SEMANTICBN: LINKING BRAIN-NETWORK UNITS TO EVIDENCE-GROUNDED DESCRIPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain-network representations for language interfaces need to preserve individual connectivity properties and make the evidence supporting each description explicit. We propose **SemanticBN**, a six-stage framework for constructing candidate brain semantic units and connecting them to evidence-grounded descriptions. Specifically, the framework combines three components: (1) self-supervised community learning forms individual soft communities and static unit vectors linked to spatial composition, within-unit connectivity, between-unit relations and available temporal records; (2) neuroscientific reference checks determine which interpretations the evidence supports and govern controlled description generation; and (3) brain–language alignment pairs unit vectors with these descriptions, while linked records support retrieval and evidence-based question answering. Developmental experiments on ADNI and major depressive disorder (MDD) separately evaluate community organization, personal-property preservation, language correspondence and evidence use. Anatomical interpretations are conditional on ROI-to-atlas provenance, which remains unverified in both cohorts. Balance regularization supports community utilization. Individualized pooling improves MDD joint two-property accuracy over shared assignments by 3.09 percentage points; fixed-reference pooling and full-connectivity regression remain stronger. In a separate information-matched, full-record experiment, MDD paraphrase brain-to-text accuracy is 7.91% with MiniLM, 12.73% with E5, 6.89% with BGE and 31.85% with symbolic targets and a supplied parser. Under archived permissions, full-context question answering achieves 92.22% correctness versus 64.44% with learned retrieval. These results support evaluating organization, property retention and language correspondence as distinct requirements for brain semantic units.

1 INTRODUCTION

Resting-state functional connectivity (FC) describes statistical associations between regional brain signals (Biswal et al., 1995). These associations form distributed networks and vary across people (Yeo et al., 2011; Finn et al., 2015; Gordon et al., 2017). A representation intended for language use must retain more than a list of connections: it needs an identifiable brain-network object, measurable properties and a defensible scope of interpretation.

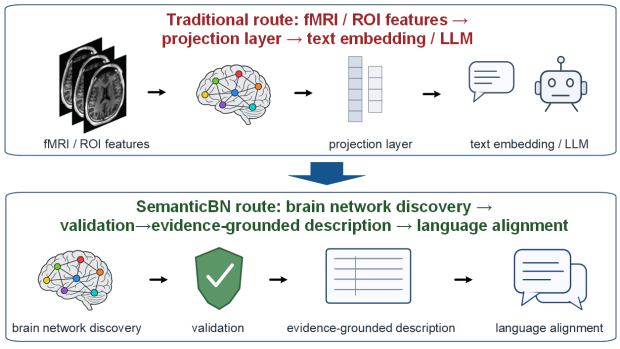


Figure 1: Scientific motivation: direct feature–language alignment (top) and candidate units supported by measurements and reference checks (bottom).

Our central question is: *What kind of brain-network representation can become a neurally meaningful brain semantic unit?* We study how learned community organization, personal connectivity measurements and neuroscience references can support controlled language expression (Figure 1).

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.

The six-stage route is *FC encoding → individual communities → candidate units and evidence → reference checks and permissions → individual descriptions → alignment, retrieval and verification* (Figure 2). Reference checks link measurable properties to supported descriptions. Alignment tests vector–description correspondence; downstream QA tests access to supported statements.

To test whether candidate representations preserve personal properties expressible in language, we further introduce common reference queries and information-matched static descriptions. This evaluation follows *static brain representation → information-matched personal-property descriptions → controlled brain–text evaluation*. Property readout is one test of this framework: the full-record protocol uses common targets and evidence-equivalent positives, while unit-level alignment pairs candidates with their own supported descriptions.

Our contributions are threefold. First, *candidate-unit organization and representation*: individual soft communities define analysis objects linking static vectors, measurements and descriptions. Second, *evidence-to-statement conversion*: spatial, connectivity and gradient records govern measured statements and conditional associations, recording missing evidence. Third, *layered evaluation*: common-property tasks, brain–text matching and evidence QA test information preservation, language correspondence and evidence use, respectively.

2 Related Work

Functional parcellations and gradients provide complementary descriptions of brain organization (Power et al., 2011; Yeo et al., 2011; Margulies et al., 2016). Individual mapping motivates subject-dependent organization (Gordon et al., 2017). Differentiable pooling and clustering learn graph assignments (Ying et al., 2018; Bianchi et al., 2020). An FC-row encoder with soft-community pooling supplies one family of representations in our controlled comparison.

BrainLM and BrainGFM study pretrained brain representations (Ortega Caro et al., 2024; Wei et al., 2026a); semantic decoding and language-consistent tokenization address brain–language correspondence (Tang et al., 2023; Wei et al., 2026b). Our targets are controlled descriptions of measured network properties. Contrastive alignment supplies a readout of these properties (Radford et al., 2021; Reimers & Gurevych, 2019), enabling comparisons with numerical and symbolic targets derived from the same measurements.

3 SemanticBN Framework

Figure 2 links six stages: FC features and soft assignments form candidates; measurements and reference checks govern their descriptions; alignment, retrieval and verification connect them to language. We describe each transition below and define its evaluation separately in Section 4.

3.1 Individual Functional Networks

For participant s, $X^{(s)} \in \mathbb{R}^{T_s \times R}$ contains T_s samples from R regions of interest (ROIs). Signed full-scan Pearson FC is

$$F_{ij}^{(s)} = \operatorname{corr}\big(X_{:i}^{(s)}, X_{:j}^{(s)}\big). \tag{1}$$

Nodes correspond to declared ROI columns and edges to statistical associations. Scan identifiers, atlas/order provenance and available preprocessing metadata accompany each matrix. ROI compatibility requires column-level provenance. Analysis starts from the supplied derivatives. A shared FC-row MLP maps each row to the 64-dimensional feature h_i in Equation 2.

3.2 Individual Soft-Community Learning

The same encoder processes every participant but produces a different assignment from each person's FC:

$$h_i^{(s)} = \mathrm{LN}\Big(\mathrm{GELU}(W_h F_{i:}^{(s)} + a_h)\Big), \qquad Q_{i:}^{(s)} = \mathrm{softmax}(W_q h_i^{(s)} + a_q). \tag{2}$$

The hidden dimension is 64 and $K = 7$. Each row of $Q^{(s)}$ sums to one, allowing graded participation across communities. This stage uses FC-only supervision. Language and common-property

experiments fix participant splits before fitting; the organization-only runs fit and evaluate complete cohorts.

The archived reconstruction is $\widehat{A} = QQ^\top$, with target

$$A_{ij} = \frac{[F_{ij}]_+ \mathbf{1}[i \neq j]}{\max(\max_{u \neq v}[F_{uv}]_+, \epsilon)}.$$

Writing $d = A\mathbf{1}$ and $2m = \mathbf{1}^\top d$, the per-scan losses are

$$L_{\text{rec}} = R^{-2}\|A - QQ^\top\|_F^2, \qquad L_{\text{bal}} = K^{-1}\sum_k \left(R^{-1}\sum_i Q_{ik} - K^{-1}\right)^2,$$

$$L_{\text{ent}} = -R^{-1}\sum_{i,k} Q_{ik}\log Q_{ik}, \qquad L_{\text{mod}} = -(2m)^{-1}\operatorname{Tr}\left[Q^\top\left(A - \frac{dd^\top}{2m}\right)Q\right]. \tag{3}$$

Training minimizes $L_{\text{rec}} + 5L_{\text{bal}} + .02L_{\text{ent}} + .05L_{\text{mod}}$, averaged over scans. Balance promotes community use, entropy sharpens memberships and modularity rewards within-community structure (Newman, 2006). The encoder reads signed FC; reconstruction targets positive off-diagonal edges with a zero diagonal. The latter conflicts with $(QQ^\top)_{ii} \geq 1/K$. The language study freezes the original community fits; comparing original, off-diagonal-only and reconstruction-free objectives throughout the downstream pipeline remains future work (Supplementary S7).

3.3 Candidate Brain Unit Construction

For community indices k, l and $w_{ij}^{kl} = Q_{ik}Q_{jl}\mathbf{1}[i \neq j]$, the signed block mean is

$$a_{kl} = \frac{\sum_{i,j} w_{ij}^{kl} F_{ij}}{\sum_{i,j} w_{ij}^{kl}}. \tag{4}$$

ROI self-pairs are excluded even between overlapping communities. Zero effective mass is recorded as missing.

The same memberships pool ROI features into a static unit vector,

$$b_k^{\text{static}} = P_b\left(\frac{\sum_i Q_{ik}h_i}{\sum_i Q_{ik}}\right) \in \mathbb{R}^{64}. \tag{5}$$

Here P_b is the identity map. The candidate record is

$$U_k^{(s)} = \left(b_k^{(s)}, E_k^{(s)}, S_k^{(s)}\right), \tag{6}$$

where E stores measurements, reference checks and availability; S contains permitted descriptions (Section 3.5). Four linked records accompany b_k (S6): *spatial composition*, ROI indices, available names, membership and normalized pooling weights; *within-unit FC*, weighted mean a_{kk}, edge SD and available training-control reference deviation; *between-unit FC*, signed profiles and supported partner/reference comparisons; *temporal measurements*, window-wise changes when time series are available. The learned vector is static; the four measurement groups remain linked records. Temporal summaries hold full-scan Q fixed and use windows in samples. Scan, ROI-order and checkpoint provenance remain attached.

3.4 Neuroscientific Reference Validation

Comparisons. Spatial checks compare weighted ROI composition with reference networks; connectivity checks compare unit and matched-reference profiles across six non-self networks within the same scan. Gradient records contain position, coverage, dispersion and compactness. Stability and disease effects extend these to five complementary check categories. References are excluded from community fitting; the complementary checks share data and anatomy, so their outcomes can be statistically dependent. S6 gives definitions and thresholds; gradient permutations remain spatially unconstrained (Burt et al., 2020).

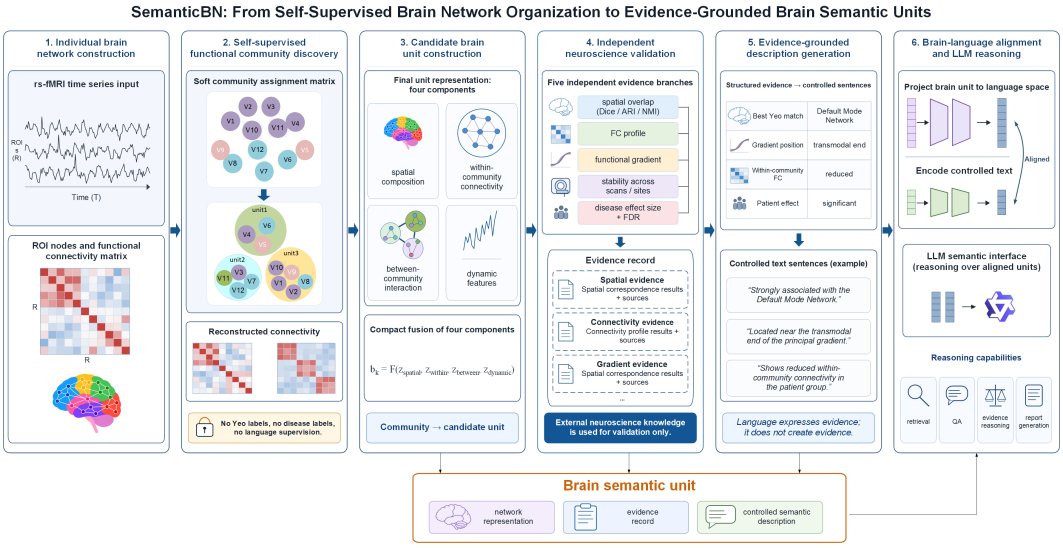


Figure 2: Six-stage SemanticBN framework. Self-supervised communities from individual FC define candidate units. Unit representations link to measurements and reference checks that govern controlled descriptions. Brain–language alignment supports unit matching; linked records support language access. Experiments separately evaluate organization, common-property recovery, language correspondence and record-based QA, specifying each representation, target and readout. Evaluated vectors are static; temporal measurements remain in linked records.

Evidence records. E_k links each check's result, reference source, interpretive scope and missing reason. Description compilation (Section 3.5) reads the resulting claim-specific permissions and denial reasons.

Description permissions. Available measurements support numerical descriptions; network associations require spatial and functional checks plus verified ROI binding. Stability requires matched repeat measurements; group differences require group statistics. Unsupported interpretations are withheld while supported measurements remain available. Current review supersedes ADNI's old verified flag: neither cohort has confirmed ROI-column–atlas binding. Anatomical and gradient permissions remain withheld; stability and disease validators remain unattached.

3.5 Evidence-Grounded Description Generation

Given E_k, the policy in Section 3.4 selects supported claim types c:

$$\mathcal{A}_k = \{c : \mathrm{Permit}(c, E_k) = 1\}, \qquad S_k = \mathrm{Verbalize}(E_k, \mathcal{A}_k).$$

These expressions specify the deterministic description compiler (S6). Type-specific templates cover spatial composition, within- and between-unit FC, training-control deviations, temporal summaries and conditionally authorized network/gradient statements. Compiled claims retain identifiers linked to source paths and values; withheld types retain denial reasons. S6 details source binding and recompilation checks.

Information-matched evaluation selects three static attribute types for common-reference descriptions S_a^{test}, with identical target content across representations (Section 4.4). These are distinct from discovered-unit evidence descriptions S_k.

3.6 Brain–Language Alignment and Evidence-Based Access

The archived connected trial (S6) pairs frozen 64-dimensional units with descriptions permitted by its original policy:

$$z_k = \mathrm{norm}\big(P_\psi\big(\,\mathrm{Standardize}_{\mathrm{train}}(b_k)\big)\big)\,, \qquad e_k = \mathrm{norm}(T(S_k)).$$

Here P_ψ is a trainable $64 \to 384$ linear projection with bias, T frozen MiniLM (Wang et al., 2020), and $\mathrm{norm}\ \ell_2$ normalization. Standardization uses training-only statistics. Symmetric contrastive training in both directions uses exact duplicate descriptions as multiple positives; only P_ψ is updated.

For access, T embeds a question to rank z_k within its scan. Retrieved records supply allowed statements and evidence to frozen Qwen2.5-7B for selection of claim and evidence identifiers. Its input consists of serialized evidence records. The verifier checks target unit, claim type, permissions and citations before rendering permitted wording, refusal or rejection. This evaluates controlled statement selection; explicit unit/field requests allow direct lookup.

The revised permissions have not been rerun through alignment or QA. Full-record evaluation (Section 4.4) uses equivalent attribute profiles as positives; it forms a separate evaluation branch.

4 EXPERIMENTS

We evaluate four aspects in order: community organization, personal-property preservation, language correspondence and downstream evidence use. The unit-level protocol (S6) and common-target full-record protocol (S2, S5) retain their own targets, candidate banks and positive definitions.

4.1 DATA AND PROTOCOLS

The community archive contains 442 ADNI scans from 130 people and 2,018 retained REST-meta-MDD participants (Mueller et al., 2005; Yan et al., 2019; Chen et al., 2022). **Both ADNI and MDD lack verified ROI-column–atlas binding.** ADNI's earlier verified flag was a local declaration; neither cohort has a reviewed link between extraction records, atlas checksums and column-bound ROI labels. The numerical comparisons condition on consistent column identity within each cohort and fixed reference weights across methods. Their outcomes support conditional numerical comparisons, with anatomical labels and spatial/gradient scores interpreted as provisional reference agreement. Access rights and secondary-use clearance remain unconfirmed and block submission (S1, S4). ABIDE and epilepsy files with unknown order are excluded. The separate 116-column disease study retains 1,976 participants.

The controlled static study uses one first archived scan per person: ADNI train/validation/test counts are 91/19/20 and MDD counts are 1,412/302/304. It reuses three frozen community/PCA fits per cohort from the common-property benchmark (seeds 7/11/19), fitted separately within each cohort. Community discovery uses AdamW at .001, weight decay .0001, batch size 256 and 20 epochs. Each text readout uses the same optimizer settings, batch size 64, at most 60 epochs and canonical-validation patience eight. Six trained conditions, two cohorts and three seeds give 36 new fits; mean-state and other-person controls require no extra fit. Protocol and source hashes were frozen before this run, but the participants had already been inspected in earlier experiments.

