

because $A\int_0^R e^{-At}\,dt = I - e^{-AR}$ and $e^{-AR} \to 0$. Integrability permits exchanging the convolution integrals, so the BOLD response area at region r is a_r times the neural response area. This proves Equation (2). For finite horizon H, the omitted neural/hemodynamic tail is an additive error in M; a gain identity for an infinite integral is not an assertion that every finite window captures the entire response.

For a nonzero row, $P_r(M)$ is the orthogonal projector onto the one-dimensional span of m_r. Multiplication by any nonzero scalar leaves this projector unchanged. Conversely, equal rank-one projectors have the same span, so $n_r = d_r m_r$ for some $d_r \neq 0$. A zero projector identifies a zero row, where d_r can be chosen as one. Applying this argument row by row proves the if-and-only-if claim in Theorem 1. The projectors are therefore a maximal invariant for the diagonal gain action, not merely an invariant summary chosen for convenience.

This equivalence relation is useful with latent neural structure. Any two neural/hemodynamic models whose integrated task responses differ only by these gains have identical projectors, regardless of their latent labels or recurrence pattern. Task conditions remain labeled external inputs. In particular, an input condition must not be reinterpreted as a direct intervention on whichever observed region has the largest response.

A.2 Weighted bases, polynomial degree, and complementary matchings

$$D_\tau(Z) = \sum_{j=0}^{k} \tau^{-j} \sum_{\substack{R\subseteq Z,\ J\subseteq[k]\\ |R|=|J|=j}} \det(\bar{M}_{R,J})^2 \tag{11}$$

Since $K_Z = \bar{M}_Z^\top \bar{M}_Z$, the coefficients of $\det(I + tK_Z)$ are elementary symmetric polynomials of the eigenvalues of K_Z. Expanding the principal minors and applying Cauchy–Binet gives

$$\det(I + tK_Z) = \sum_{j=0}^{k} t^j \sum_{\substack{J\subseteq[k]\\ |J|=j}} \det\left((\bar{M}_Z^\top \bar{M}_Z)_{J,J}\right) \tag{12}$$

$$= \sum_{j=0}^{k} t^j \sum_{\substack{R\subseteq Z, J\subseteq[k]\\ |R|=|J|=j}} \det(\bar{M}_{R,J})^2. \tag{13}$$

All coefficients are nonnegative. The largest nonzero coefficient has index $\mathrm{rank}(\bar{M}_Z)$, which equals $\mathrm{rank}(M_Z) = \mathrm{rank}(T_Z)$ because the nonzero row normalizations and vascular gains do not change rank. Consequently,

$$\lim_{\tau\downarrow 0} \frac{c_\tau(Z)}{\log(1/\tau)} = \mathrm{rank}(T_Z). \tag{14}$$

At a finite τ, the coefficients weight response bases by squared volume; the representation preserves more information than the polynomial degree alone.

For the graphical interpretation, order the augmented vertices as $W = (U, V)$ and define

$$B^+ = \begin{bmatrix} 0 & 0 \\ C & B \end{bmatrix}, \qquad (I - B^+)^{-1} = \begin{bmatrix} I & 0 \\ (I-B)^{-1}C & (I-B)^{-1} \end{bmatrix}. \tag{15}$$

The observed task transfer is the Z, U block of this inverse. The nullity identity for complementary blocks of an invertible matrix gives

$$\operatorname{rank}\left([(I - B^{+})^{-1}]_{Z,U}\right) = |Z| + |U| - |W| + \operatorname{rank}\left((I - B^{+})_{W\setminus U, W\setminus Z}\right). \tag{16}$$

For generic coefficients, the rightmost rank is the matching rank of its support. The support contains ones on its full diagonal and allowed neural/task edges elsewhere. A nonzero determinant term corresponds to a matching; with independent free edge coefficients, its polynomial does not vanish identically. Equivalently, the inverse block represents a gammoid whose rank counts vertex-disjoint directed paths (Ingleton & Piff, 1973). This proves Equation (16). Stability restricts the weights to an open domain and does not change the generic polynomial identity. Rank cancellations at exceptional parameter values remain possible; Equation (4) is computed from the measured response and remains well defined there.

The identified object. The identified object is the task-response orbit under regional gain changes, summarized by a specified finite collection of output sets. The relation to path/edge ranks supplies a structural interpretation of the leading degree, while the continuous coefficients retain physiological information beyond the rank signature.

A.3 Continuous reserve bounds and joint tightening

Weyl's singular-value inequality (Horn & Johnson, 2012) implies

$$\max\{\hat{\sigma}_j - \epsilon_Z, 0\} \le \sigma_j(\bar{M}_Z) \le \hat{\sigma}_j + \epsilon_Z. \tag{17}$$

Applying the increasing function $s \mapsto \log(1+s^2/\tau)$ and summing proves Equation (5). Its derivative is $2s/(\tau + s^2) \le 1/\sqrt{\tau}$, giving the useful continuity bound

$$|c_\tau(\hat{\bar{M}}_Z) - c_\tau(\bar{M}_Z)| \le \frac{\min(|Z|, k)}{\sqrt{\tau}}\epsilon_Z. \tag{18}$$

The main uncertainty result therefore remains meaningful near a rank change. Inferring exact algebraic rank from finite noisy data would require an additional spectral separation condition, which the continuous reserve bound does not use.

For submodularity, write a nonzero row projector as $v_r v_r^\top$ with $\|v_r\| = 1$. The matrix determinant lemma yields

$$c_\tau(A \cup \{r\}) - c_\tau(A) = \log\left(1 + v_r^\top(\tau I + K_A)^{-1} v_r\right), \quad r \notin A. \tag{19}$$

If $A \subseteq B$, then $K_A \preceq K_B$ and inverse order is reversed. The marginal increment is thus no smaller for A than for B, proving diminishing returns, and hence submodularity. Each increment is nonnegative and at most $\log(1 + 1/\tau)$, proving monotonicity and the per-region upper bound. Zero rows have zero increment.

For any family closed under union and intersection, the actual capacity vector satisfies all the inequalities used in tightening. Each propagated lower bound follows from necessary inequalities and existing valid lower/upper bounds, and each propagated upper bound does likewise. Induction over updates therefore preserves the true vector, and an outer box is all the monotone decoder requires.

Normalization uncertainty. Suppose a raw row has a confidence radius $\|\hat{m}_r - m_r\| \le \delta_r$. If both rows are nonzero,

$$\left\| \frac{\hat{m}_r}{\|\hat{m}_r\|} - \frac{m_r}{\|m_r\|} \right\| \le \min\left\{2, \frac{2\delta_r}{\|\hat{m}_r\|}\right\} =: \eta_r. \tag{20}$$

For an unresolved zero/near-zero row we use the conservative value two. A simultaneous bound for an output set is $\epsilon_Z = (\sum_{r\in Z} \eta_r^2)^{1/2}$, since the Frobenius norm dominates the operator norm. This makes low-response regions uncertain instead of converting normalized noise into apparently reliable reserve. Tighter directly bootstrapped operator bounds may be used, but their coverage must then be verified empirically; it does not follow from the deterministic error-bound theorem.