All results are developmental. B→T ranks descriptions from brain inputs; T→B reverses this direction. For full-record text matching, primary outcomes use the held-out paraphrase template. Scores average seven references and three seeds within person before 10,000 paired sign flips and 5,000 participant bootstraps. Benjamini–Hochberg (BH) correction (Benjamini & Hochberg, 1995) covers 28 contrasts: individualized versus seven alternatives, two directions and two cohorts. Intervals condition on fixed models and candidate banks; shared-bank dependence and model-refitting uncertainty fall outside their coverage. Supplementary S2 gives implementation details, canonical scores and all paired contrasts.

4.2 COMMUNITY ORGANIZATION AND REFERENCE CHECKS

We report adjusted Rand index (ARI), one-to-one matched ROI overlap (Hubert & Arabie, 1985) and coverage-weighted gradient dispersion (Margulies et al., 2016). They instantiate the organization checks in Section 3.4, conditional on provisional column-to-atlas mapping.

Table 1 and Figure 3 use pooled ADNI/MDD fits and same-cohort evaluation (seeds 7/17/27), separate from the cohort-specific static fits. SemanticBN metrics use $\arg\max_k \bar{Q}_{ik}$, where $\bar{Q}$ averages scans within each cohort; repeated ADNI scans receive scan-level weighting. BrainLM-style and BrainGFM-style are locally trained architecture-inspired controls. S3 specifies their differences

Table 1: Conditional within-cohort template comparisons (three-seed means). Both cohorts' column-to-atlas mappings remain unverified. ARI and overlap are higher-is-better; gradient dispersion is lower-is-better. These are provisional reference-agreement metrics; full SDs and aggregation rules are in S3.

	ADNI			MDD		
Method	ARI	Overlap	Grad.	ARI	Overlap	Grad.
Random	.011	.282	.234	.011	.282	.234
Spectral	.222	.447	.153	.388	.588	.145
K-means FC	.264	.471	.136	.442	.612	.114
NMF	.057	.329	.177	.426	.647	.132
BrainLM-style	.270	.471	.126	.403	.584	.104
BrainGFM-style	.130	.361	.198	.244	.471	.149
SemanticBN	.293	.510	.147	.426	.627	.122

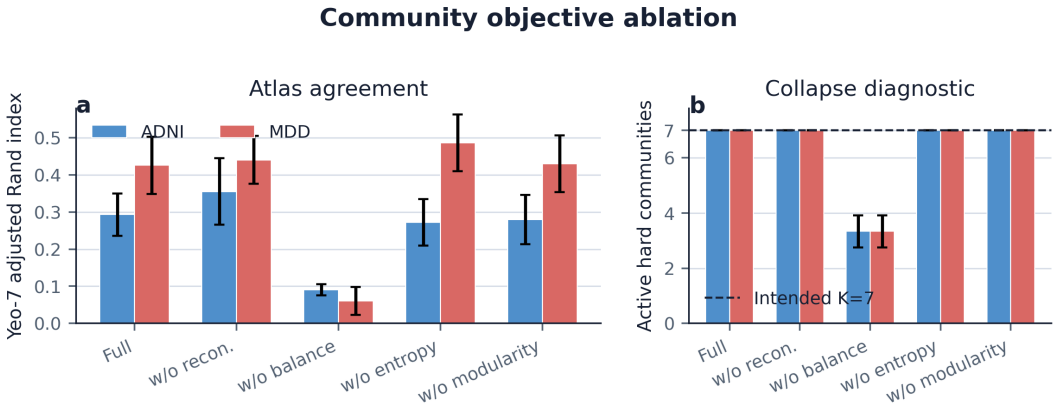


Figure 3: Within-cohort objective ablation. Active counts nonempty labels in the hard cohort-average assignment. Error bars are seed SD. Removing reconstruction improves template ARI; anatomical correspondence remains provisional.

from official pretrained models. Atlas scores remain conditional on the source-order verification described above.

Removing balance reduces each cohort template's active count from 7.00 to 3.33 across seeds and ADNI/MDD ARI from .293/.426 to .090/.061 (Figure 3). Removing reconstruction increases ARI to .355/.441, and removing entropy gives .272/.487. Balance supports community utilization at the cohort-template level. This organization-only ablation evaluates the loss components; downstream measurements, descriptions and alignment retain the original community fits (Supplementary S3, S7).

4.3 Personal-Property Preservation

Common numerical targets. All representations share fixed reference weights $Y \in \mathbb{R}_+^{R\times 7}$, treated as numerical column weights under the stated provenance assumptions. The common signed FC block means are $G_{ac} = \sum_{i\neq j} Y_{ia}Y_{jc}F_{ij} / \sum_{i\neq j} Y_{ia}Y_{jc}$. For each reference set a, the three target properties are

$$y_a = (G_{aa},\ G_{a,a_1} - G_{a,a_2},\ \mathrm{SD}_{c\neq a}\, G_{ac})\,. \tag{7}$$

For $a \in \{1,\dots,7\}$, partners are $a_j = 1 + ((a-1+j) \bmod 7)$, $j = 1, 2$. SD uses divisor six over external partners. Gold is computed from F and fixed reference weights Y, independently of learned Q: the targets are indexed by reference set a. For each cohort, reference and attribute, the $1/3$ and $2/3$ quantiles of *all training participants*, pooling diagnostic groups, define low/middle/high bands $g_a \in \{0,1,2\}^3$. The band reference includes both patient and control groups in the training cohort.

Table 2: Common-source joint two-property accuracy and structured top-1 retrieval (%, three-seed means). All methods use the same gold attributes. Complete SDs, ADNI retrieval and additional banks are in Supplementary S5.

	ADNI joint	MDD joint	MDD B→T	MDD T→B
Individualized	41.19	69.03	81.55	84.63
Shared	39.52	65.95	77.40	79.65
Fixed reference	42.14	71.57	83.74	86.75
Individual K-means	45.48	67.15	78.73	81.39
Individual spectral	43.33	62.73	72.24	76.11
FC-PCA/MLP	48.81	69.17	68.75	75.38
Raw-FC ridge	62.14	92.58	81.16	86.61

Matched reference-query records. For fair cross-partition comparison, supplied reference index a differs from learned-community index k. With the same weights Y as the common targets, r_a is a reference-weighted mixture of multiple discovered units b_k. Concatenating seven queries yields the evaluation record:

$$r_a = \sum_k \alpha_{ak} b_k^{\text{static}}, \qquad \alpha_{ak} = \frac{\sum_i Y_{ia} Q_{ik}}{\sum_{i,k'} Y_{ia} Q_{ik'}}, \qquad r = [r_1; \ldots; r_7] \in \mathbb{R}^{448}. \tag{8}$$

This readout is invariant to community-index permutations. Individualized, training-mean shared and fixed-reference assignments reuse the same frozen ROI encoder within each cohort and seed. FC-PCA compresses all 4,005 off-diagonal FC edges to at most 448 training-fitted components, with zero padding when rank is lower. This matches output dimensionality while allowing each representation to retain different connectivity information.

Numerical readout. Separate $448 \to 128 \to 21$ GELU heads predict seven references' three numerical attributes, using training-only feature/target standardization and validation-MSE stopping (up to 200 epochs, patience 20; S5). They reuse the cohort/seed source fits; per-scan clustering adds controls.

The property-joint accuracy in Table 2 requires correct internal-FC and external-contrast bands. Structured retrieval compares predicted numerical attributes with explicit attribute descriptions in 16-person, same-reference banks; all candidates sharing three bands are positive. Mean positive fractions are 10.58% (ADNI) and 10.44% (MDD). Raw-FC ridge is a full-information linear control; exact source aggregation supplies a tool upper bound.

After averaging queries and seeds within person, MDD individualized-minus-shared joint two-property accuracy is 3.09 percentage points (95% bootstrap CI [1.90,4.31], BH $q = .00019$ across the 48-comparison family). Differences use unrounded scores. ADNI's corresponding difference is 1.67 points ([-1.67,5.24], $q = .449$). The gain concerns MDD property preservation relative to shared pooling. Fixed-reference MDD accuracy is 2.54 points higher than individualized, and raw-FC ridge achieves the highest joint attribute accuracy in both cohorts.

Restricted-input follow-up. With matched parameter counts, a single 64-dimensional reference mixture gives MDD joint accuracy 70.08%, versus 69.03% for all seven queries. A training-matched discovered community gives 57.61%, versus 67.50% for all seven communities: -9.88 points, 95% CI $[-11.39, -8.33]$, $q = .00060$ (S2). This tests pooled-vector access to common-reference attributes; each ROI encoder still reads whole-brain FC.

Supporting tasks. Dynamic augmentation changes MDD category-paraphrase retrieval from .9522 to .9518 and diagnosis balanced accuracy from .5759 to .5626; adjusted dynamic group tests are nonsignificant. Separate disease identification yields AUROC .5477 versus .6663 for raw-FC logistic regression (S8).

4.4 Brain–Language Correspondence and Controls

Unit-level description retrieval. The matched follow-up to Section 3.6 uses mixed descriptions under archived permissions and exact-text targets. With 2,128 MDD candidates and exact-text pos-

Table 3: Full-record static-property paraphrase retrieval (%, mean ± seed SD). Every output can use all 448 input features; the supplied reference index selects a 20-person ADNI or 304-person MDD bank. Equivalent three-band profiles are positive. Symbolic uses the same individualized record and a supplied parser. Canonical scores and paired contrasts are in S2.

	ADNI		MDD	
Condition	B→T	T→B	B→T	T→B
Individualized	12.86 ± 3.11	15.00 ± 2.86	$7.91 \pm .33$	4.90 ± 2.07
Shared	15.71 ± 2.58	16.19 ± 2.18	$7.69 \pm .47$	6.27 ± 1.04
Fixed reference	15.24 ± 2.89	14.52 ± 2.97	$7.66 \pm .14$	5.95 ± 2.14
FC-PCA	6.19 ± 1.09	9.05 ± 4.19	$6.23 \pm .41$	$4.23 \pm .90$
Symbolic control	$25.24 \pm .41$	28.81 ± 3.22	31.85 ± 1.12	50.94 ± 2.28
Shuffled training text	10.24 ± 3.93	7.62 ± 2.30	$4.23 \pm .82$	4.84 ± 2.00
Training-mean state	7.38 ± 1.09	$9.36 \pm .00$	$5.34 \pm .38$	$4.71 \pm .00$
Other person's state	$6.43 \pm .71$	7.86 ± 3.11	$4.31 \pm .57$	$4.61 \pm .90$

itives, individualized B→T/T→B is .75/.66%, versus 1.61/3.23% for FC-MLP: means over three alignment initializations with seed-7 communities fixed (S6).

Controlled descriptions. Define $S_a^{\text{test}} = \text{Template}(g_a)$; the superscript labels the controlled-description task. The canonical wording is *Relative to training tertiles, internal FC is low; external FC contrast is middle; across-partner FC dispersion is high.* A second template changes sentence structure and band wording. Both grammars round-trip all 27 discrete profiles exactly. The benchmark text excludes continuous values, network names, participant IDs, diagnoses, gradients and temporal measurements. The bank supplies reference identity; S_k retains its separate evidence-permission policy.

Full-record alignment and scoring. To control readout capacity, conditions freeze source features and train separate $448 \to 128 \to 7 \times 384$ GELU readouts with identical architecture and per-reference normalization (404,224 parameters each). Every z_a reads $r = [r_1; \ldots; r_7]$: all seven queries can inform every prediction. Inputs use training statistics; MiniLM encodes canonical S_a^{test} into normalized 384-dimensional targets (Wang et al., 2020). This tests property correspondence at the record level. Unlike exact-text positives in Section 3.6, positives here share the same three attribute bands.

For participant s, reference a, and $\mathcal{P}_a(s) = \{t : g_a^{(t)} = g_a^{(s)}\}$, training averages the following loss over references and both directions:

$$L_{B\to T}^{a} = -\frac{1}{N}\sum_{s} \log \frac{\sum_{t\in\mathcal{P}_a(s)} \exp(z_{sa}^{\top} e_{ta}/\tau)}{\sum_{t} \exp(z_{sa}^{\top} e_{ta}/\tau)}, \qquad \tau = .07, \tag{9}$$

where N is minibatch size and $e_{ta} = \text{norm}(T(S_a^{\text{test},(t)}))$ uses canonical text. Batches contain one cohort; canonical validation loss selects stopping. The supplied reference index a selects both output slot z_a and candidate bank; interpreting a natural-language request to identify a is outside this task. Banks contain all held-out people at that reference: 20 for ADNI, 304 for MDD. Equal profiles count as positives across people and templates. Scores within 10^{-7} of the maximum receive uniform expected top-1 credit, independent of candidate order. Chance is the mean positive fraction.

Further controls shuffle training text across people, replace test features by the training mean, or substitute another person's test representation.

Table 3 tests S_a^{test} matching from full records, conditional on reference a. The 20/304-person banks include repeated descriptions, with 11–16/26–27 unique profiles per reference for ADNI/MDD. Expected accuracy from uniformly selecting a candidate is 9.36/4.71%.

Individualized MDD B→T exceeds shuffled text by 3.68 points (95% CI [2.60,4.76], $q = .00047$) and another person's state by 3.60 points ([2.63,4.62], $q = .00047$), supporting a contribution from personal information. Relative to shared assignments, the differences are +.22 points in B→T ($q = .560$) and -1.36 in T→B ($q = .00245$). Shared assignments also have higher point estimates in both ADNI directions.

Three-encoder cohort follow-up. With fixed brain inputs and optimization, 72 fits compare four representations, three encoders and three seeds in both cohorts across five paraphrase variants (S2). On the original MDD paraphrase, individualized B→T/T→B is 7.91/4.90% (MiniLM), 12.73/9.01% (E5) and 6.89/5.26% (BGE). E5 exceeds MiniLM by 4.82/4.10 points (both BH $q = .00057$); its individualized advantage over shared pooling is direction-dependent.

4.5 Encoding and Optimization Diagnostics

Target-encoding control. The symbolic control maps three bands to 27 fixed orthonormal 384-dimensional targets, keeping brain features, head, optimizer and loss fixed. Its supplied template parser recovers bands exactly. With individualized features, MDD B→T/T→B improves by 23.93/46.04 points (both $q = .00047$). This contrast changes both target geometry and parsing.

Text-target correspondence before brain readout. Over 27 equally weighted profiles, paraphrase-to-canonical top-1 is 7.41/14.81/7.41% for MiniLM/E5/BGE. Substituting correct canonical targets for brain outputs preserves cohort banks, duplicates and all equivalent positives. MiniLM canonical-to-paraphrase / reverse scores are 11.43/17.14% (ADNI) and 2.87/8.60% (MDD), versus 100% canonical self-matching. These gold-dependent substitutions isolate wording sensitivity (S2).

Encoding and optimization effects. Public true-attribute-code probes compare MiniLM, E5 and BGE (Wang et al., 2022; Beijing Academy of Artificial Intelligence, 2023). MiniLM canonical accuracy rises from 22.40% ($\tau = .07$, 60 epochs) to 95.24% ($\tau = .005$, 240 epochs); paraphrase accuracy remains 3.70%. A parser/geometry factorial separates their effects (S2). On unseen attribute combinations, E5 scores 34.92% versus 14.29% for symbolic targets.

4.6 Evidence-Based Question Answering

Full seven-unit context yields 92.22% E2E correctness versus 64.44% for learned top-three retrieval on 180 requests using the matched seed-7 index and archived permissions (S6); always refusing yields 41.67%. The original pilot's 70.00% uses a different projection/run. Correctness requires exact claims and citations or correct required refusal under the archived policy. Explicit unit/field requests can query authorized records directly; property retrieval is evaluated separately.

Common-property Qwen QA gives .42% individualized joint correctness at .70% measurement coverage on MDD, and zero joint correctness on ADNI (S5).

Structured, provenance and full dual-evidence checks make identical acceptance decisions on unmodified executor outputs (S9). Additional rejection beyond provenance occurs in constructed delivery or wording faults. Earlier direct-selection filtering trades coverage for lower selective risk. These checks remain auxiliary.

5 Conclusion

SemanticBN uses individual functional communities as analysis units, linking network representations, measurement evidence and controlled descriptions in one record. Balance supports community utilization; individualized pooling improves MDD two-property accuracy over shared pooling, while fixed-reference and full-FC controls remain stronger. Personal state aids matching; partition rankings vary across cohorts and retrieval directions. Symbolic targets with a supplied parser outperform MiniLM in this protocol. Full-context QA outperforms learned retrieval.

Scope and next steps. These developmental findings use previously inspected cohorts and supplied reference identity. Follow-ups cover three coarse attributes, three encoders, five wording variants, public codes and gold-target substitution. Single-unit sufficiency depends on input scope: matched communities and reference mixtures retain different information. Functional meaning requires verified ROI provenance, appropriate gradient nulls, repeatability and independent description assessment (S7). Temporal fusion, Qwen diagnostic fine-tuning and open-ended reports call for separate evaluations.

AI USE STATEMENT

OpenAI Codex assisted conceptual development, experiment design, mathematical diagnostics, implementation, data-processing code, experiment execution, literature searches, interpretation, drafting, translation and checking. Evaluation questions and synthetic query fixtures were agent-authored. The conceptual illustration was supplied by the user; the framework diagram adapts an earlier project figure with AI-assisted label edits. Result plots use recorded numerical results. Qwen is an evaluated model in the supporting studies. Human authors are responsible for data handling, analyses, citations and conclusions; final human review and approval remain unconfirmed.

REFERENCES

Beijing Academy of Artificial Intelligence. bge-small-en-v1.5: Model card, 2023. URL https://huggingface.co/BAAI/bge-small-en-v1.5.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 874–883, 2020. URL https://proceedings.mlr.press/v119/bianchi20a.html.

Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995.

Joshua B. Burt, Markus Helmer, Maxwell Shinn, Alan Anticevic, and John D. Murray. Generative modeling of brain maps with spatial autocorrelation. *NeuroImage*, 220:117038, 2020. doi: 10.1016/j.neuroimage.2020.117038.

Xiao Chen, Bin Lu, Hui-Xian Li, Xue-Ying Li, Yu-Wei Wang, Francisco Xavier Castellanos, Li-Ping Cao, Ning-Xuan Chen, Wei Chen, Yu-Qi Cheng, et al. The DIRECT consortium and the REST-meta-MDD project: towards neuroimaging biomarkers of major depressive disorder. *Psychoradiology*, 2(1):32–42, 2022. doi: 10.1093/psyrad/kkac005.

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.

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, Jacqueline M. Hampton, Rebecca S. Coalson, Annie L. Nguyen, Kathleen B. McDermott, Joshua S. Shimony, Abraham Z. Snyder, Bradley L. Schlaggar, Steven E. Petersen, Steven M. Nelson, and Nico U. F. Dosenbach. Precision functional mapping of individual human brains. *Neuron*, 95(4):791–807, 2017.

Keqi Han, Yao Su, Lifang He, Liang Zhan, Sergey Plis, Vince Calhoun, and Carl Yang. Rethinking functional brain connectome analysis: do graph deep learning models help. *npj Artificial Intelligence*, 2:19, 2026. doi: 10.1038/s44387-025-00067-x.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075.

Daniel S. Margulies, Satrajit S. Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M. Huntenburg, Georg Langs, Gleb Bezgin, Simon B. Eickhoff, F. Xavier Castellanos, Michael Petrides, Elizabeth Jefferies, and Jonathan Smallwood. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579, 2016.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005. doi: 10.1016/j.jalz.2005.06.003.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103.

Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed Rizvi, Matteo Rosati, Christopher Averill, James Cross, Prateek Mittal, Emanuele Zappala, Rahul Dhodapkar, Chadi Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RwI7ZEfR27.

Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Barnes, Jessica A. Church, Alecia C. Vogel, Timothy O. Laumann, Fran M. Miezin, Bradley L. Schlaggar, and Steven E. Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. URL https://arxiv.org/abs/2212.03533.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Lifang He, and Yu Zhang. A brain graph foundation model: Pre-training and prompt-tuning across broad atlases and disorders. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=PeGHkAaRxs.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D. Calhoun. fMRI-LM: Towards a universal foundation model for language-aligned fMRI understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6931–6940, 2026b.

Chao-Gan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tong-Jian Bai, Qi-Jing Bo, Jun Cao, Guan-Mao Chen, Ning-Xuan Chen, Wei Chen, et al. Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proceedings of the National Academy of Sciences*, 116(18):9078–9083, 2019. doi: 10.1073/pnas.1900390116.

B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zollei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/e77dbaf6759253c7c6d0efc5690369c7-Abstract.html.

SUPPLEMENTARY MATERIAL

S1 records data and metric definitions. S2 supplies the static alignment protocol, complete paired family and encoding diagnostics; S3–S5 cover organization, provenance and common properties. S6 documents unit-level alignment and record QA, and S7 states structural limits. S8 summarizes earlier supporting experiments; S9 retains the auxiliary query audit. Each study keeps its original targets and evaluation scope.

S1 DATA, METRICS AND ARTIFACT MAP

Cohorts and source status. The supplied MDD derivatives contain 2,034 time series: the historical 90-column route retains 2,018 people, while the separate 116-column disease route retains 1,976 (1,077 MDD, 899 controls) after excluding 58 constant-ROI cases. The ADNI community archive contains 442 scans from 130 people; unit-description studies retain one archived scan per person. ABIDE (1,035 records) and epilepsy (49 files) are excluded. These are local inventory counts. Both included cohorts lack verified column-to-atlas binding (S4).

The files were supplied as REST-meta-MDD/DIRECT and ADNI derivatives (Yan et al., 2019; Chen et al., 2022; Mueller et al., 2005). The DIRECT agreement and ADNI agreement describe access, attribution and publication obligations; authors must establish which signed terms apply to the actual transfer and intended use. Project-specific access/transfer authorization, applicable secondary-use approval or exemption, consortium attribution/publication review and final all-author approval remain unconfirmed in the materials reviewed. This is an outstanding verification step, not a finding that authorization is absent. Submission and external distribution remain blocked pending documented author confirmation. Extraction records, atlas checksums and column-bound labels require a separate scientific-provenance check; private approval documents stay outside the anonymous package.

Metric definitions. ARI is chance-adjusted Rand agreement with Yeo hard labels. Overlap (Ov.) is the fraction of eligible ROIs matched correctly under optimal one-to-one label assignment. Grad. is coverage-weighted within-community gradient SD; Active counts nonempty labels in the hard cohort-average template (S3). Spatial interpretations remain conditional on S4's provenance status.

B→T and T→B are top-1 retrieval unless otherwise specified. Full-record text matching (S2), structured common-property matching (S5) and unit-description matching (S6) retain different targets, banks and positive rules. Property-joint accuracy requires both internal-FC and external-contrast tertile bands to be correct. In record-selection QA, answerable accuracy requires exact claims and evidence; required-refusal accuracy uses forbidden requests; E2E divides their summed correct outcomes by all requests. Retrieval coverage uses answerable targets; measurement coverage uses accepted measurement answers over source-answerable measurement requests. Rejection or a retrieval-induced false refusal remains an error.

AUROC denotes area under the ROC curve; balanced accuracy (BA) averages sensitivity and specificity. MAE uses training-SD normalization where stated; Brier is mean squared probability error. Values are fractions unless marked as percentages; percentage differences are percentage points. Seed SD measures initialization variation. Participant tests aggregate repeated queries/seeds within person; S9 instead resamples question texts after record averaging. Each analysis specifies its own BH correction family.

Materials and access boundary. `semanticbn_review_draft.zip` contains this S1–S9 supplement, active manuscript sources, core S2–S5 code and dependencies, aggregate comparisons, frozen configurations, split counts and run instructions. `semanticbn_code.zip` provides the standalone code attachment. `scripts/review_smoke.py` tests synthetic FC, property definitions, both description grammars, equivalence-aware scoring and readout gradients without participant files.

The attachment retains all 28 static and 48 common-property paired comparisons, public-template embeddings, all 24 initial diagnostic fits and 48 mechanism trajectories, plus 96 follow-up validation histories and 74 additional paired comparisons. Participant identities, FC arrays, descriptions, scores and weights remain restricted. Cohort replay requires authorized derivatives and the original frozen

artifacts; synthetic checks verify implementation only. Secondary protocols are summarized in S6, S8 and S9, with local artifact locations for traceability; their complete execution artifacts are outside the attachment. Distribution remains subject to the author confirmations above.

S2 STATIC FULL-RECORD ALIGNMENT

The original evaluation object is the compressed record r in Equation 8. Each r_a mixes learned communities, and the head jointly reads all seven queries (Section 4.4). The supplied index a selects an output slot and a same-reference candidate bank. Scores measure whole-record attribute matching conditional on that index. Section S2.3 separately restricts pooled-vector access for numerical recovery; language-based reference identification remains outside these tasks.

This section supplies details beyond the main evaluation protocol. Source splits and properties are unchanged from the common-property benchmark (S5), while alignment heads and text targets are new. FC-derived gold and training thresholds are recomputed from the frozen source. Cutoffs are the $1/3$ and $2/3$ quantiles of all training participants within each cohort, reference and attribute, with diagnostic groups pooled. Values and cutoffs are rounded to six decimals: low is below the first cutoff, middle includes the first but excludes the second, and high includes the second. These empirical training-cohort bands are distinct from healthy-control norms; ties can make the three groups unequal in size.

Exact description grammars. The canonical template is: *Relative to training tertiles, internal FC is [level]; external FC contrast is [level]; across-partner FC dispersion is [level].* The levels are low, middle and high. The held-out template is: *Using training bands, within-set connectivity is [level]. The difference between two external connections is [level], and dispersion over the six external partners is [level].* Its levels are lower, intermediate and upper. Each template round-trips every three-band state, yielding 54 successful checks. The text contains only the discrete bands; the bank supplies the reference query. This protocol tests an agent-authored change of wording on a fixed set of attributes.

Implementation details. The main text specifies the common head and optimization. Its frozen inputs come from the six cohort/seed source fits described in Supplementary S5. ADNI FC-PCA has 90 active components and zero padding to 448. The archive key `fc_mlp` stores PCA inputs to the readout. Training sample SDs are floored at 10^{-6}. MiniLM is locally cached and mean-pooled over valid tokens; overflow is rejected. Seed-matched readouts share initialization and minibatch order; canonical-validation stopping selects epochs 1–48 across the 36 fits.

Control construction. The symbolic vectors are generated by QR orthogonalization with seed 20260926; the known grammar maps either template to the same band code. Shuffled text uses a training-person derangement and takes all seven reference targets from the same donor; validation and test labels stay correct. The mean-state control evaluates the individualized head at zero standardized input. The other-person control deranges test projections without changing their true-person targets. These last two controls require no extra fit.

Candidate multiplicity. The seven ADNI banks contain 14/13/14/11/14/16/13 distinct profiles; MDD contains 27/27/27/27/27/26/26. Under the main-text equivalence and tie rules, random-candidate expected accuracy is 9.3571% and 4.7110%, respectively. Candidate multiplicity determines the distinction between bank size and unique-profile count.

Table S1: Canonical static retrieval (%, mean ± seed SD). Paraphrase results are in main-text Table 3; the parser-based symbolic condition is unchanged by template.

	ADNI		MDD	
Condition	B→T	T→B	B→T	T→B
Individualized	15.95 ± 0.41	21.19 ± 2.70	14.87 ± 0.43	19.03 ± 7.96
Shared	15.24 ± 1.80	16.90 ± 2.18	14.82 ± 0.14	18.11 ± 2.23
Fixed reference	14.76 ± 1.80	18.81 ± 2.97	15.15 ± 0.24	20.66 ± 2.02
FC-PCA	13.81 ± 1.80	21.90 ± 0.41	10.68 ± 0.24	14.40 ± 2.43
Symbolic control	25.24 ± 0.41	28.81 ± 3.22	31.85 ± 1.12	50.94 ± 2.28
Shuffled training text	8.10 ± 1.09	8.10 ± 3.52	4.98 ± 0.70	4.23 ± 0.08
Training-mean state	8.33 ± 0.41	9.36 ± 0.00	4.70 ± 0.00	4.71 ± 0.00
Other person's state	4.76 ± 1.09	8.33 ± 2.30	4.06 ± 0.41	4.50 ± 0.88

Complete paired family. The following tables report every contrast in the main-text 28-test family, with differences defined as individualized minus alternative. Sign-flip and bootstrap resampling both use seed 20260926. Neither reference rows nor seeds count as independent participants. These conditional intervals do not resample model fitting or shared candidate banks.

Table S2: ADNI paired paraphrase contrasts: individualized minus alternative, percentage points. BH correction spans both cohorts and directions (28 tests).

Alternative	Direction	Difference	95% CI	Paired p	BH q
Shared	B→T	-2.86	$[-5.24, -0.71]$	0.04340	0.07148
Shared	T→B	-1.19	$[-4.76, 2.14]$	0.60264	0.64900
Fixed reference	B→T	-2.38	$[-4.05, -0.95]$	0.01690	0.03420
Fixed reference	T→B	0.48	$[-2.62, 3.57]$	0.88431	0.90291
FC-PCA	B→T	6.67	$[1.67, 12.14]$	0.02100	0.03920
FC-PCA	T→B	5.95	$[1.67, 10.24]$	0.01490	0.03420
Symbolic control	B→T	-12.38	$[-20.00, -4.76]$	0.00560	0.01568
Symbolic control	T→B	-13.81	$[-19.05, -8.33]$	0.00010	0.00047
Shuffled training text	B→T	2.62	$[-1.90, 8.10]$	0.41806	0.53207
Shuffled training text	T→B	7.38	$[2.86, 11.90]$	0.00820	0.02087
Training-mean state	B→T	5.48	$[-0.24, 11.90]$	0.10629	0.14881
Training-mean state	T→B	5.64	$[0.56, 10.95]$	0.05259	0.08181
Other person's state	B→T	6.43	$[1.90, 11.19]$	0.01710	0.03420
Other person's state	T→B	7.14	$[1.19, 12.86]$	0.03660	0.06404

Table S3: MDD paired paraphrase contrasts: individualized minus alternative, percentage points. BH correction spans both cohorts and directions (28 tests).

Alternative	Direction	Difference	95% CI	Paired p	BH q
Shared	B→T	0.22	$[-0.36, 0.80]$	0.48005	0.56006
Shared	T→B	-1.36	$[-2.16, -0.58]$	0.00070	0.00245
Fixed reference	B→T	0.25	$[-0.30, 0.81]$	0.35816	0.47755
Fixed reference	T→B	-1.05	$[-1.75, -0.36]$	0.00320	0.00995
FC-PCA	B→T	1.68	$[0.77, 2.62]$	0.00060	0.00240
FC-PCA	T→B	0.67	$[-0.02, 1.38]$	0.07919	0.11670
Symbolic control	B→T	-23.93	$[-25.85, -21.99]$	0.00010	0.00047
Symbolic control	T→B	-46.04	$[-47.87, -44.19]$	0.00010	0.00047
Shuffled training text	B→T	3.68	$[2.60, 4.76]$	0.00010	0.00047
Shuffled training text	T→B	0.06	$[-0.75, 0.89]$	0.90291	0.90291
Training-mean state	B→T	2.57	$[1.54, 3.62]$	0.00010	0.00047
Training-mean state	T→B	0.19	$[-0.44, 0.84]$	0.55564	0.62232
Other person's state	B→T	3.60	$[2.63, 4.62]$	0.00010	0.00047
Other person's state	T→B	0.30	$[-0.45, 1.02]$	0.45625	0.55544

Reproduction checks. `scripts/run_static_alignment.py` stores frozen source/code hashes, selected heads and participant scores. The audit reloads all heads and recomputes gold, normalization and scores; maximum projection and participant-score discrepancies are both zero. Tests cover the 54 grammar round trips, training-only normalization, orthogonal targets, derangements, equivalent positives, ties and gradients.