A.4 INTERVAL-BASED LEARNING AND CONFORMAL CALIBRATION

Every spline feature in Equation (7) is nondecreasing and every corresponding weight is nonnegative. Hence $g_\theta(z, \tilde{\ell}) \le g_\theta(z, c) \le g_\theta(z, \tilde{u})$ for every c in the box. A squared residual is convex in the scalar prediction, so its maximum over this prediction interval is attained at an endpoint. This proves the worst-endpoint formula in Equation (8). The two endpoint predictions are affine in the parameters; the maximum of their squared residuals is convex. Nonnegative spline coefficients and quadratic regularization preserve convexity.

For conformal coverage (Lei et al., 2018; Romano et al., 2019), condition on the data used to fit the representation settings and decoder. Under exchangeability, the calibration scores and the new test score have an exchangeable rank. The prescribed order statistic makes the probability that the test score exceeds q_α at most α, with ties giving conservative coverage. The event that the test score is at most q_α is precisely membership in the untruncated interval in Equation (9). Intersecting with $[0, 100]$ does not remove a valid AQ outcome. This completes Theorem 2. The result is marginal coverage over the calibration distribution.

Numerical checks. The reference checks evaluate recurrent neural graphs with seven neural vertices and three task roots. Random edge supports have probability 0.24; neural weights are rescaled to an infinity norm at most 0.65, ensuring stability of the continuous-time system. The task-root support is sampled independently. Across the generated graphs, multiple output/input queries per graph compare the transfer-block rank to the complementary matching identity. An additional configuration test applies heterogeneous positive amplitudes and unit-modulus row factors to verify normalized spectral invariance. These unit-modulus factors extend the real-area calculation to the corresponding frequency-domain invariance.

Continuous reserve bounds are checked under known operator-norm deviations between 0.005 and 0.4 across 1,500 matrices. Random capacity functions over four output blocks test submodular tightening. Additional diagnostic tests check exact-rank envelopes under an explicitly imposed singular-value separation and discrete matroid envelopes. Those separated-rank checks are supplementary; the main continuous method does not assume the separation condition.

B COHORTS, ACQUISITION, AND IMPLEMENTATION

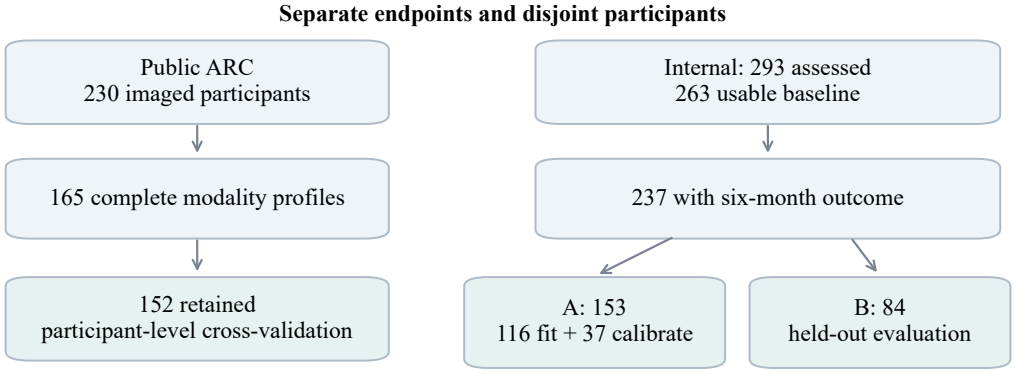


Figure S1: **Cohort organization.** ARC supports current-impairment evaluation; the internal cohort supports six-month prognosis. Internal development, calibration, and held-out evaluation are disjoint. Public imaging availability is counted from files, not from the larger participant metadata table.

B.1 PUBLIC ARC RELEASE AND ENDPOINT ALIGNMENT

The release is pinned to OpenNeuro version 1.0.2 through its versioned dataset DOI. The source descriptor reports 230 imaged participants and 902 sessions (Gibson et al., 2024). A repository audit

identifies 894 naming-task series from 193 participants and 508 resting-state series from 192 participants. Of the 193 task participants, 192 have a T1 series and a nonmissing AQ, and 165 additionally have resting-state fMRI. The root metadata file contains 245 participant rows; its denominator is distinct from those modality-complete subsets.

The root event file has two trial types, `image` and `abstract`. Each is modeled with its own FIR columns. Event order is not converted into unreported item identity, and repeated runs do not create additional independent task conditions. The maximum algebraic task rank is therefore two. Continuous reserve can still vary through the relative alignment of regional task-response directions, as shown in Figure 3a.

The assessment fields include `wab_aq` and `wab_days`. We match the eligible imaging visit nearest the assessment and retain the timing difference as a covariate. The released root table supplies a participant-level AQ rather than a visit-resolved six-month recovery target. All visits and series belonging to a participant remain within the same outer fold. Nine participants are excluded for excessive motion or insufficient usable frames, and four for unsuccessful lesion-aware registration, leaving 152 in the complete-modality analysis. Public conclusions concern severity association and response robustness, not a six-month treatment-response label.

B.2 Internal cohort and prospective target

The two-center protocol includes adults aged 18–85 years with a first left-hemispheric ischemic stroke, clinically documented aphasia, and baseline imaging 2–6 weeks after onset. Over a 36-month enrollment period, clinical eligibility and MRI-safety assessment identify 293 consenting adults for baseline acquisition. Pre-existing progressive neurological disease, previous major stroke, and MRI contraindications are exclusions. Missing required structural imaging ($n = 7$), excessive motion ($n = 15$), and incomplete task acquisition ($n = 8$) remove 30 participants under a prespecified primary-exclusion hierarchy. Of the 263 participants with usable baseline data, 26 lack an in-window six-month AQ, leaving 237. The screening log distinguishes acquisition failure from incorrect task responses and records follow-up noncompletion separately.

Table S1: **Internal cohort partition.** Center membership is fixed before outcome modeling. The split within center A is stratified by baseline severity, without using six-month outcomes.

Partition	Center	Participants	Use
Development cohort	A	153	Fitting and calibration
Model fitting	A	116	Inner selection and final fitting
Calibration	A	37	Predictive interval radius only
Held-out center	B	84	Locked evaluation
Total	Two	237	No participant overlap

Baseline structural acquisition includes T1, FLAIR, and a lesion mask. Resting-state and task BOLD use 3-T imaging with a target 1.2-s repetition time and 2.5-mm isotropic resolution. The task battery consists of two 12-minute runs, with 240 jittered events in total, balanced across six conditions. The conditions are image naming, abstract-image viewing, spoken-word comprehension, semantic association, sentence-to-picture matching, and repetition. A brief practice session checks instruction comprehension and tolerability. Naming and repetition use overt speech with synchronized MR-compatible audio recording; response-scored comprehension and matching trials use a predefined response interface. Abstract-image viewing is passive and has no correctness score. Permitted motor accommodations preserve the stimulus and response rule within each condition. Task responses, omissions, and acquisition interruptions are recorded separately. Anatomical output groups and task coding are fixed independently of the recovery outcome.

The primary outcome is WAB-AQ measured six months after stroke onset within a ± 30-day window. Baseline AQ is assessed within seven days of baseline imaging and is available to the longitudinal decoder, preventing imaging features from receiving credit merely for rediscovering baseline impairment. Clinical features include age, sex, education, time after onset, baseline AQ, lesion volume, output-parcel tissue fractions, and documented speech-language therapy minutes up to the prediction landmark. Exposure after the landmark is not supplied to the baseline prognostic model.

B.3 PRESPECIFIED REGIONAL OUTPUT SETS

The six groups contain three anatomically matched language-related regions each (Table S2). They are defined before label access. The analysis queries all unions of these groups, giving $2^6 - 1 = 63$ nonempty output sets. The same anatomical correspondence is used across datasets, while the number of task conditions remains two for ARC and six for the internal battery. Different outcome heads and condition dimensions are retained; the two datasets are never pooled as one homogeneous sample.

Table S2: **Anatomical output groups.** Names describe fixed parcel correspondences. Lesion fractions and unavailable tissue are recorded separately.

Group	Regional correspondence
Left posterior temporal	Posterior superior temporal, posterior middle temporal, supramarginal
Left anterior temporal	Anterior superior temporal, anterior middle temporal, temporal pole
Left inferior frontal	Pars opercularis, pars triangularis, pars orbitalis
Left speech production	Ventral precentral, anterior insula, supplementary motor
Right temporal	Superior temporal, middle temporal, anterior temporal
Right frontal/parietal	Inferior frontal, ventral precentral, supramarginal

B.4 IMPLEMENTATION, UNCERTAINTY, AND COMPARATORS

FIR estimation. Task inputs are scaled to unit event area. Regional FIRs are fitted over 0–40 s, after nuisance regression and temporal prewhitening. The design includes motion parameters, their derivatives, low-order drift, and available physiological terms. Models are fitted within runs before combining compatible estimates. A condition-number check detects designs unable to separate the chosen FIR columns. We extend the horizon to 48 s in sensitivity analysis, and treat differences beyond estimation variability as a tail-error contribution. A single canonical HRF is not imposed on all regions.

Confidence-envelope construction. We use 200 run/block bootstrap replicates, with blocks of 24 s, repeating response fitting and normalization. The maximum error statistic across the queried output sets supplies a simultaneous budget. The deterministic bounds in Proposition 1 are conditional on that budget covering the response error; empirical coverage of the budget is assessed in the controlled evaluation. Low-norm rows receive conservative normalization uncertainty. An empty joint envelope triggers expansion of the response budgets, not replacement by a convenient point graph. Updates continue until the maximum bound change is below 10^{-6} or 30 sweeps have completed. Bounds are clipped only by valid nonnegativity and spectral maximum constraints.

Resolution and decoder settings. The main setting is $\tau = 0.25$, with checks at 0.1, 0.5, and 1.0. There are four nonnegative spline increments per reserve coordinate and an unrestricted linear clinical term. Ridge strength is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ in participant-grouped development folds. Continuous covariates are standardized using fitting data only. The final convex problem uses a projected/bound-constrained optimizer with convergence tolerance 10^{-8}. The full envelope objective is compared with squared loss on the coordinate midpoint using exactly the same knots and covariates.

Comparable information and capacity. Clinical-plus-lesion ridge uses all clinical and tissue-loss features but no functional data. FC elastic net adds the upper triangle of the observed-region resting correlation matrix. Task-FIR ridge adds integrated regional task responses. The brain network transformer receives the same observed-region connectivity and task features; embedding widths 32/64 and one/two transformer blocks are selected inside the training folds, with dropout 0.1/0.3 and weight decay 10^{-4}/10^{-3}. The point-reserve decoder uses the same normalized spectral summaries as LANTERN, isolating the role of uncertainty-aware learning. Baseline output intervals receive their own disjoint calibration procedure. None of the baselines receives a true latent graph.

Computation. For a query containing m_Z rows and k conditions, the reserve can be evaluated from a $k \times k$ Gram matrix at cost $O(m_Z k^2 + k^3)$. The 63-query representation is thus inexpensive

for $k \in \{2, 6\}$. Set-envelope propagation operates on 64 masks including the empty set, with at most 2,016 unordered pairs per sweep. No factorial column search, latent-variable enumeration, or complete equivalence-class traversal is required. The released implementation favors transparent NumPy operations over specialized accelerators.

C Complete clinical prediction and uncertainty results

C.1 Participant-level outcome comparisons

Figure S2 shows the locked center-B predictions and paired errors. Figure S3 complements Table 1 with bootstrap intervals for every comparator and endpoint.

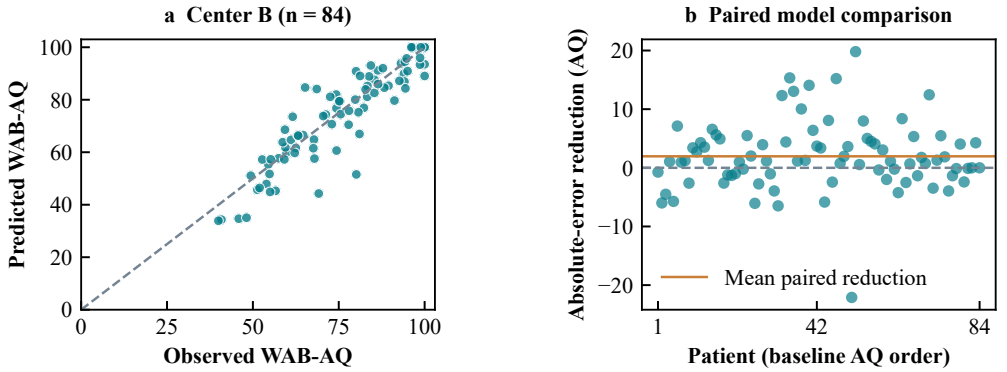


Figure S2: **Longitudinal predictions and paired errors.** (a) Locked predictions for 84 center-B participants; the dashed line is perfect agreement. (b) Transformer absolute error minus LANTERN absolute error, ordered by baseline AQ. Positive values favor LANTERN; the orange line is the mean paired reduction.