S2.1 Text and Exact-Attribute Diagnostics

The cohort experiment combines feature preservation, text-target geometry and finite-budget readout fitting. Post-review text and exact-code diagnostics isolate parts of this process using the 27 public three-band profiles. Variants and seeds were fixed before inspecting their scores. A later no-fit follow-up adds attribute-level errors and correct-target substitution in the existing cohort banks. These developmental probes leave the cohort heads, losses, earlier results and 28-test family unchanged.

Encoders and controlled wording changes. We freeze all-MiniLM-L6-v2, E5-small-v2 and BGE-small-en-v1.5 (Wang et al., 2020; 2022; Beijing Academy of Artificial Intelligence, 2023); all output 384-dimensional normalized vectors. MiniLM and E5 use attention-masked mean pooling; E5 prepends `query:` to both sides for symmetric similarity. BGE uses its CLS representation without a passage-retrieval instruction. Public model revisions, software versions and code hashes are fixed in the diagnostic protocol. No text is truncated: maximum token counts are 39/41/39, respectively.

The canonical description is unchanged. Five variants separate: (i) lexical substitution only, using lower/intermediate/upper in the canonical sentence; (ii) the earlier paraphrase sentence form with low/middle/high retained; (iii) the original combined paraphrase; (iv) clause order only, dispersion–internal–contrast with the canonical words; and (v) an additional narrative form: *The internal FC falls in the [level] training band. For the difference between two external connections, the band is [level]. The spread of FC over six external partners falls in the [level] band.* Variant (v) uses lower/intermediate/upper. All 162 descriptions round-trip exactly. The forms use agent-authored controlled grammars. Their full text is included in the code attachment.

Text-only matching. For each encoder, the exact canonical embedding replaces the learned brain projection. Each bank contains all 27 unique profiles, equally weighted. We rank the other template by cosine similarity and reverse the direction, retaining uniform credit for ties within 10^{-7}. Chance is $1/27 = 3.70\%$. This directly evaluates text-target correspondence with the brain encoder and learned readout removed. Table S4 reports all five variants. Canonical self-matching is 100% in both directions for all encoders, establishing distinguishability within the canonical template.

Table S4: Text-only profile matching (%). Each cell is canonical-to-variant / variant-to-canonical top-1 over 27 profiles. No participant data, fit or random seed is involved.

Variant	MiniLM	E5-small-v2	BGE-small-en-v1.5
Lexical only	3.70 / 18.52	18.52 / 29.63	14.81 / 33.33
Sentence form only	3.70 / 18.52	40.74 / 25.93	33.33 / 18.52
Original combined paraphrase	3.70 / 7.41	14.81 / 14.81	7.41 / 7.41
Clause order only	29.63 / 18.52	37.04 / 48.15	25.93 / 18.52
Additional narrative	7.41 / 7.41	7.41 / 7.41	7.41 / 3.70

Each profile has six counterfactual foils differing in exactly one band. For the original combined paraphrase, a foil scores at least as high as the correct description within the tie tolerance for 85.19/74.07/92.59% of canonical queries (MiniLM/E5/BGE). Canonical off-diagonal cosine means are .9769/.9780/.9495 and centered covariance effective ranks are 2.29/2.43/2.37. These diagnostics describe tightly grouped targets and wording-sensitive rankings. Per-profile credit, per-attribute accuracy and foil margins are included in the public-code results; the interventions below examine their learnability.

Attribute errors within text retrieval (analysis A). A no-fit follow-up reuses these frozen embeddings to distinguish adjacent-band mistakes from low–high reversals. Rows of each 3×3 confusion matrix are the true bands and columns are the bands of the top-ranked description, with uniform expected counts over ties. Each attribute has 27 queries, nine per true band. Table S5 reports marginal accuracy; joint accuracy requires all three attributes. These scores enumerate the complete synthetic profile set.

Table S5: Attribute accuracy in original-paraphrase text retrieval (%). C and P denote canonical and paraphrase. Joint accuracy requires all three bands to match.

Encoder	Direction	Internal	Contrast	Dispersion	Joint
MiniLM	C→P	37.04	33.33	33.33	3.70
	P→C	59.26	33.33	29.63	7.41
E5	C→P	51.85	48.15	55.56	14.81
	P→C	55.56	44.44	48.15	14.81
BGE	C→P	62.96	33.33	44.44	7.41
	P→C	37.04	37.04	37.04	7.41

MiniLM C→P selects a high-dispersion description for all 27 queries. Its internal/contrast/dispersion low–high reversal rates are 29.63/0.00/33.33%; adjacent-band error rates are 33.33/66.67/33.33%, with 27 as each denominator. Thus the errors include a preference for particular output bands, rather than a uniform reversal of direction. For a more local test, each query is paired separately with each of its two foils differing only in one specified attribute. Over 54 comparisons per attribute, MiniLM correctly prefers the gold description 70.37/61.11/61.11% of the time; pairwise ties receive half credit. Beating some local foils is easier than ranking first among all 27 profiles. Lexical-only and sentence-form-only MiniLM C→P both score 3.70% (Table S4), so both controlled wording changes can disrupt ranking. These probes characterize target-space errors and wording sensitivity.

Correct-target substitution in the original banks (analysis B). For every held-out record s and supplied reference a, set the query to $t_{\mathrm{can}}(c_{sa})$, where c_{sa} is its correct three-band code, and retrieve among $\{t_{\mathrm{para}}(c_{ua})\}_u$. Reverse retrieval queries paraphrase vectors against the canonical substitutes. We call the original same-reference evaluator, keeping all 20 ADNI or 304 MDD entries, their multiplicities, equivalent-profile positives and 10^{-7} tie rule. The local follow-up uses only cached attributes, training-derived cutoffs and split assignments; it fits no model and leaves earlier results unchanged. MiniLM uses the exact original float32 text cache. E5 and BGE use the public embeddings above with the same profile assignments and banks.

Table S6: Gold-dependent target substitution in original cohort banks (%). Canonical self-matching is 100% in both directions for every row. The queries use the correct canonical targets as oracle inputs.

Bank	Encoder	C→P	P→C	Chance
ADNI (20)	MiniLM	11.43	17.14	9.36
	E5	31.43	20.00	9.36
	BGE	11.43	16.43	9.36
MDD (304)	MiniLM	2.87	8.60	4.71
	E5	13.53	15.70	4.71
	BGE	8.08	8.18	4.71

Unlike the equal-weight 27-profile probe, these banks reflect the existing test distributions and missing profiles. Their chance rates are mean positive fractions, not $1/N$. A perfect canonical vector can have a poor cross-wording ranking. This is not a mathematical ceiling: learned queries need not equal canonical vectors and may rank paraphrases better, as the 7.91% MDD individualized score already illustrates. Accordingly, the difference between this probe and a fitted model is not an additive estimate of brain-side error.

Encoder repetition and reproducibility (analysis C). The three encoders repeat both no-fit probes under matched evidence and candidate rules. This encoder-level sensitivity analysis precedes the cohort retraining in Section S2.3. The separate exact-code fits below test readout optimization on synthetic codes. Full confusion matrices, all five variants and cohort aggregates are in `results/text_target_probe`; `scripts/run_text_target_probe.py--audit` reproduces the public-template metrics offline. Adding `--cohort-caches` also replays the local cohort substitutions from separately controlled property caches. Individual arrays and profile-frequency tables are excluded from the attachment. Author clearance and both cohorts' spatial-provenance status remain unchanged.

Exact-attribute input with a matched readout. The next probe supplies all correct bands directly. Seven numerical references each contribute nine one-hot scalars (three properties, three levels); their 63 scalars are zero-padded to 448. The head is the original $448 \to 128 \to 7 \times 384$ network with 404,224 trainable parameters. Optimizer, temperature, batch size, 60-epoch cap, eight-epoch patience and canonical-validation stopping match the cohort readout. Each seed uses identical inputs, initialization and minibatch order across the three encoders and the orthogonal symbolic target. Exact-code inputs isolate readout learning from brain-feature preservation.

The seen-profile regime uses 324 training and 108 validation records, balanced over all 27 profiles independently at each reference; each test reference enumerates all 27 profiles once. These are synthetic code records. A separate compositional regime excludes from both training and validation the nine profiles satisfying $(c_1 + c_2 + c_3) \bmod 3 = 0$, with $c_j \in \{0, 1, 2\}$, and tests only those profiles. Each attribute level remains present during training. The two regimes have different banks and chance rates (3.70% versus 11.11%), so their absolute percentages are not directly comparable. Validation uses only seen codes and canonical wording. All 24 fits (four targets, two regimes, seeds 7/11/19) completed; selected epochs are 47–60. There was no post-test tuning or best-seed selection.

Table S7: Exact-attribute readout top-1 (%, mean ± seed SD). O denotes the learned projection of true synthetic attribute codes. The symbolic condition uses an exact parser and a fixed orthogonal codebook. All additional variants and per-seed histories are in the attachment.

		Canonical		Original paraphrase	
Regime	Target	O→T	T→O	O→T	T→O
Seen	MiniLM	$22.40 \pm .31$	36.68 ± 2.50	$4.94 \pm .61$	6.00 ± 1.22
	E5	30.16 ± 2.95	27.69 ± 2.44	21.16 ± 1.40	$10.93 \pm .31$
	BGE	28.75 ± 2.20	42.50 ± 3.83	$8.11 \pm .31$	$8.82 \pm .61$
	Symbolic	66.49 ± 3.01	69.84 ± 3.70	66.49 ± 3.01	69.84 ± 3.70
Unseen	MiniLM	$33.33 \pm .00$	45.50 ± 1.83	$11.11 \pm .00$	17.99 ± 1.83
	E5	35.98 ± 1.83	45.50 ± 3.30	34.92 ± 1.59	26.46 ± 2.42
	BGE	$33.86 \pm .92$	56.61 ± 6.42	23.28 ± 1.83	22.22 ± 4.20
	Symbolic	14.29 ± 4.20	10.58 ± 2.42	14.29 ± 4.20	10.58 ± 2.42

Interpretation of the diagnostic controls. Text-only matching identifies wording sensitivity at the target level. The exact-attribute probe additionally identifies finite-budget readout error with the required information explicitly available. Symbolic targets improve seen-profile learning, while E5 performs better on the unseen-profile test. Orthogonal profile IDs provide separation but leave relationships between training and held-out states unspecified. These outcomes motivate the geometry, parsing and optimization interventions below. Their scope is three compact English encoders and five controlled variants; assigning portions of cohort error to these factors requires interventions on the cohort models.

Public-input reproduction. `scripts/run_alignment_diagnostics.py` performs all 24 fits after freezing its protocol. The code package includes templates, public-template embeddings, all run histories and summary statistics under `results/alignment_diagnostics`. `scripts/audit_alignment_diagnostics.py` checks hashes, replays all 18 text tables and recomputes aggregates from every saved run, offline without encoder weights. This audit does not replay checkpoints. Rerunning the fits requires the specified public encoders, not clinical data.

Unlike the restricted cohort reproduction above, this diagnostic can be independently rerun from the attached code and public model weights.

S2.2 Separating the Symbolic Control's Advantages

The symbolic control combines orthogonal geometry with an exact parser. To distinguish their effects, we freeze a second public-code protocol using the released template cache, the same seen-profile exact-code inputs, three encoders and seeds 7/11/19. There are 48 training trajectories: three raw encoder targets plus one shared orthogonal target, four training objectives and three seeds. Each trajectory supplies two validation-selected checkpoints. All predefined settings and earlier results are retained; checkpoint selection uses canonical validation.

Geometry crossed with supplied parsing. Let $C \in \mathbb{R}^{27\times 384}$ contain the normalized canonical text prototypes and S the original orthogonal profile vectors. Raw/no-parser evaluates actual alternative embeddings. Raw/parser resolves either wording to its canonical embedding, changing no training target or fitted head. Orthogonal/no-parser applies $W = C^{+}S$ to each actual text embedding and normalizes it. Only canonical profiles fit W; alternative wording is excluded. Orthogonal/parser resolves both wordings to S. All 27 canonical profile identities occur in training, so this is a canonical-profile-supervised intervention, not an unsupervised transform or unseen-code test. Both parser cells apply the supplied grammar. The parser maps each candidate's text, not the query's gold answer.

The pseudoinverse cutoff is fixed at 10^{-7} of the largest singular value. All canonical matrices retain row rank 27; their condition numbers are 5675/2696/1697 (MiniLM/E5/BGE). Canonical mapping error is below 10^{-6}, but map operator norms are 1104/524/335. Thus orthogonalizing canonical targets can amplify out-of-template perturbations. Canonical S is identical across encoders, so one orthogonal head is shared rather than counted as three independent fits.

Table S8: Geometry–parser factorial on the original paraphrase (%, mean ± seed SD), at $\tau = .07$ and 60 epochs. Inputs are true attribute codes. O denotes their learned projection. The orthogonal/parser result is shared, not three replications.

Encoder	Geometry	Supplied parser	O→T	T→O
MiniLM	Raw	No	$4.94 \pm .61$	6.00 ± 1.22
	Raw	Yes	$22.40 \pm .31$	36.68 ± 2.50
	Orthogonal	No	$4.76 \pm .53$	$3.53 \pm .61$
	Orthogonal	Yes	66.49 ± 3.01	69.84 ± 3.70
E5	Raw	No	21.16 ± 1.40	$10.93 \pm .31$
	Raw	Yes	30.16 ± 2.95	27.69 ± 2.44
	Orthogonal	No	3.88 ± 1.70	$4.94 \pm .61$
	Orthogonal	Yes	66.49 ± 3.01	69.84 ± 3.70
BGE	Raw	No	$8.29 \pm .61$	$9.17 \pm .61$
	Raw	Yes	28.92 ± 2.50	41.45 ± 3.44
	Orthogonal	No	$4.41 \pm .61$	4.76 ± 1.06
	Orthogonal	Yes	66.49 ± 3.01	69.84 ± 3.70

The supplied parser makes both wordings use identical targets while preserving canonical learning error. Orthogonal targets improve canonical learning under the original objective; actual paraphrase accuracy remains close to chance. The factorial therefore measures the interaction between target separation and exact parsing within this controlled readout.

Loss, temperature and optimization budget. We prescribe contrastive temperatures .07, .02 and .005, plus pointwise cosine regression $\mathrm{mean}[1 - \cos(z, t)]$. All other training settings remain fixed: 404,224-parameter head, AdamW learning rate .001, weight decay .0001, batch size 64, and seed-matched inputs, normalization, initialization and minibatch order. Every trajectory runs 240 epochs without patience stopping; the 60-epoch checkpoint uses only its first 60 epochs. Within each prefix, selection minimizes the original canonical-validation contrastive loss at $\tau = .07$, irrespective of the

training objective. This common criterion does not use paraphrases, but need not be optimal for every loss. Unlike the preceding 24-fit probe, patience stopping is disabled; the small BGE baseline difference is retained rather than overwriting the earlier result.

Table S9: Canonical exact-code O→T accuracy (%, mean ± seed SD). Raw text geometry is unchanged across loss/budget conditions. The orthogonal column uses the shared target; canonical wording makes parsing irrelevant. All prescribed settings are reported.

Training objective	Epoch cap	MiniLM	E5	BGE	Orthogonal
Contrastive .07	60	$22.40 \pm .31$	30.16 ± 2.95	28.92 ± 2.50	66.49 ± 3.01
	240	$24.87 \pm .53$	40.56 ± 4.04	40.04 ± 3.60	65.26 ± 3.19
Contrastive .02	60	45.86 ± 1.53	66.84 ± 5.56	71.08 ± 4.44	66.67 ± 5.05
	240	52.03 ± 2.50	76.01 ± 1.10	75.13 ± 3.22	69.14 ± 3.44
Contrastive .005	60	$78.84 \pm .53$	$87.65 \pm .61$	$95.06 \pm .31$	55.38 ± 5.40
	240	95.24 ± 1.40	98.06 ± 1.10	$99.12 \pm .31$	58.02 ± 6.74
Cosine regression	60	78.84 ± 2.65	58.91 ± 3.19	76.37 ± 1.33	57.85 ± 3.01
	240	92.95 ± 1.86	$71.96 \pm .53$	84.48 ± 1.33	56.44 ± 2.14

Table S10: Actual original-paraphrase O→T accuracy for the same raw-text trajectories as Table S9 (%, mean ± seed SD). No supplied parser is used. Strong canonical recovery does not remove the wording mismatch.