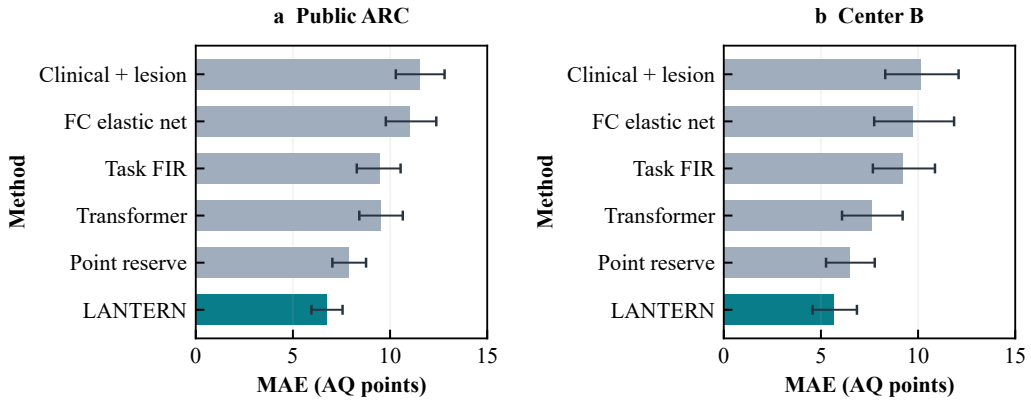


Figure S3: **Complete outcome comparisons.** MAE and 95% participant-bootstrap intervals for current AQ in ARC and six-month AQ at center B. These are the same estimates summarized in Table 1.

C.2 Retention analysis and calibration geometry

The retention score is the difference between the two uncalibrated decoder endpoints. It is computed before clipping to $[0, 100]$, so boundary proximity does not create artificially narrow scores. Sorting does not use a test outcome. We evaluate 11 retained fractions between 0.5 and 1, using the ceiling of the requested fraction times 84. The point-reserve comparator is evaluated on the same

retained patients to separate patient selection from predictor quality. A random-ranking reference has expected MAE equal to the full-cohort MAE. Methods whose residual-calibrated intervals have constant untruncated width do not receive an artificial confidence ranking from AQ clipping.

Pointwise curve bands resample the retained patients. For the primary 50%-retention improvement, each paired bootstrap resamples the full cohort and recomputes the score ordering before comparing retained and full-cohort errors. Coverage–width curves use the same 37 calibration participants at center A for each method, with its own order statistic at each nominal level. They evaluate prediction sets, whereas the retention curves evaluate selected point predictions. Neither operation changes model weights using center B.

Figure S4 illustrates calibrated intervals and the distribution of query-specific tightening. Mean contraction across the clinical queries is 0.93%; the contraction is heterogeneous across patients and output sets, with larger effects concentrated where overlapping sets provide redundant information.

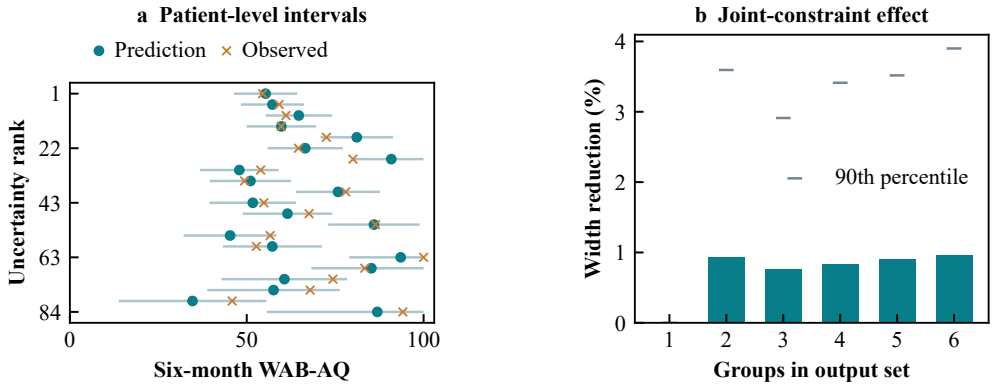


Figure S4: **Individual prediction intervals and query-specific tightening.** (a) Nominal 90% AQ intervals for 21 patients sampled uniformly along the uncertainty ranking. (b) Mean reserve-interval contraction by output-set size; horizontal marks show its 90th percentile. Patient-level prediction widths and reserve-feature contraction are different quantities.

Acquisition information and calibration. Figure S5a compares methods in an independent 40-repetition response evaluation as event counts increase. It complements Figure 4b, which varies task dimension under one quotient-reserve formula on 24 fitting-cohort profiles. Figure S5b reports interval calibration at center B across nominal levels. Each clinical coverage value is the number covered divided by 84. Interval radii use only the 37 calibration participants at center A; at nominal 90% coverage the selected score is the 35th order statistic. Sparse task data create wider response uncertainty, which is carried to the reserve envelope. The joint constraints reduce incompatible uncertainty across overlapping output sets. The output calibration step addresses the residual gap between mechanistic reserve and individual language outcome; it does not redefine the task dimension or change the two-condition public design.

D Observation stress tests and task information

D.1 Gain–noise and integration-horizon stress tests

For each response matrix, we apply a regional gain pattern v_r and an additive noise component E, setting $M' = \mathrm{diag}(\exp(sv))M + \sigma E$. The grid uses $s \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and $\sigma \in \{0, 0.025, 0.05, 0.10, 0.15\}$. Each condition is evaluated on 24 independent response profiles, with matched gain and noise parameters across the grid. Raw-response reserve is $\log\det(I + M^\top M/\tau)$; quotient reserve uses Equation (4). Drift is divided by the appropriate baseline value. The additive noise scale is fixed in response units and does not grow together with vascular gain.

The positive-gain control independently rescales columns of regional time series before recomputing Pearson connectivity. Numerically, maximum Pearson change is below 10^{-15}, as is quotient-reserve

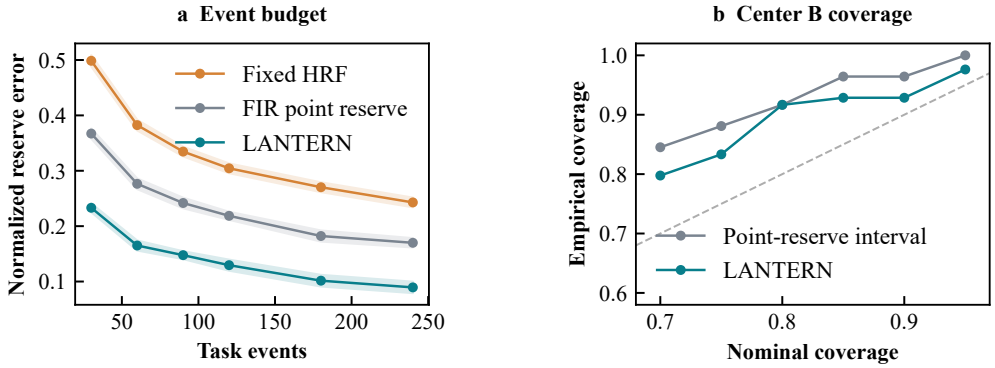


Figure S5: **Measurement budget and calibration.** (a) Normalized reserve error versus task-event count, with 95% confidence bands over 40 repetitions. (b) Coverage in 84 held-out participants at center B, calibrated on 37 disjoint center-A participants. The point-reserve interval omits response-envelope propagation.

drift. In the delay experiment, regional hemodynamic responses are normalized to unit area and shifted by up to 16 s. Condition-dependent neural time constants range from 1 to 7 s. Convolution is integrated on a 0.1-s grid with horizons of 16, 24, 32, 40, and 48 s. Each horizon is compared with the same full-area reserve; there is no change in neural connectivity between horizons.

Figure S6 complements the crossed gain–noise comparison in Figure 3. Pure row scaling changes raw responses but preserves quotient reserve; Pearson correlation is independently invariant to positive rescaling. Figure 4c reports the delay–horizon grid. This experiment isolates omitted response tails while keeping the full-area target fixed; its horizon is the integration window within an FIR estimate, not the scan duration.

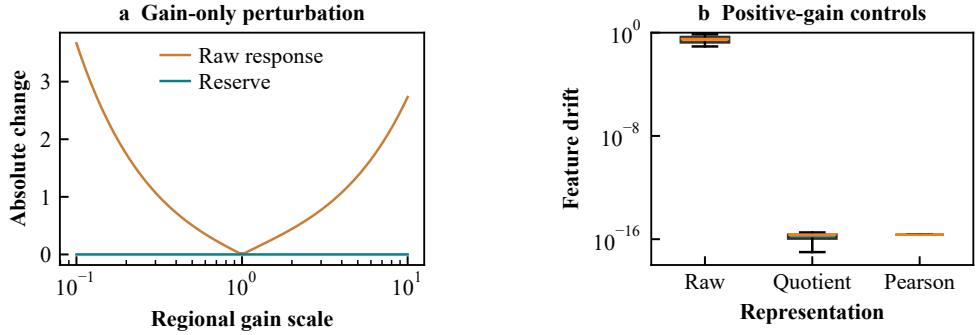


Figure S6: **Gain-only controls.** (a) Direct heterogeneous row rescaling changes the raw response while preserving normalized reserve. (b) Pure positive-gain controls at log-gain heterogeneity 0.8: raw log-volume changes, whereas quotient reserve and Pearson connectivity remain invariant under rescaling. Pearson drift is the maximum entrywise correlation change. Delayed-response integration appears in Figure 4c.

D.2 Task diversity and information budget

Following the distinction between data quantity and feature reliability in precision functional mapping (Gordon et al., 2017), Figure 4 separates descriptive task diversity from estimation precision. Panel (a) uses all 237 internal response profiles and all 63 nonempty subsets of the six conditions; the subsets are fixed without clinical outcomes. Panel (b) is a response-estimation sensitivity analysis on the 116 fitting participants. Each task-dimension/event-budget cell uses the same 24 profiles selected at a fixed four-participant stride within the fitting split; the profile index and chosen task

subset are recorded. The budgets are 60, 90, 120, 180, and 240 events, and the additive response-noise standard deviation is multiplied by $\sqrt{240/n_{\text{events}}}$. This changes the retained information budget, not the acquisition duration. All task dimensions use their own known spectral maximum to normalize reserve.

For a single nonzero task-response coordinate, $P_r = 1$. Hence $c_\tau(Z) = \log(1 + |Z|/\tau)$ when every queried row is present, regardless of response amplitude. This gives both the zero variation and the zero reserve error of the one-condition control. With unavailable rows, this value can still encode the number of observed rows; direction diversity remains absent. The task-diversity analysis is descriptive and does not select an outcome model. ARC retains its released two-condition design throughout.

Resolution and response horizon. Table S3 varies the spectral resolution and integration horizon on the development resplits. The 40-s and 48-s estimates give similar outcome errors; moderate changes in τ preserve the predictive pattern. Increasing the horizon is useful for capturing delayed responses, while an unnecessarily long horizon also accumulates estimation noise.

Table S3: **Resolution and horizon sensitivity.** Mean six-month AQ MAE over 20 development resplits at center A; the calibration participants and center B are not used to select a setting.

Resolution τ	FIR horizon: 40 s	FIR horizon: 48 s
0.1	6.67	6.66
0.25	6.46	6.49
0.5	6.60	6.60
1	6.80	6.79

Boundary conditions of the observation model. Gain invariance is defined for diagonal regional scaling of a nonzero integrated response. Response tails beyond the FIR horizon enter the error budget, and changes in task engagement act through C itself, not through a gain. Mixing across regional measurements beyond diagonal gain calls for a broader quotient. Testing these regimes apart from ordinary gain heterogeneity keeps the robustness claims aligned with the observation model.

E Anatomical interpretation and individual consistency

Anatomical allocation and individual consistency. To interpret the reserve of the full measured network, let $G_1, \ldots, G_6$ be its anatomical groups and $G_S = \bigcup_{j \in S} G_j$. The exact Shapley allocation (Shapley et al., 1953) of the reserve set function is

$$\phi_g = \sum_{S \subseteq [6] \setminus \{g\}} \frac{|S|!(5 - |S|)!}{6!} \big[c_\tau(G_{S \cup \{g\}}) - c_\tau(G_S)\big]. \tag{21}$$

The 63 existing queries suffice to evaluate all six allocations. They are nonnegative, sum to $c_\tau(X)$, and inherit regional-gain invariance. Thus $100\phi_g/c_\tau(X)$ describes each group's share of the measured reserve. We repeat this calculation within the response bootstrap. These allocations describe the reserve set function; clinical associations are evaluated independently.