Training objective	Epoch cap	MiniLM	E5	BGE
Contrastive .07	60	$4.94 \pm .61$	21.16 ± 1.40	$8.29 \pm .61$
	240	$4.06 \pm .31$	21.16 ± 1.91	$8.11 \pm .81$
Contrastive .02	60	$4.06 \pm .31$	18.52 ± 2.31	$7.76 \pm .31$
	240	$4.41 \pm .81$	24.16 ± 2.50	$6.70 \pm .81$
Contrastive .005	60	$5.29 \pm .92$	18.87 ± 2.44	$5.82 \pm .92$
	240	$3.70 \pm .92$	20.81 ± 2.91	$7.05 \pm .31$
Cosine regression	60	$4.59 \pm .31$	$16.58 \pm .81$	$7.41 \pm .00$
	240	$3.70 \pm .00$	$18.52 \pm .00$	$7.41 \pm .00$

MiniLM's canonical accuracy rises from 22.40% to 78.84% by lowering temperature at 60 epochs; extending the original .07 trajectory reaches 24.87%. Combining .005 with 240 epochs yields 95.24%, retaining the same text encoder, target geometry and attribute information. Cosine regression reaches 92.95% at 240 epochs. Canonical accuracy is thus sensitive to optimization even with fixed text targets. For the orthogonal target, lowering temperature from .07 to .005 at 60 epochs changes accuracy from 66.49% to 55.38%, demonstrating a geometry-dependent response. The accompanying paraphrase scores distinguish canonical learning from wording transfer.

Is template shift just a common offset? For each alternative wording, estimate $\delta = \mathrm{mean}_{c \in \mathcal{C}}[t_{\mathrm{alt}}(c) - t_{\mathrm{can}}(c)]$ on nine calibration profiles with $\sum_j c_j \bmod 3 = 0$. Subtract δ before normalization and evaluate only the remaining 18 profiles, using identical candidate banks before and after correction. No scored alternative description fits the offset. The nine calibration descriptions are an additional resource, not an uncalibrated robustness claim. Chance is $1/18 = 5.56\%$; these scores must not be compared directly with the preceding 27-candidate text-only table.

Table S11: Canonical-to-variant text matching on 18 non-calibration profiles (%). Cells show uncorrected / offset-corrected scores. No learned readout, clinical observations or seed averaging is involved.

Variant	MiniLM	E5	BGE
Lexical only	5.56 / 16.67	22.22 / 33.33	22.22 / 38.89
Sentence form only	5.56 / 55.56	38.89 / 50.00	44.44 / 38.89
Original combined paraphrase	16.67 / 5.56	16.67 / 27.78	5.56 / 11.11
Clause order only	33.33 / 22.22	27.78 / 22.22	22.22 / 5.56
Additional narrative	11.11 / 11.11	11.11 / 22.22	11.11 / 22.22

Offset correction raises MiniLM's sentence-form matching from 5.56% to 55.56% and lowers combined-paraphrase matching from 16.67% to 5.56%. The response varies across encoders and templates, identifying a template-dependent role for mean displacement. The attachment includes reverse-direction scores, displacement residuals, all variants and seed histories. Its auditor verifies all 48 trajectories, both validation-selected budgets, source hashes, matched controls, calibration and 1,728 aggregate metric cells from saved outputs. Checkpoint replay is a separate operation; training can be repeated from the attached public embeddings.

S2.3 Frozen-Cohort Follow-up: Input Access and Text Encoders

This separately frozen developmental follow-up reuses the original cohort splits, three community seeds, gold properties, training cutoffs and public text caches. It completes 24 numerical and 72 language fits. All 96 fits use training/validation alone and are locked before a separate test-evaluation command. Original fits and scores remain archived unchanged. No diagnosis enters these objectives.

Restricting pooled-vector access. For each cohort/seed, maximum-weight one-to-one matching uses only training-mean memberships $\bar{Q}_{\mathrm{train}}$, with score

$$D_{ak} = \frac{2\sum_i Y_{ia}\bar{Q}_{\mathrm{train},ik}}{\sum_i Y_{ia} + \sum_i \bar{Q}_{\mathrm{train},ik}}.$$

This fixes a community index $\pi(a)$ for every reference across all splits. Four $448 \to 128 \to 21$ heads have 60,181 parameters and seed-matched initialization. Full heads read seven reference mixtures r_a or seven matched unit vectors $b_{\pi(a)}$. Local heads use only the corresponding 64-dimensional block for each output, setting all other block contributions to zero inside the first layer. Shared biases remain; outputs use their own final-layer rows. Other-block perturbation and gradient tests verify isolation. AdamW, normalization and stopping match S5's numerical protocol (200-epoch cap, patience 20). The head accesses one pooled vector, but that vector still derives from whole-FC-row node features; the gold describes the common reference, not the matched community's own FC.

Table S12: Restricted-access numerical readout (%, three-seed mean ± SD). Two/three denotes joint correctness of two/three attribute bands. Full and local have equal trainable parameter counts but different input access.

Input / access	ADNI two	ADNI three	MDD two	MDD three
Reference mixture / full	41.19 ± 2.70	25.48 ± 0.82	69.03 ± 1.52	53.10 ± 0.60
Reference mixture / local	52.14 ± 2.58	33.33 ± 3.22	70.08 ± 0.65	54.56 ± 0.33
Matched community / full	39.29 ± 6.55	25.24 ± 2.06	67.50 ± 1.36	49.08 ± 1.04
Matched community / local	37.62 ± 2.97	21.67 ± 4.54	57.61 ± 0.08	40.44 ± 0.78

The primary numeric family contains three contrasts per cohort: reference local–full, community local–full, and community full–reference full. Seeds and references are averaged within person before 10,000 sign flips and 5,000 bootstraps; BH covers all six contrasts. Reference-local minus full is +10.95 points on ADNI (95% CI [4.76,17.14], $q = .0111$) and +1.05 on MDD ([-.27,2.35], $q = .1873$). Community-local minus full is -1.67 on ADNI ([-7.38,3.58], $q = .5764$) and -9.88 on MDD ([-11.39,-8.33], $q = .00060$). Thus dimensionality alone does not explain recovery: reference mixtures and matched communities encode different task information. This establishes a task-specific single-vector limit, not general failure of individual communities.

Actual cohort alignment with three encoders. Four representations (individualized, shared, fixed reference and FC-PCA), three encoders, two cohorts and three seeds give 72 new fits. Each uses the original 404,224-parameter head, temperature .07, AdamW .001, weight decay .0001, batch 64, 60-epoch cap and patience eight. Canonical validation alone selects stopping. Encoder revisions/pooling and all six template embeddings are frozen from the public diagnostic cache; MiniLM canonical/paraphrase targets reuse the original cohort bytes. Canonical and all five alternative forms are scored in the original 20/304-person banks. The original combined paraphrase is the primary wording. Identical temperature controls the protocol, not encoder-specific optimization difficulty.

Table S13: Three-encoder cohort paraphrase retrieval (%, mean ± seed SD). All four representations and all encoders are retained; no test-selected encoder replaces the original main comparison.

Encoder	Representation	ADNI B→T	ADNI T→B	MDD B→T	MDD T→B
MiniLM	Individualized	12.86 ± 3.11	15.00 ± 2.86	$7.91 \pm .33$	4.90 ± 2.07
	Shared	15.71 ± 2.58	16.19 ± 2.18	$7.69 \pm .47$	6.27 ± 1.04
	Fixed reference	15.24 ± 2.89	14.52 ± 2.97	$7.66 \pm .14$	5.95 ± 2.14
	FC-PCA	6.19 ± 1.09	9.05 ± 4.19	$6.23 \pm .41$	$4.23 \pm .90$
E5	Individualized	17.14 ± 1.43	16.43 ± 6.19	$12.73 \pm .53$	$9.01 \pm .90$
	Shared	15.00 ± 1.24	14.52 ± 3.22	$12.30 \pm .22$	10.03 ± 1.19
	Fixed reference	$16.19 \pm .41$	15.95 ± 2.06	$11.98 \pm .20$	9.13 ± 1.86
	FC-PCA	11.90 ± 2.30	$13.33 \pm .82$	$9.51 \pm .57$	8.88 ± 1.02
BGE	Individualized	16.19 ± 2.18	18.33 ± 1.80	$6.89 \pm .21$	5.26 ± 1.65
	Shared	15.00 ± 1.89	13.10 ± 1.80	$7.32 \pm .10$	7.53 ± 2.79
	Fixed reference	$16.19 \pm .41$	16.67 ± 2.97	$7.06 \pm .17$	6.19 ± 2.59
	FC-PCA	16.43 ± 1.43	15.00 ± 3.98	$7.79 \pm .28$	8.08 ± 1.75

The text family has 68 seed-averaged participant contrasts: E5/BGE minus MiniLM within each representation, and individualized minus each other representation within each encoder, in both directions/cohorts. The same resampling scheme applies, with BH across all 68. MDD individualized E5 minus MiniLM is +4.82 points B→T (95% CI [3.45,6.17]) and +4.10 T→B ([3.16,5.06]), both $q = .00057$. E5 individualized minus shared is +.43 B→T ($q = .4094$) and -1.02 T→B ($q = .0348$); versus fixed reference, B→T is +.75 ($q = .0309$). The encoder effect is clearer than a consistent individualized-partition advantage. ADNI E5/BGE gains over MiniLM are nonsignificant after correction. All 456 aggregate cells, 74 paired comparisons and 96 validation histories accompany `results/review_followup`; `scripts/audit_review_followup.py` replays public aggregates and selection. The optional `--checkpoints` audit additionally replays all 96 local models using restricted participant artifacts. Both cohorts' provenance and authorization requirements remain unchanged.

S3 Community Organization and Ablations

Fitting and aggregation scope. The historical pooled fits use the main-text community architecture and loss, with dataset-balanced sampling and seeds 7/17/27. They differ from the cohort-specific training-only source models of the primary static study. The pooled organization models fit all 442 ADNI scans and 2,018 MDD scans and are evaluated on those same cohorts. These are structural diagnostics on the fitting cohorts. For a cohort with n_C scans, SemanticBN constructs

$$\bar{Q}_{ik} = n_C^{-1} \sum_{r=1}^{n_C} Q_{ik}^{(r)}, \qquad \hat{c}_i = \arg\max_k \bar{Q}_{ik}, \qquad A_C = \big|\{\hat{c}_i : i = 1, \ldots, R\}\big|.$$

ARI and matched overlap compare $\hat{c}$ with valid Yeo-labeled ROIs; Grad. averages coverage-weighted gradient SD over valid hard communities. Active reports the cohort-template count A_C. ADNI repeated scans contribute separately to the scan-weighted $\bar{Q}$. Simple clustering baselines use cohort-mean FC, while foundation-style baselines cluster cohort-mean ROI embeddings. Error bars summarize variation across initialization seeds. Anatomical metrics for both ADNI and MDD remain conditional on provisional ROI ordering.

Table S14: Complete pooled common-90-column template comparison (mean ± sample standard deviation over three seeds), fitted and evaluated within the same cohorts. Ov. is optimal one-to-one Yeo-7 overlap and Grad. is hard-template gradient dispersion; lower Grad. is better. Deterministic group baselines have zero seed variation where applicable.

	ADNI			REST-meta-MDD		
Method	ARI	Ov.	Grad.	ARI	Ov.	Grad.
Random	.011 ± .006	.282 ± .012	.234 ± .004	.011 ± .006	.282 ± .012	.234 ± .004
Agglomerative	.119	.400	.121	.339	.576	.109
Spectral	.222	.447	.153	.388	.588	.145
K-means FC	.264 ± .038	.471 ± .031	.136 ± .018	.442 ± .037	.612 ± .024	.114 ± .005
NMF	.057	.329	.177	.426	.647	.132
BrainLM-style	.270 ± .021	.471 ± .020	.126 ± .005	.403 ± .039	.584 ± .049	.104 ± .010
BrainGFM-style	.130 ± .023	.361 ± .024	.198 ± .007	.244 ± .024	.471 ± .024	.149 ± .003
SEMANTICBN	.293 ± .057	.510 ± .053	.147 ± .027	.426 ± .077	.627 ± .049	.122 ± .007

The BrainLM-style baseline masks FC entries and predicts their values from an ROI encoder. The BrainGFM-style baseline combines masked graph reconstruction with a contrastive objective over two graph views. Both use the same processed FC inputs and training seeds as SEMANTICBN. They are deliberately labeled "style" baselines because they do not reproduce the original models' pretraining scale, data, complete architectures, or released weights (Ortega Caro et al., 2024; Wei et al., 2026a).

S3.1 OBJECTIVE ABLATIONS

Table S15: Within-cohort template ablations (mean ± sample standard deviation over three seeds). Active averages the cohort-template count A_C across seeds; it is identical for both cohorts in these runs. Seven labels are intended.

	ADNI		MDD		
Objective	ARI	Overlap	ARI	Overlap	Active
Full	.293 ± .057	.510 ± .053	.426 ± .077	.627 ± .049	7.00 ± .00
w/o reconstruction	.355 ± .089	.537 ± .060	.441 ± .064	.631 ± .056	7.00 ± .00
w/o balance	.090 ± .015	.384 ± .024	.061 ± .037	.337 ± .030	3.33 ± .58
w/o entropy	.272 ± .062	.486 ± .060	.487 ± .077	.694 ± .077	7.00 ± .00
w/o modularity	.280 ± .066	.502 ± .053	.430 ± .077	.627 ± .048	7.00 ± .00

The active count describes the hard cohort template. Removing reconstruction improves template ARI in both cohorts; the original objective is retained rather than selected by these descriptive scores.

S4 ATLAS SENSITIVITY AND PROVENANCE

Table S16: Provisional AAL90/AAL116 sensitivity on the same 1,976 MDD participants, fitted and evaluated within cohort. The last column is the fraction of group-average soft communities with unadjusted permutation $p \leq .05$, not a per-person FDR-qualified authorization rate.

Atlas	Subjects	SEMANTICBN ARI	K-means ARI	Best overlap	Unadjusted $p \leq .05$
AAL90	1976	.436 ± .054	.440 ± .038	.651 ± .049	.952 ± .082
AAL116	1976	.487 ± .073	.445 ± .040	.670 ± .070	.762 ± .082

Each atlas-sensitivity seed evaluates gradient compactness on cohort-average soft $\bar{Q}$, using 200 ROI-label permutations and unadjusted lower-tail p values. The denominator is the number of communities with finite tests; the table reports the mean and sample SD over three seeds. The archived

claim route instead used individual $Q^{(s)}$, 199 permutations, within-scan BH correction and the then-recorded provenance flags. Its historical count of 233 gradient-permitted units is a different statistic and is superseded as current authorization by the assessment below. Neither null preserves spatial autocorrelation.

Table S17: Current provenance assessment. The package's submission-status record supersedes historical verified flags; the older audit contains MDD, ABIDE and epilepsy inventory checks but omitted ADNI.

Dataset	Status	Evidence and unresolved requirement
ADNI	Unverified	Local declaration of the first 90 AAL ROIs; upstream extraction records, atlas checksum and column-bound ROI labels lack a verified binding. Historical anatomical permissions are superseded.
REST-meta-MDD	Unverified	The 2,034 series have 116 columns, 1,604 IDs match the phenotype file, and published reuse describes AAL116. Local extraction records, atlas checksum and column-bound ROI labels lack a verified binding.
ABIDE	Unknown atlas/order	Two 90-column variants show mean same-index absolute correlation .907 and optimal identity match .968, but no metadata identifies the shared order as AAL90. Excluded.
Epilepsy	Unknown atlas/order	All 49 files have 90 columns, but ROI names, atlas image, extraction script, and a control group are absent. Excluded.