Regional response diagnostics. To separate anatomical contribution from measurement precision, we also compute a leave-one-region-out diagnostic for $c_\tau(X) > 0$:

$$d_r = 100 \, \frac{c_\tau(X) - c_\tau(X \setminus \{r\})}{c_\tau(X)}. \tag{22}$$

This is the percentage of full-set reserve lost when a measured response row is removed. It is gain invariant and nonnegative, but is not an additive allocation. The response bootstrap supplies its interval width, and the two task runs supply its absolute-agreement reliability. We map these three quantities to fixed Schaefer parcel correspondences on matched fsLR32k surfaces using surfplot and BrainSpace (Gale et al., 2021; Vos de Wael et al., 2020). Clinical prediction continues to use the original 63 reserve coordinates.

E.1 Anatomical allocation and clinical adjustment

Equation (21) averages each group's marginal reserve over all possible insertion orders. Efficiency follows by telescoping the increments in every order; nonnegativity follows from monotonicity. Because every capacity coordinate is gain invariant, so is each allocation. We report participant-average shares and their 95% intervals from 2,000 participant bootstrap resamples. Within-participant allocation uncertainty is estimated by repeating normalization, query construction, and allocation across the 200 response bootstrap replicates.

Cortical values are assigned directly to the 18 matched Schaefer parcel labels. Surface display uses the paired Conte69/fsLR32k meshes and parcellation supplied together by BrainSpace; no continuous interpolation is applied to parcel labels. Atlas IDs, actual RAS centroids, and all regional metrics are tabulated in Appendix E.3. For the clinical association, the ordinary least-squares adjustment models for full-network reserve and six-month AQ are fitted independently using only the 116 fitting participants. They include an intercept, baseline AQ, age, and $\log(1 + \text{lesion volume})$. Their coefficients are frozen before computing center-B residuals. Spearman association and its participant-bootstrap interval use the 84 held-out residual pairs.

Figure S7 separates regional contribution, measurement precision, and cross-run reliability. Mean regional deletion loss spans 2.47–3.25% of full reserve; regional ICCs span 0.60–0.91. These finer-scale estimates complement the group-level ICCs in Figure 6b.

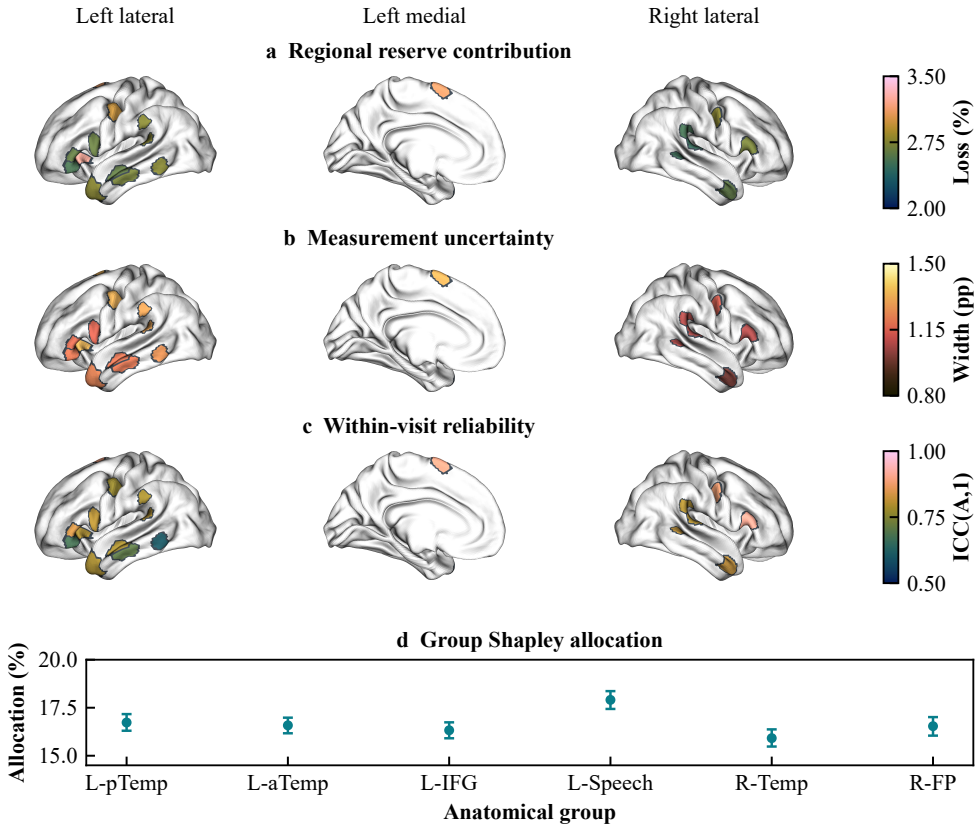


Figure S7: **Anatomical reserve, precision, and repeatability.** Fixed Schaefer parcel correspondences on Conte69/fsLR32k surfaces. (a) Mean regional deletion loss as a percentage of full reserve. (b) Mean width of the regional 95% response-bootstrap interval, in percentage points. (c) Regional ICC(A,1) across two runs. Gray cortex is outside the selected parcels; color has no significance threshold. (d) Group Shapley shares with 95% participant-bootstrap intervals. Maps use surfplot/BrainSpace and 84 held-out participants.

Table S4: **Clinical context for the matched-case display.** Full-network reserve is normalized by its known spectral maximum.

Case	Age	Baseline AQ	Lesion (mL)	Reserve	Six-month AQ
A	49	78.32	9.05	0.87	93.86
B	49	77.31	10.32	0.60	82.96

E.2 Cross-run reliability and individual reserve profiles

The two baseline task runs are analyzed on their own. Each profile contains all 63 nonempty output-group unions, normalized by their known spectral maxima. For each coordinate, we pool the two runs of the 116 center-A fitting participants to estimate the mean and population standard deviation, with a standard-deviation floor of 0.025. These fixed statistics standardize both center-B runs. To assess individual profile shape, we then subtract each standardized profile's coordinate mean before computing cross-run cosine similarity. This removes its mean standardized level from the matching statistic. The resulting 84×84 matrix compares each run-1 participant with every run-2 participant at center B, using the same row and column order (Figure S8a). Each row's maximum selects one candidate; top-1 matching is 70/84 (83.3%), as reported with the pairwise distributions in Figure 6c. Without profile centering, the same protocol identifies 79.8%. No outcome, clinical covariate, or unavailable-tissue mask is appended to the matching vector. Pairwise distributions are descriptive; the identification denominator is 84 patient queries. This extends the within-person comparison used in connectome fingerprinting (Finn et al., 2015) to a task-reserve representation.

For each anatomical group we calculate the single-measure, two-way absolute-agreement coefficient ICC(A,1), with 95% intervals from 1,000 participant bootstrap draws. The observed group range is 0.82–0.92. Since both runs belong to one visit, these estimates characterize within-visit consistency.

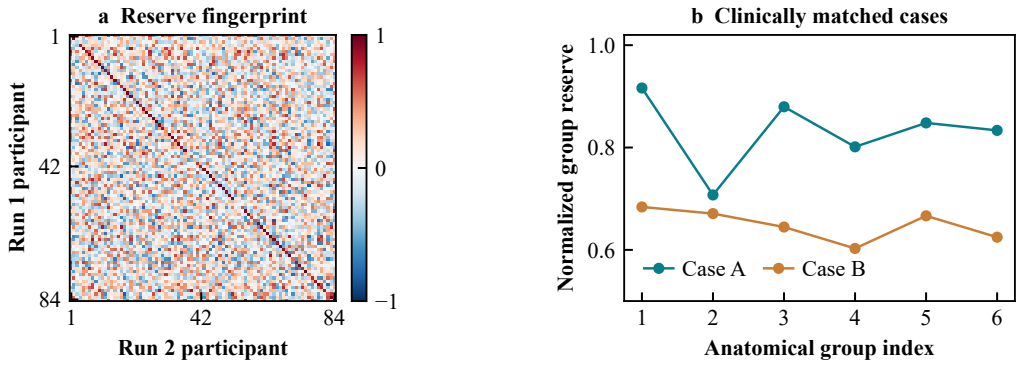


Figure S8: **Repeatability and clinically matched reserve profiles.** (a) Cross-run cosine similarity of centered, standardized reserve profiles; the diagonal compares the same patient. (b) Group reserve for two participants selected using baseline AQ, lesion volume, and reserve contrast, without their follow-up outcomes. Group indices follow Table S2. Table S4 supplies the clinical context.

Matched-case selection. Among center-B pairs, we require baseline AQ to differ by at most five points and lesion volume by at most 10 mL. The displayed pair has the largest full-network reserve contrast among those eligible pairs. Follow-up AQ is revealed only after selection. The pair illustrates a contrast within the cohort-wide analysis; it does not define a new patient subtype.

E.3 Atlas-resolved diagnostics

Fixed atlas correspondence and map semantics. The 18 anatomical anchors define a hemisphere-constrained, one-to-one nearest-centroid correspondence to the Schaefer2018 400-parcel, seven-network atlas. The assignment uses only anatomy and is fixed before outcome analysis. Table S5 records the parcel IDs and official FSL-MNI152 RAS centroid coordinates. The corresponding fsLR32k vertex labels and Conte69 surface meshes come from the same BrainSpace

Table S5: **Atlas correspondence and spatial diagnostics.** Centroids are official FSL-MNI152 RAS coordinates in millimeters. Loss is the mean percentage of full-set reserve removed with that row; width is the mean 95% bootstrap width in percentage points. All metrics use 84 held-out participants.

Region	Parcel	x	y	z	Loss (%)	Width (pp)	ICC
L-pSTG	37	-58	-36	16	2.77	1.30	0.80
L-pMTG	133	-60	-48	-10	2.75	1.26	0.60
L-SMG	95	-62	-36	34	2.75	1.30	0.78
L-aSTG	152	-56	-8	-14	2.63	1.19	0.79
L-aMTG	150	-62	-18	-20	2.71	1.19	0.71
L-TP	124	-44	12	-34	2.76	1.22	0.79
L-IFGop	105	-52	8	14	2.65	1.15	0.80
L-IFGtri	172	-48	28	0	2.64	1.17	0.83
L-IFGorb	170	-46	32	-10	2.54	1.19	0.68
L-vPreCG	47	-52	-6	44	2.97	1.33	0.77
L-aINS	99	-32	24	0	3.25	1.35	0.79
L-SMA	189	-6	10	64	3.10	1.33	0.89
R-STG	237	60	-24	10	2.47	1.06	0.80
R-MTG	373	64	-38	0	2.50	1.08	0.81
R-aTemp	367	50	8	-32	2.62	1.01	0.82
R-IFG	378	54	24	6	2.71	1.07	0.91
R-vPreCG	248	52	-6	38	2.77	1.14	0.88
R-SMG	297	58	-32	24	2.53	1.08	0.80

resource release. Values are assigned by parcel ID, preserving left/right order and vertex correspondence; there is no volume-to-surface interpolation or spatial smoothing.

Unmeasured parcels and the medial wall remain masked. Map boundaries identify the selected parcels; a nonzero map value is not a significance declaration. Each metric uses one fixed scale across all its views, and the matched cases share the same reserve scale. The left medial view displays the medial frontal output that is partly hidden in the lateral view. Actual inputs, engine versions, color limits, parcel correspondences, and surface arrays accompany the source package.

E.4 Nonredundant regional reserve and its precision

For the unit response direction v_r and full-set Gram matrix K_X, the matrix determinant lemma gives

$$c_\tau(X) - c_\tau(X \setminus \{r\}) = -\log\left[1 - v_r^\top (\tau I + K_X)^{-1} v_r\right]. \tag{23}$$

The quantity inside the logarithm is positive for $\tau > 0$. This expression is evaluated for every region and checked against direct deletion and recomputation. It measures the uniqueness of a response direction given all remaining measured outputs; redundancy means that these marginal contributions need not sum to the full reserve. It is distinct from the order-averaged, additive group Shapley allocation in Equation (21).

Each participant has 200 response-bootstrap realizations, with normalization and the full-set denominator recomputed in every realization. We display the participant-average width between the 2.5th and 97.5th percentiles, in percentage points of full reserve. Regional ICC(A,1) is computed from the two run-specific deletion profiles, with 2,000 participant-bootstrap intervals. This separates a region's contribution from its precision and its repeatability.

E.5 Individual profiles and acquisition quality

Figure S9 displays the same pair selected by baseline AQ, lesion volume, and reserve contrast in Table S4. Every parcel in a group carries that group's normalized reserve, and both patients use the same color limits. The participant-query heatmap retains all 84 held-out participants; its order depends only on full reserve. The acquisition-quality panel relates mean framewise displacement to mean regional bootstrap width ($\rho = 0.29$), providing a measurement interpretation for heterogeneous uncertainty. It does not select patients or tune the outcome head.