Equal ROI count or statistical column similarity does not identify an atlas. **Numerical and anatomical validity are separate.** Within-cohort numerical comparisons assume stable column identity across participants and splits, a consistent preprocessing/indexing convention, and identical fixed column weights across methods. Under these stated assumptions, attribute values and conditional representation comparisons remain interpretable as numerical results. The reference sets are not confirmed functional networks, and their spatial or gradient agreement cannot establish independent neuroscience validation. Anatomically interpretable pooling or transfer additionally requires verified cross-cohort ROI correspondence. Both current input manifests are provisional; frozen caches, scores and hashes retain their historical state for reproducibility.

Dataset source versus analysis cohort. REST-meta-MDD is a DIRECT consortium resource (Yan et al., 2019; Chen et al., 2022); the supplied 2,034-person derivative is distinct from the source project's complete cohort. The historical 90-column filter retains 2,018 people, whereas the current all-116-column filter retains 1,976. ADNI historical analyses use 442 scans from 130 people, so scan and participant counts are not interchangeable. ABIDE's reported distribution route is the Preprocessed Connectomes Project, but the two local 90-column variants lack the extraction metadata needed to establish atlas order. The 49 epilepsy files are not treated as a public dataset and contribute no reported experiment. A source-paper description does not by itself establish the local files' identity, transfer permission or ethical approval. The provenance statuses above concern scientific traceability, not legal or institutional clearance.

S5 Common-Source Property Benchmark

Gold and source representations. This earlier benchmark uses Equation 7, the participant splits and training-only rounding/band rules of the primary static study. Its separately fitted community/PCA models supply that study's frozen features. Its own readout predicts numerical attributes rather than MiniLM vectors. Both studies fit band cutoffs on the full training cohort, pooling diagnostic groups. For both ADNI and MDD, fixed reference weights define numerical targets conditional on the column-order assumptions in S4.

Matched architectures and strong controls. Source community fitting follows the main-text training schedule and seeds, using training FC only. Within a seed, individualized, shared, fixed-reference, per-scan K-means and per-scan spectral representations freeze this same ROI encoder.

Shared Q is the mean training assignment; it still pools each person's own ROI features. K-means uses FC rows; spectral clustering uses positive FC with zero diagonal, seven clusters, ten initializations and the same seed. Fixed-reference assignments row-normalize Y; unmapped ROIs receive uniform membership.

The reference query is Equation 8. Each neural property head is Linear(448,128)/GELU/Linear(128,21), with 60,181 trainable parameters, training-only feature/target standardization, AdamW .001, weight decay .0001, batch size 64, at most 200 epochs and validation-MSE patience 20. The same seed-specific minibatch order is used across these heads; test scores never select stopping.

FC-PCA/MLP uses the same head after training-only PCA of the 4,005 signed FC edges. It retains at most 448 components, limited to training sample count minus one (90 for ADNI), with zero padding to the common input width. Raw-FC ridge instead uses all 4,005 standardized edges and validation-selects its penalty from $.1, 1, 10, 100, 1000$. This full-input linear control uses every FC edge to recover held-out participants' attributes. Direct source aggregation is an explicit exact-information tool bound. These controls are especially informative because the first two targets are linear FC summaries.

Structured property retrieval. The neural heads predict numerical attributes. A candidate description is represented by its three explicit numerical attributes; normalized squared distance compares them with predicted attributes. Candidate banks contain 16 held-out people from the same cohort and reference set, with the query person included at a shuffled position. Every candidate sharing all three gold tertile bands is positive, irrespective of identity. The reverse direction ranks predicted brain attributes for the same bank. Candidates and their order are fixed across methods/seeds; 32/128-candidate extensions apply only when sufficient people exist. This protocol evaluates structured property matching. The mean positive fraction for 16-way banks is 10.58% on ADNI and 10.44% on MDD, rather than automatically $1/16$.

Table S18: Common-source controls: joint two-property accuracy and structured 16-candidate retrieval (%, mean ± SD across three seeds). B/T denote predicted brain attributes/explicit description attributes. Seed averages are not independent participants. Negative controls retain the true-person gold while replacing predictions.

Cohort	Representation	Joint properties	B→T	T→B
ADNI	Individualized	41.19 ± 2.70	45.24 ± 2.06	53.33 ± 4.12
	Shared	39.52 ± 1.80	41.43 ± 3.27	49.52 ± 9.85
	Fixed reference	42.14 ± 0.71	43.33 ± 1.80	51.43 ± 2.47
	Individual K-means	45.48 ± 2.18	44.05 ± 2.51	55.95 ± 6.40
	Individual spectral	43.33 ± 1.80	43.57 ± 3.98	54.76 ± 6.64
	FC-PCA/MLP	48.81 ± 5.77	33.57 ± 3.27	56.43 ± 3.57
	Raw-FC ridge	62.14 ± 0.00	56.43 ± 0.00	65.71 ± 0.00
	Training mean	10.00 ± 0.00	9.29 ± 0.00	12.14 ± 0.00
	Cross-person shuffle	10.48 ± 0.41	8.33 ± 3.93	6.19 ± 2.30
MDD	Individualized	69.03 ± 1.52	81.55 ± 0.88	84.63 ± 0.91
	Shared	65.95 ± 1.80	77.40 ± 0.25	79.65 ± 1.14
	Fixed reference	71.57 ± 1.51	83.74 ± 0.85	86.75 ± 0.97
	Individual K-means	67.15 ± 0.48	78.73 ± 0.35	81.39 ± 0.25
	Individual spectral	62.73 ± 0.78	72.24 ± 0.40	76.11 ± 1.27
	FC-PCA/MLP	69.17 ± 0.70	68.75 ± 0.94	75.38 ± 0.83
	Raw-FC ridge	92.58 ± 0.00	81.16 ± 0.00	86.61 ± 0.00
	Training mean	12.22 ± 0.00	8.88 ± 0.00	10.43 ± 0.00
	Cross-person shuffle	10.93 ± 0.38	8.71 ± 0.75	9.84 ± 0.68

Paired representation statistics. Seven queries and three seeds are averaged within person before 10,000 paired sign flips and 5,000 participant bootstrap draws. The BH family contains all 48 comparisons: eight non-source alternatives versus individualized, two cohorts, and joint-property/B→T/T→B outcomes. MDD's individualized-minus-shared differences are 3.09 points for joint properties (95% CI [1.90,4.31]), 4.15 for B→T ([3.10,5.20]) and 4.98 for T→B ([3.92,6.12]);

all have $q = .00019$. ADNI's corresponding joint-property difference is 1.67 points ([-1.67,5.24], $q = .449$); its T→B difference is 3.81 ([.95,6.67], $q = .0445$). Fixed-reference joint accuracy exceeds individualized on MDD by 2.54 points ([1.35,3.68], $q = .00019$). Raw-FC ridge exceeds it by 20.95 points on ADNI ([10.95,30.00], $q = .00178$) and 23.54 on MDD ([21.96,25.11], $q = .00019$). These findings support limited information preservation, not an overall performance advantage for learned individual communities. Deterministic ridge results repeated under each seed are not independent training replications.

Including the nonlinear external-dispersion band gives MDD three-property accuracy of 53.10% for individualized, 48.61% shared, 55.20% fixed-reference and 59.26% ridge (descriptive seed means). Their mean absolute errors normalized by training SD are .251, .277, .235 and .183. In 128-candidate MDD banks, B→T/T→B scores are 61.23/64.83%, 56.80/60.50%, 64.99/68.26% and 67.48/72.93%, respectively; the mean positive fraction is 5.18%. ADNI's 20-person test set supports the 16-candidate evaluation. These additional descriptive scores are outside the specified 48-test family; all method/seed outputs are retained in the audit.

Auxiliary common-property QA. The archived QA crosses individualized/shared representation with evidence control off/on and applies control to six additional representations. All 324 held-out people receive seven joint-property queries, one missing-contrast probe and one unsupported-diagnosis probe: 2,916 requests per condition and 29,160 responses in total. Readout seed 7 is fixed. Each regular answer selects one of nine two-band conjunctions and cites four entries: both estimates and both training-tertile records. IDs are pseudorandom and shuffled.

Evidence control permits a measurement answer only when both estimate-plus/minus-median-validation-residual intervals stay within single bands. This is an empirical policy, not calibrated uncertainty. The verifier checks permissions, displayed estimates and exact citations; source-FC gold is used only for scoring. Frozen Qwen2.5-7B-Instruct uses FP16, greedy decoding, repetition penalty 1.1, 512 output tokens and a 4,096-token input ceiling (actual inputs 706–820). The completed ASCII-identifier grammar permits wrong identifiers and empty-array refusals. All outputs finish below the cap; 99.93% satisfy the schema.

Measurement denominators are 140 ADNI and 2,128 MDD queries; E2E adds 40 and 608 refusal probes. The full on/off factorial below reports raw and post-verifier joint correctness against source gold, with accepted measurement coverage.

Cohort	Assignment	Control	Raw joint	System joint	Coverage
ADNI	Shared	Off	0.00	0.00	100.00
	Individualized	Off	0.71	0.71	100.00
	Shared	On	1.43	0.00	0.00
	Individualized	On	1.43	0.00	0.00
MDD	Shared	Off	0.52	0.52	99.86
	Individualized	Off	0.70	0.70	99.81
	Shared	On	1.60	0.47	0.56
	Individualized	On	1.50	0.42	0.70

Table S19: All controlled common-property QA methods (%). Raw claim ignores citations; raw/system joint require the source-gold conjunction and exact citations. Coverage is accepted measurement-answer coverage. System E2E also credits required refusals; an always-refuse system earns 22.22%. The direct-source row still uses Qwen and must not be confused with the deterministic tool bound.

Cohort	Representation	Raw claim	Raw joint	Sys. joint	Coverage	Sys. E2E
ADNI	Individualized	12.14	1.43	0.00	0.00	22.22
	Shared	10.71	1.43	0.00	0.00	22.22
	Fixed reference	12.14	2.14	0.00	0.00	22.22
	Individual K-means	12.14	1.43	0.00	0.00	22.22
	Individual spectral	13.57	0.00	0.00	0.00	22.22
	FC-PCA/MLP	12.14	0.71	0.00	0.71	22.22
	Raw-FC ridge	12.14	1.43	0.00	0.00	22.22
	Direct-source text	11.43	4.29	4.29	4.29	25.56
MDD	Individualized	10.67	1.50	0.42	0.70	22.55
	Shared	11.04	1.60	0.47	0.56	22.59
	Fixed reference	10.95	1.50	0.66	0.80	22.73
	Individual K-means	10.95	1.36	0.47	0.70	22.59
	Individual spectral	11.04	1.27	0.19	0.33	22.37
	FC-PCA/MLP	10.39	1.60	0.47	0.70	22.59
	Raw-FC ridge	11.47	1.55	1.46	1.55	23.36
	Direct-source text	10.90	1.83	1.83	1.83	23.65

The controlled individualized MDD system accepts 15 measurement answers, nine jointly correct; shared accepts 12, ten correct. Neither accepts an ADNI measurement answer. All controlled systems reject unsupported probes. MDD individualized E2E is 22.55%, close to the 22.22% always-refuse control. Of 2,067 schema-valid answered individualized-on measurement outputs, 1,776 cite two IDs instead of the required four; only 227/2,128 requests select a claim matching the displayed estimates. The gate permits 1,133/2,128 measurements, so low final coverage reflects more than its interval width.

Paired QA findings and replay. Seven measurement queries are averaged within person. Each outcome uses 22 planned BH-corrected contrasts across cohorts. Individualized-minus-shared system joint correctness is $-.05$ points on MDD (95% CI $[-.33, .23]$, $q = 1$); the control-by-individualization interaction is $-.23$ ($[-.61, .14]$, $q = .914$). ADNI joint correctness with control is zero for both assignments. Equal-answer-count risk comparisons have only 7/304 MDD people for individualized versus shared and 15/304 or 12/304 for control on versus off; ADNI risk is unestimable. These small accepted subsets provide no useful-coverage advantage. All outcome families and missing estimates remain in the local audit.

`scripts/run_common_properties.py` provides prepare, train, prepare-qa, qa and summarize stages. `results/common_properties/` retains plans, source hashes, training references, predictions, participant scores and unchanged responses; completed QA uses `qa_ascii/`. The 384-response grammar timing run is excluded. Audits reconstruct gold and normalization, verify hashes, recompute all 48 representation contrasts and replay QA scoring. Checkpoint inference is a separate reproduction step.

S6 Mixed-Description Alignment and Downstream QA

Unit records and archived permissions. This study implements the unit-level route in Section 3.6: frozen 64-dimensional vectors, mixed permitted descriptions, exact-text positives and an affine projection. The full-record experiment instead uses common attribute targets and equivalent-band positives. QA evaluates downstream record access. Explicit unit/field requests can use direct lookup; property matching is evaluated separately.

`semanticbn.claims/1.0` binds the brain vector to four-dimensional evidence, per-unit checks, allowed claims, denial reasons, controlled descriptions and a record-scoped evidence reg-

istry. Consumers recompile permissions and statements from source values; hashes check artifact consistency. `BSU_XX` is a record identifier, with permissions assigned by claim type.

Table S20: Detailed fields of the four linked records summarized in Section 3.3. Record availability is distinct from inclusion in the static vector.

Dimension	Recorded measurements
Spatial composition	ROI indices, available names, assignment and normalized pooling weights; atlas correspondence only with verified provenance.
Within-unit FC	Weighted mean a_{kk}, edge SD, and training-control reference deviation when available.
Between-unit FC	Signed profile $(a_{kl})_{l\neq k}$, strongest positive partner, and reference-network profile where supported.
Temporal dynamics	Window-wise blocks, within-FC SD, RMS of external-link temporal SDs, adjacent-profile RMS change; network-state occupancy and transitions when available.

The archived training-control deviations are separate from the all-training-cohort tertile bands in S2 and S5. The seed-7 archive retains one scan per person, with ADNI 91/19/20 and MDD 1,412/302/304 train/validation/test participants.

Historical permission yield (superseded). All permission counts in this section are historical software outputs. Among 15,036 units from 2,148 participants, the old policy marked 14,803 measurement-only, 173 additionally gradient-permitted and 60 additionally reference-associated units. The 60 ADNI reference-associated and 233 gradient-permitted units are not currently validated anatomical or gradient evidence: S4 supersedes the old local atlas flag for both cohorts. Disease-difference and membership-stability permissions were zero. Cached descriptions and QA gold retain the original policy; revised permissions have not been propagated through a new alignment/QA run.

Reference checks and gradient definition. For ROI-to-reference fractions Y_{ia}, soft spatial support is $f_{ka} = \sum_i Q_{ik}Y_{ia} / \sum_i Q_{ik}$. The archived association gate required a verified atlas flag, training-group hard overlap $\geq .35$, individual matched-reference support $\geq .35$, same-scan six-partner FC-profile correlation $\geq .50$, gradient coverage $\geq .20$ and compactness $q \leq .05$. These complementary checks share data and anatomy; they do not form a calibrated independent-evidence score.

For gradient coordinate g_i, ROI coverage c_i, eligible ROIs $\mathcal{I} = \{i : g_i, c_i \text{ finite}, c_i \geq .20\}$ and $w_{ik} = Q_{ik}c_i\mathbf{1}[i \in \mathcal{I}]$,

$$\mu_k = \frac{\sum_i w_{ik}g_i}{\sum_i w_{ik}}, \qquad \sigma_k = \sqrt{\frac{\sum_i w_{ik}(g_i - \mu_k)^2}{\sum_i w_{ik}}}, \qquad \gamma_k = \frac{\sum_i w_{ik}}{\sum_i Q_{ik}}.$$

Position and dispersion require two positively weighted eligible ROIs. Each of 199 null draws permutes g_i within $\mathcal{I}$ at fixed weights; lower-tail $p = (1 + \#\{\sigma_k^{\text{null}} \leq \sigma_k\})/200$, followed by within-scan BH correction. This null is spatially unconstrained. Current provenance withholds anatomical and gradient claims in both cohorts; fixed-Q temporal records and separate group tests supply neither membership repeatability nor individual disease authorization.

Original alignment and QA. There are 10,521/2,247/2,268 train/validation/test unit-description pairs. Frozen MiniLM produces normalized 384-dimensional mean-pooled embeddings, rejecting overlong inputs. Training-only standardized brain vectors feed an affine projection, trained with symmetric exact-description multi-positive contrastive loss, temperature .07, AdamW .001, at most 60 epochs and validation patience eight. The original run batches 128 shuffled unit pairs. Its cohort-wide B→T/T→B scores are 4.29/5.71% on ADNI and .89/.85% on MDD, versus 55.71/54.29% and 85.95/92.67% in seven-unit within-scan banks.

The QA pilot asks nine predefined claim kinds about the first unit of the first ten held-out people per cohort: spatial composition, internal FC, training-reference deviation, strongest partner, temporal

variability, reference association, gradient compactness, disease difference and membership stability. Of 180 requests, 105 are answerable and 75 require refusal under the archived policy. Frozen Qwen2.5-7B-Instruct selects claim/evidence IDs from serialized records. A verifier checks unit, claim kind and exact support; rejected outputs and retrieval-induced false refusals remain failures. The types-only identifier grammar permits arbitrary wrong IDs and empty-array refusals. The initial incompatible JSON-schema decoder run is excluded. This is fixed-template statement selection.

S6.1 Matched Representation Follow-Up

Fixed inputs and architectures. The comparison reuses all 2,148 participants, seed-7 assignments Q, archived permissions, descriptions and MiniLM targets: 1,503/321/324 train/validation/test participants. Seeds 7/11/19 change only alignment initialization and minibatch order. SemanticBN freezes its 64-dimensional vectors and learns an affine text projection. FC-linear pools 90-dimensional signed FC rows under the same Q. FC-MLP applies two shared 90-dimensional Linear/ReLU/dropout(.2) layers before the same pooling and projection. The dual-pathway adaptation (Han et al., 2026) sums a linear pooled-FC branch and a temporal graph branch: width-three 1D convolution with 64 channels, ReLU, valid-length max pooling, a 64-dimensional linear layer and one four-head GAT on positive FC edges with self loops. Unit-pooled graph features receive an affine projection. Trainable counts are 24,960/34,944/51,324/68,608, respectively.

The FC-MLP and dual-pathway encoders train on descriptions; SemanticBN's community encoder stays frozen. The dual-pathway control adapts RethinkingBCA commit `cde613fbc687e791141a637c9ad4039e41437ddb`, using PyTorch Geometric 2.6.1 GATConv, smaller temporal widths and unit-description readouts. It is a retrieval adaptation, rather than an official pretrained-model reproduction.

All methods use exact-description multi-positive contrastive loss, temperature .07, AdamW .001, weight decay .0001, batches of 16 scans (112 pairs), at most 60 epochs and validation patience eight. Seed-matched minibatch orders and training-only normalization are shared. Temporal inputs preserve original row order and length, with within-scan ROI standardization and masked padding. Recomputed FC matches archived off-diagonal FC within 4.42×10^{-7} for ADNI and 2.99×10^{-8} for MDD. Both ADNI and MDD retain unresolved upstream column-to-atlas binding; no TR equivalence is assumed.

Table S21: Matched individual-description retrieval (%, mean ± sample SD over three initialization seeds). Cohort-wide pools contain 140 ADNI or 2,128 MDD held-out units; within-scan pools contain seven. B/T denote brain/text. Methods reuse each participant's frozen $Q^{(s)}$ and targets, not one population-wide partition.

Cohort	Representation	Cohort B→T	Cohort T→B	Scan B→T	Scan T→B
ADNI	SemanticBN	4.29 ± 0.00	4.52 ± 0.82	53.10 ± 0.41	53.81 ± 1.09
	FC-linear	4.29 ± 0.71	2.62 ± 1.80	47.38 ± 0.41	40.71 ± 0.71
	FC-MLP	8.33 ± 2.51	8.10 ± 3.67	56.90 ± 2.18	54.76 ± 5.17
	Dual-pathway	3.81 ± 1.09	5.71 ± 0.71	38.57 ± 3.98	52.86 ± 0.71
MDD	SemanticBN	0.75 ± 0.05	0.66 ± 0.17	85.56 ± 0.23	92.51 ± 0.05
	FC-linear	0.55 ± 0.07	0.63 ± 0.07	82.58 ± 0.18	91.34 ± 0.14
	FC-MLP	1.61 ± 0.18	3.23 ± 0.24	93.33 ± 0.14	96.77 ± 0.21
	Dual-pathway	0.53 ± 0.03	0.89 ± 0.05	82.61 ± 0.25	91.87 ± 0.12

Participant scores average seven units and three alignment seeds before 10,000 paired sign flips and 5,000 participant bootstraps. BH correction covers 24 contrasts: three alternatives, two cohorts and four retrieval metrics. Intervals condition on fixed splits, communities and candidate banks. Exact duplicate descriptions are positives; evidence-equivalent paraphrases use the separate S2 protocol.

Table S22: FC-MLP minus matched SemanticBN in cohort-wide retrieval. Differences and bootstrap intervals are percentage points, after averaging units and initialization seeds within participant. BH correction uses all 24 retrieval comparisons, not only these four rows.

Cohort	Direction	Difference	95% CI	Paired p	BH q
ADNI	B→T	4.05	$[-.24, 8.57]$	.126	.251
ADNI	T→B	3.57	$[-1.43, 8.82]$	.237	.379
MDD	B→T	.86	$[.39, 1.33]$	.00030	.00120
MDD	T→B	2.57	$[1.93, 3.20]$	.00010	.00048

Matched QA contexts and decoding. All eight contexts below use the same 180 questions and records, with seed 7 fixed for learned indexes. Qwen remains frozen in FP16, using greedy decoding, repetition penalty 1.1, a 512-token output cap, 32,768-token context ceiling and the audited identifier grammar. Follow-up inference uses vLLM 0.6.6.post1, four concurrent sequences and batches of 16 on L40S GPUs. Top-three inputs average 6,198–6,289 tokens, full context 14,146 and no evidence 176; these are shared limits rather than equal context lengths.

The original brain projection yielded 70.00% E2E; its unchanged-prompt follow-up yielded 67.78% with changed inference batching, including two capped outputs. The newly matched SemanticBN projection yields 64.44% and uses the different scan-batched alignment schedule above. These run identities remain distinct. The matched-index run has no capped or schema-invalid outputs. Direct-text retrieval ranks the same archived descriptions with MiniLM, without a prepended unit-ID field.

Table S23: Downstream matched record-selection QA. End-to-end (E2E) accuracy credits exact claim-and-citation answers or correct required refusals, over all 180 requests. Retrieval coverage uses the 105 answerable requests. All conditions use the same questions; the full-context advantage is retained as a negative result for learned retrieval.

Context/index	E2E (%)	Retrieved/answerable
SemanticBN, matched	64.44	50/105
FC-linear	60.00	41/105
FC-MLP	61.67	43/105
Dual-pathway	65.56	51/105
Archived brain index, rerun	67.78	62/105
Direct text	61.67	45/105
All seven units	92.22	105/105
No evidence	41.67	0/105

Matched SemanticBN retrieves 50/105 answerable targets and answers 44 of those correctly (88.00% conditional joint accuracy). Full context supplies all 105 targets and yields 96 jointly correct answers; five unsupported answers among 75 forbidden requests are rejected. No-evidence E2E is the 75/180 always-refuse baseline. QA thus favors complete records over the evaluated retrieval routes. A deterministic lookup using the already supplied unit/claim fields solves 180/180 requests by the same permission rules; this is an implementation control.

Paired QA analysis and artifacts. Nine outcomes are averaged within each of 20 people before 10,000 paired sign flips and 5,000 bootstrap draws. BH correction spans all seven alternatives below. These intervals describe the fixed question set and records; they exclude retraining and question-author uncertainty.

Table S24: Alternative minus matched SemanticBN E2E, participant-paired. Differences and 95% bootstrap CIs are percentage points; p is two-sided, with BH correction across all seven rows.

Alternative	Difference	95% CI	Paired p	BH q
FC-linear	-4.44	$[-11.11, 1.67]$	.271	.474
FC-MLP	-2.78	$[-14.44, 8.89]$	.663	.747
Dual-pathway	1.11	$[-7.22, 9.44]$	.747	.747
SemanticBN (archived)	3.33	$[0.00, 6.67]$	.111	.260
Direct text	-2.78	$[-12.78, 6.67]$	.636	.747
All seven units	27.78	$[18.89, 36.11]$	.00010	.00035
No evidence	-22.78	$[-31.11, -14.44]$	.00010	.00035

`scripts/run_semantic_pipeline.py` and `results/semantic_contract/` retain the original contracts, projection and pilot. Follow-up outputs and paired/citation/cohort breakdowns remain under `results/recent_comparisons/`; frozen-pilot audits are in `results/paper_revision/`. The `--recent-comparisons` mode of `scripts/audit_paper_revision.py` checks target/index hashes, recomputes all 12 retrieval runs from saved embeddings and replays full/visible answerability and verifier decisions. Historical outputs remain unchanged.

S7 Structural Properties and Validation Priorities

Reconstruction diagonal. In Equation 3, $\sum_k Q_{ik} = 1$ implies $(QQ^\top)_{ii} \geq 1/K$ while $A_{ii} = 0$. The normalized reconstruction loss therefore includes a diagonal contribution of at least $1/(RK^2)$, minimized by uniform memberships. Entropy minimization favors sharp assignments. FC-normalization and modularity denominators are floored at 10^{-6}; entropy clamps memberships below at 10^{-8}.

S3's objective ablations evaluate organization only. A matched downstream comparison of the original, off-diagonal-only and reconstruction-free objectives remains unrun. It must select using a prespecified validation criterion and refresh assignments, features, community-dependent measurements, permissions, descriptions and alignment scores. FC sources, reference queries, gold attributes, training bands and participant splits can remain fixed. Current language results retain the original fits.

Static versus temporal scope. For centered series $\bar{X}$ and a common time-point permutation P, $(P\bar{X})^\top(P\bar{X}) = \bar{X}^\top\bar{X}$. Pearson FC and deterministic FC-only representations are invariant, whereas contiguous-window measurements depend on temporal order. This distinguishes the static vectors from the temporal records; S8 summarizes the separate dynamic study.

Functional interpretation requires verified ROI provenance, spatially appropriate gradient nulls, matched-unit repeat observations and independently assessed descriptions. S2 now tests restricted pooled-vector access; it does not isolate spatially local measurements or validate functional meaning. The remaining checks are outstanding evaluations rather than implemented components of the main algorithm.

S8 Supporting Analyses

These earlier experiments use their own representations, targets and split protocols. They are summarized here to retain the findings relevant to the main study without conflating them with S2's static matching task. Original aggregate outputs remain in the local experiment directories.

Historical category retrieval. Frozen MiniLM targets were seven predefined network-category descriptions, expanded to 35 candidates with fixed text negatives. The group-level gate used overlap $\geq .35$ and gradient coverage $\geq .20$. Alignment used AdamW .001, weight decay .0001, temperature .07, a 150-epoch cap, patience 20 and seeds 7/17/27. Shared, individualized and fixed-reference paraphrase B$\rightarrow$T scores were $.760 \pm .058$, $.871 \pm .027$ and $.932 \pm .009$; these are scan-averaged category scores, not individual-property description matching. Fixed reference remained stronger.

Source-only category transfer also favored fixed reference: paraphrase B→T was .795/.963 for individualized/fixed ADNI→MDD and .706/.767 for MDD→ADNI. Different source gates and unverified ROI correspondence limit comparison across directions. These historical experiments supply neither single-unit semantic validation nor cross-cohort validation of the connected QA route.

Figures S1 and S2 visualize the historical category results and participant-paired gains, respectively.

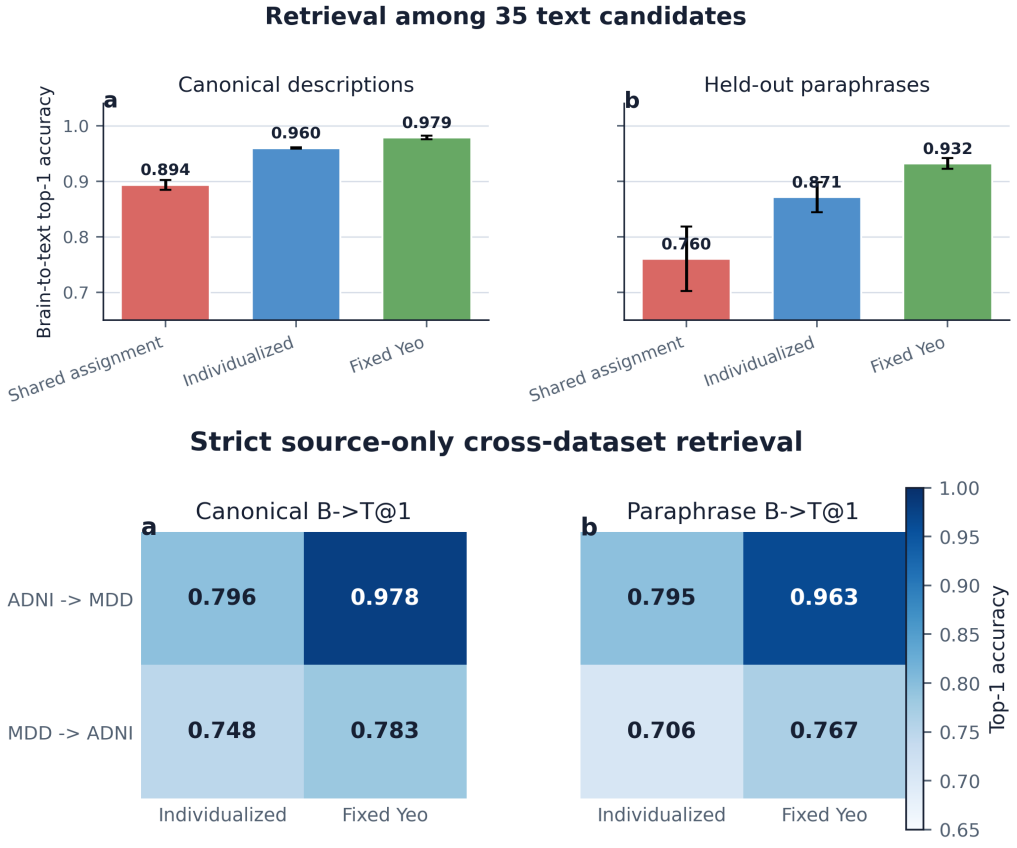


Figure S1: Historical network-category retrieval. Top: 35-text canonical and paraphrase accuracy, with seed-SD error bars. Bottom: source-only cross-cohort transfer. These targets are predefined category descriptions, separate from S2's three-property descriptions. "Fixed Yeo" denotes the historical fixed-reference control; ROI-to-atlas correspondence remains unverified in both cohorts.

Group associations. The ADNI analysis aggregates scans within 130 people labeled NC/EMCI/LMCI (48/48/34), rather than AD versus controls. Across three seeds, it has 0/75 FDR-significant feature tests. MDD Welch tests yield 52/75 on 2,018 people (maximum $|d| = .236$); diagnosis coefficients from OLS with HC3 errors, controlling age, sex, motion and site on 1,604 matched people, yield 26/75 (maximum standardized $|\beta| = .183$). BH correction is within each seed's 25-test family. These exploratory group associations do not authorize individual disease statements; anatomical interpretations remain conditional on S4.

Figure S3 separates adjusted effect size from corrected significance.