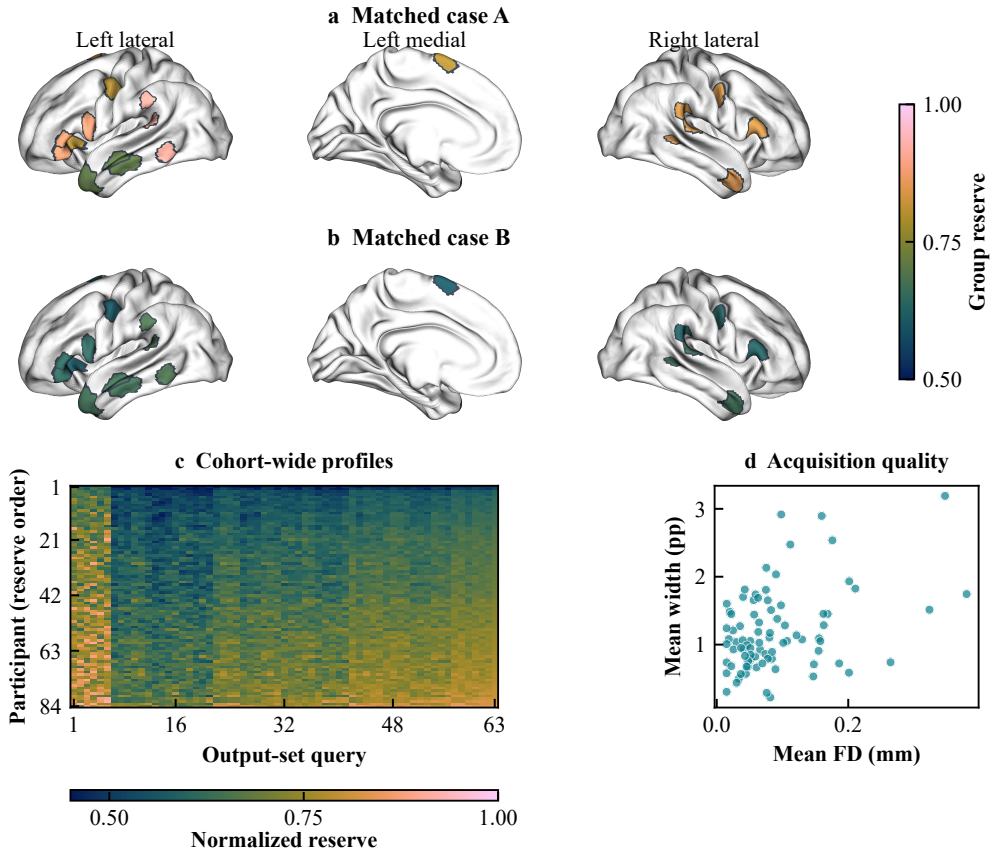


Figure S9: **Individual cortical profiles in the context of the full cohort.** (a,b) Matched cases A and B, selected using baseline information and reserve contrast, share an identical cortical mapping and color scale. Values are normalized group reserve, distinct from the deletion metric in Figure S7. (c) All 84 held-out participants and 63 output queries, ordered by full reserve and then by query cardinality/bitmask. (d) Mean regional uncertainty width versus mean framewise displacement. The case pair's clinical measurements appear in Table S4.

F UNAVAILABLE OUTPUTS AND RESERVE-COMPLETION BOUNDS

Let $O \subseteq X$ be available output regions and $m = |X \backslash O|$. Monotonicity and the per-region increment bound imply

$$c_\tau(O) \le c_\tau(X) \le \min\{c_\tau(O) + m \log(1 + \tau^{-1}),\ c_{\max}(X)\}. \tag{24}$$

All three quantities are divided by the same $c_{\max}(X)$, so the target remains the complete 18-region reserve. These are deterministic completion bounds conditional on the measured rows. Their coverage diagnostic is distinct from the nominal clinical prediction coverage.

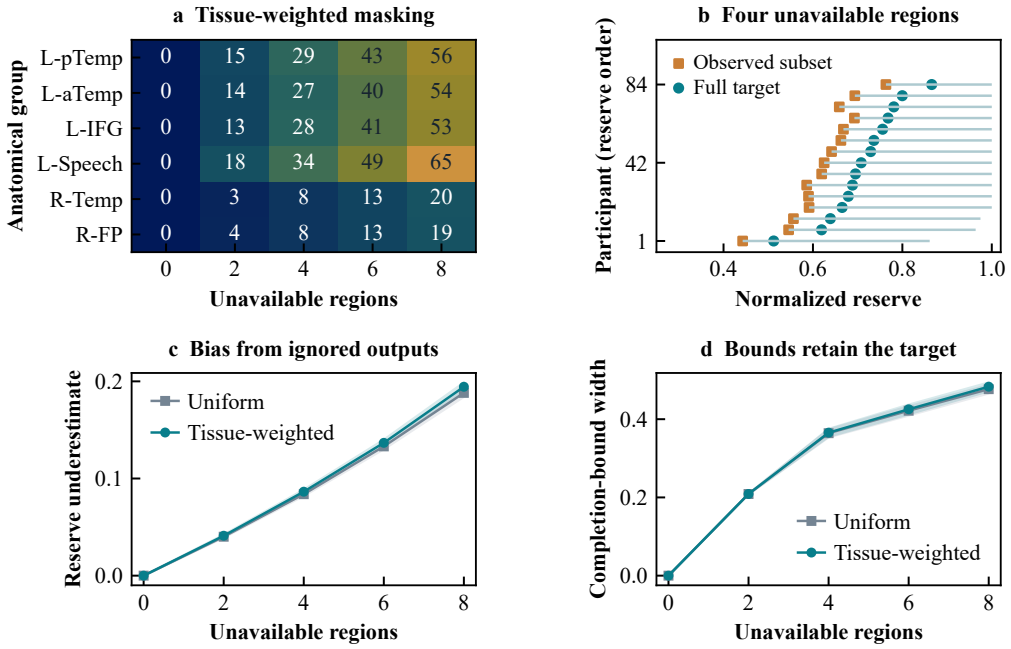


Figure S10: **Reserve uncertainty under unavailable regional measurements.** (a) Mean masking frequency within each anatomical group, in percent, for the tissue-weighted regime. (b) Completion intervals and observed-subset values for 15 patients sampled uniformly along full-reserve order, with four unavailable regions. The complete target is shown for evaluation. (c,d) Subset underestimation and completion width as the number of unavailable regions increases; shading is the 95% participant-bootstrap interval for each mean. The same full-network normalization is used in every panel.

For each of the 84 participants, we mask 0, 2, 4, 6, or 8 rows and sample 40 realizations at each count. Uniform masking is compared with probabilities proportional to $0.03 + (1 - f_{ir})$, where f_{ir} is the available-tissue fraction. Each realization removes access to regional measurements without altering the underlying response structure; it does not model a new lesion or change neural dynamics. Masks use neither the six-month outcome nor a prediction residual. We first average across realizations within each participant and then use 2,000 participant bootstrap resamples for the reported bands.

All complete-set targets lie inside Equation (24). Under tissue-weighted masking, the mean subset underestimate increases from 0.086 at four unavailable outputs to 0.194 at eight; the mean completion width grows from 0.366 to 0.483. Uniform masks give a similar trend. The experiment shows why availability must be retained explicitly: a missing measurement creates a range for the full-network target; it is not evidence of low reserve in the missing region.