Dynamic records and downstream augmentation. The same 442 ADNI/2,018 MDD scans use fixed full-scan Q. Sliding-window FC produces signed 7×7 block means $C^{(w)}$, excluding ROI self-pairs. The primary window/step is 40/10 stored samples. For $s_{kl} = \mathrm{SD}_w(C_{kl}^{(w)})$, the three summaries are

$$v_k = s_{kk}, \qquad e_k = \sqrt{\frac{1}{6}\sum_{l \neq k} s_{kl}^2}, \qquad d_k = \sqrt{\frac{\sum_{w=1}^{W-1}\sum_{l=1}^{7}(C_{kl}^{(w+1)} - C_{kl}^{(w)})^2}{7(W-1)}}.$$

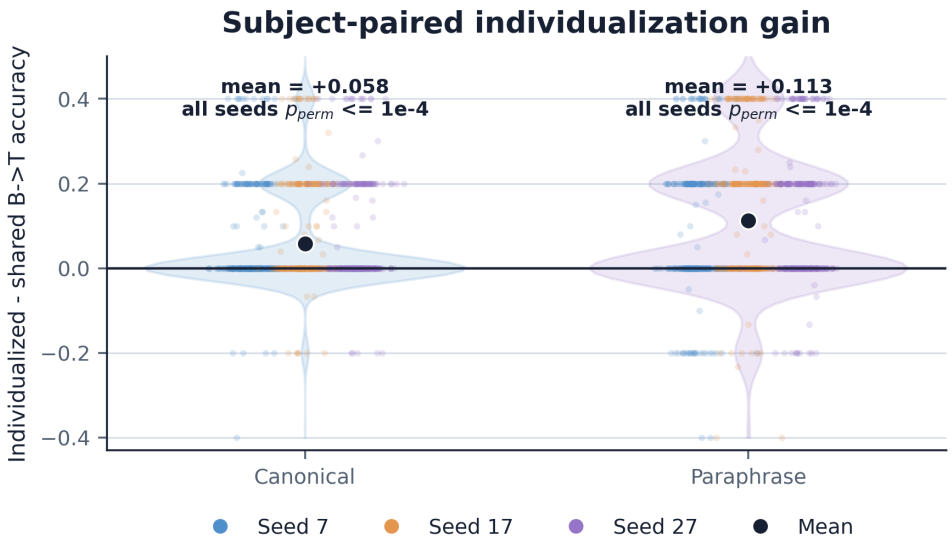


Figure S2: Participant-paired individualized-minus-shared category-retrieval accuracy. Each point is one participant within a seed; repeated scans are averaged before comparison. Violins show the distributions and black points their means: .058 canonical and .113 paraphrase, distinct from the scan-averaged differences above. Each seed's sign-flip test gives $p \leq 10^{-4}$. The same participants recur across seeds and are not independent additional samples.

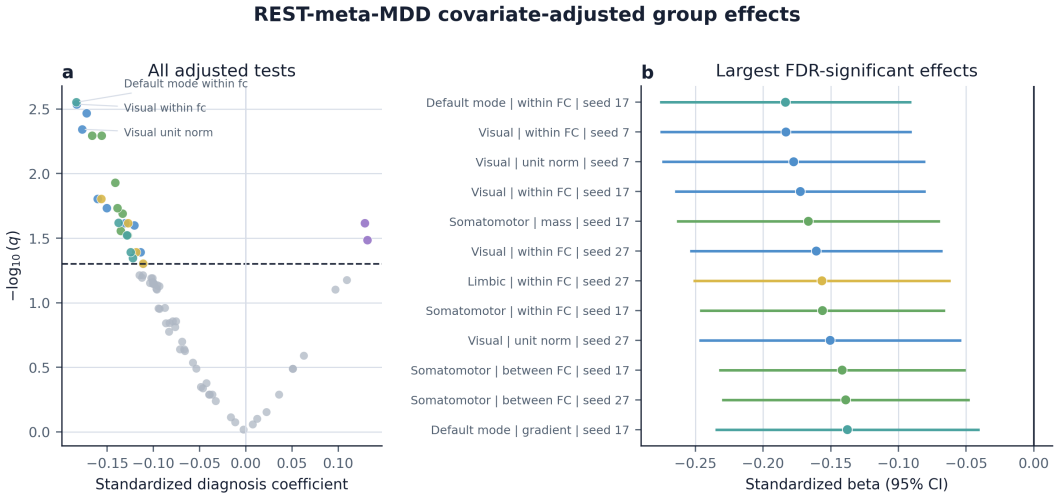


Figure S3: Exploratory MDD group associations (left) and the twelve strongest FDR-selected adjusted effects (right). Coefficients control age, sex, motion and site; intervals use HC3 standard errors. BH correction is within each seed's 25-test family. Network names are provisional reference labels, and seed-specific units are not matched across fits. These are group-level associations, not validated individual disease interpretations.

TR and censoring metadata remain unbound, so timing uses stored samples; these features measure connection variation at fixed memberships. Three/four/five-state K-means uses training windows only, with four states fixed as primary. Window sensitivities use 30/10, 50/10 and 40/5; split-half estimates share full-scan Q. For 20 held-out people per cohort, 199 shared-phase surrogates preserve auto/cross spectra and full-scan FC; BH correction spans 60 tests per cohort/assignment. Individualized MDD rejects none; its split-half correlations are .169/.187/.258 for within variation, external variation and adjacent change.

The subject-disjoint 70/15/15 splits and seeds 7/17/27 compare static, appended-dynamic and shuffled-dynamic inputs. Retrieval appends nine measurements to each 64-dimensional unit; diagnosis augments 28 static block means with 21 compact dynamic features, four occupancies and switching rate. Training-only standardized, class-balanced logistic regression selects $C \in \{.01, .1, 1, 10\}$ by validation BA. MDD category-paraphrase retrieval changes from .9522 to .9518, and BA from .5759 to .5626. Dynamic group tests yield 0/45 significant results in each analysis family (15 tests per seed). The frozen Qwen-1.5B numeric-argmax pilot scores 17.5% on 120 questions, separate from the later 7B statement-selection tasks. These findings provide no stable downstream augmentation benefit.

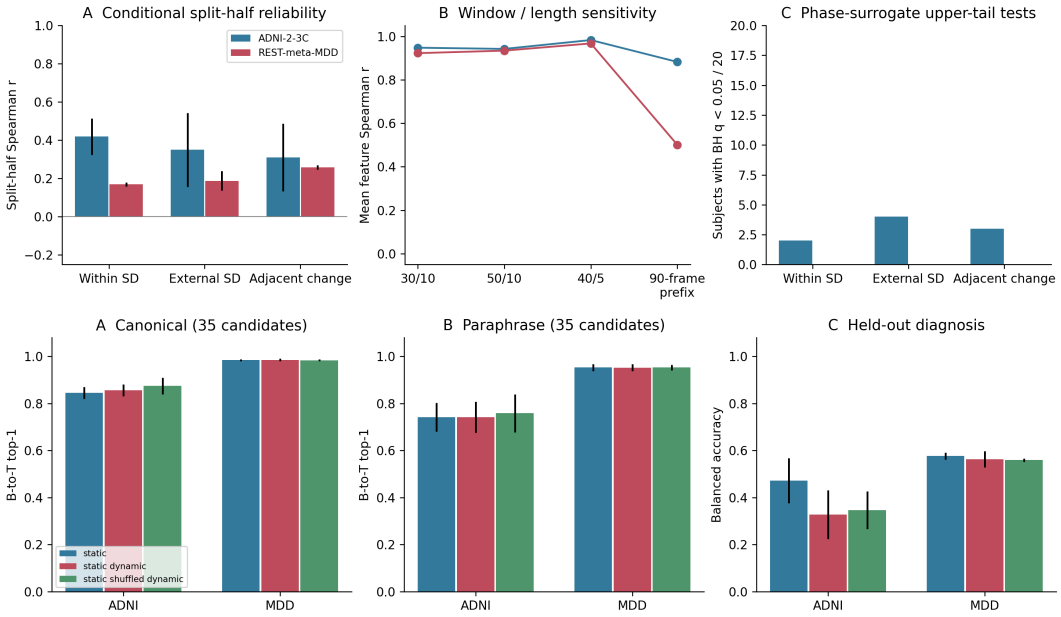


Figure S4: Dynamic records and downstream augmentation. Top: conditional split-half correlations, window/length sensitivity and phase-surrogate rejections. Windows and steps use stored samples; split halves share the full-scan assignment. Bottom: static, dynamic-augmented and shuffled-dynamic inputs for category retrieval and diagnosis. Error bars show seed SD. MDD shows no stable augmentation gain; ADNI diagnosis labels are NC/EMCI/LMCI. These analyses measure connectivity variation at fixed memberships, rather than independent scan–rescan stability.

Figure S4 combines the dynamic checks with their downstream comparisons.

Cross-archived-series recovery. A separate ADNI proxy uses the first two distinct base-series IDs for 95 people, with five participant folds (61/15/19 train/validation/test), seeds 7/11/19 and feature ceilings 28/112/448. Only two pairs have confirmed acquisition intervals (81/91 days); the task is cross-archive recovery, not verified short-interval repeatability. Source-only community fits feed training-standardized ridge heads with validation-selected penalties; targets are the other archive's 21 reference attributes. At the primary 448-scalar budget, individualized/shared/fixed-reference normalized MAE is .8277/.8288/.8276; a 21-scalar source-summary ridge gives .8203 and the training mean .8280. Individualized minus the validation-selected compact baseline is $-.00776$ (95% CI $[-.02706, .01088]$, BH $q = .715$ over 36 contrasts), with improvement in two of five folds. The prespecified 5% improvement screen fails.

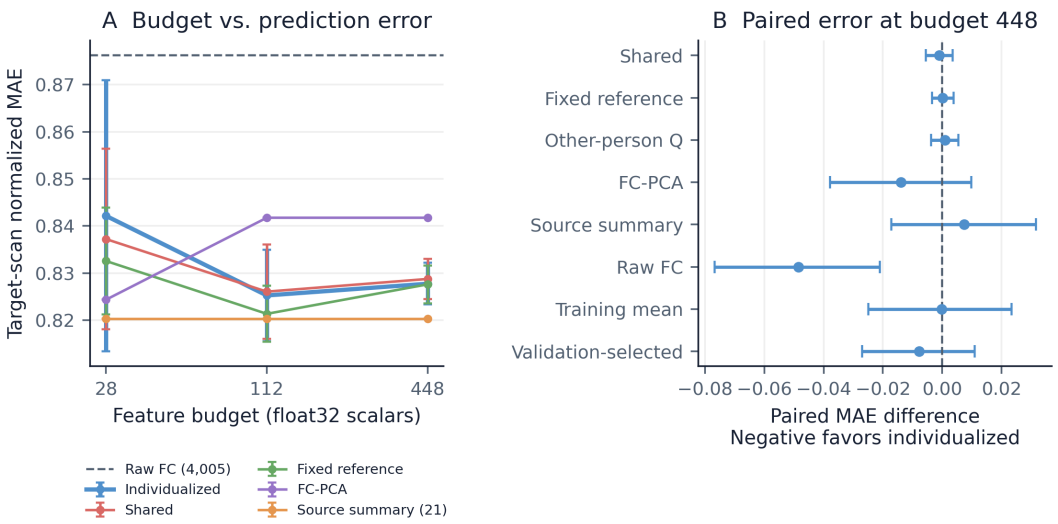


Figure S5: Cross-archived-series attribute recovery in 95 ADNI participants. Left: normalized MAE across feature ceilings, with seed-SD error bars. Right: paired differences at the primary 448-scalar ceiling, with conditional participant-bootstrap 95% CIs; negative differences favor individualized pooling. The raw-FC control uses 4,005 values, outside these ceilings. Acquisition intervals are confirmed for only two archive pairs, so this evaluates cross-archive recovery rather than verified test–retest reliability.

Figure S5 shows the budget sensitivity and paired uncertainty.

Separate MDD disease-recognition study. The 116-column route retains 1,976 people, split 1,383/293/300 with 1,077 cases and 899 controls overall. Training-mean FC K-means defines seven shared units. Each unit stores internal mean/SD and six external means: 56 entries containing 35 unique measurements. Seven 8–32–8 GELU encoders feed additive score heads, numeric decoders and MiniLM projectors. Disease loss is class-balanced BCE; numeric MSE and same-unit text contrastive losses have weights .1 each and temperature .07. Training uses AdamW .001, weight decay .0001, batch 128, at most 120 epochs and validation-BCE patience 20; BA thresholds use validation predictions. Raw-FC controls read all 6,670 edges. Logistic regularization is selected on validation AUROC from $C = .001, .01, .1, 1, 10$. All primary methods are retained below.

Table S25: MDD/control classification on 300 test participants. Mean ± seed SD; deterministic rows have zero initialization variation. BA: balanced accuracy. Brier is lower-is-better.

Method	AUROC	BA	Brier
Raw FC / logistic	.6663	.6186	.2352
Raw FC / MLP	.6564 ± .0253	.5910 ± .0171	.2534
Measured state / logistic	.5770	.5373	.2489
Measured state / MLP	.5518 ± .0087	.5202 ± .0284	.2506
Covariates / logistic	.4974	.4862	.2512
Training prevalence	.5000	.5000	.2459
State: disease only	.5488 ± .0150	.5191 ± .0183	.2509
State: + numeric	.5472 ± .0134	.5141 ± .0086	.2509
State: + text	.5497 ± .0154	.5255 ± .0163	.2509
State: + numeric + text	.5477 ± .0138	.5150 ± .0111	.2509
State: + numeric + shuffled text	.5472 ± .0133	.5177 ± .0148	.2509

Seeds are 7/11/19; deterministic logistic rows have no independent initialization repeats. On participant-wise seed-averaged scores, joint-state AUROC is .5547, distinct from the .5477 mean per-seed AUROC. The paired joint-minus-raw-logistic difference is −.11163 (5,000 stratified par-

ticipant bootstraps: 95% CI $[-.18358, -.03447]$; 5,000 within-person swaps, BH $q = .0230$ over ten contrasts). Five site-held-out sensitivity folds give pooled out-of-fold AUROC .5288 for joint states versus .6330 for raw logistic; pooled scores can also reflect between-fold calibration. These developmental comparisons favor the full-connectivity controls.

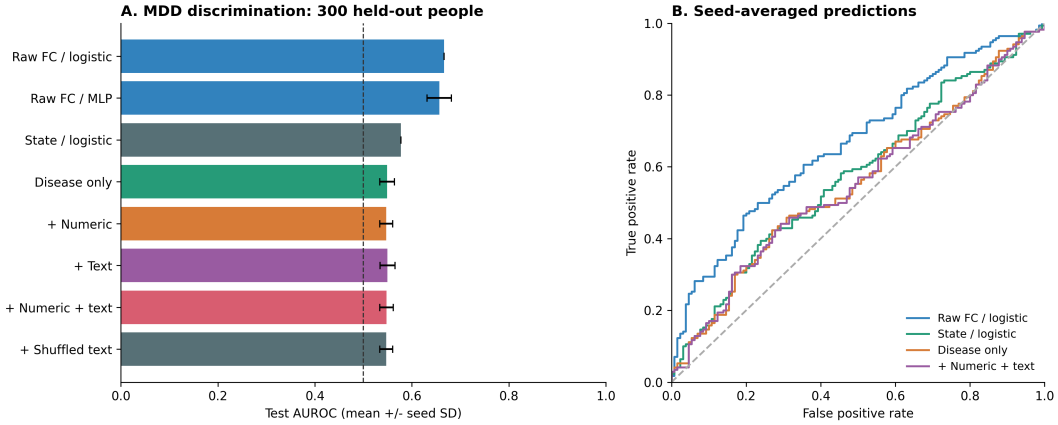


Figure S6: Separate MDD/control classification study on 300 held-out participants. Left: mean per-seed AUROC, with seed-SD error bars. Right: ROC curves use participant-wise seed-averaged scores, so their AUROCs differ from the means on the left. Raw-connectivity logistic regression remains stronger than the joint state model. This disease-recognition task is separate from the main static brain–text matching protocol.

Figure S6 visualizes the classification comparison and ROC curves.

A subsequent training-only block-PCA control preserves 1/2/4/8 components in each of 28 connection blocks. Its primary AUROC is .5639/.6143/.6436/.6428; all four ranks are reported, with no test-based selection. Post-run paired gains over measured-state logistic are nonsignificant after the eight-contrast BH correction. Equal feature count alone therefore does not establish equal retained information.

Local audit trail. The retained historical analyses are implemented by the category-language runners, `scripts/run_disease_effects.py`, the dynamic-signature runner, `scripts/run_representation_value.py`, `scripts/run_brain_states.py` and `scripts/run_predictor_evidence.py`. Their original protocol, result and audit files remain unchanged in the project. The compact figures retain the principal comparisons; repeated per-seed narrative and noncore QA tables are omitted. The review attachment's executable coverage is specified in S1.

S9 AUXILIARY QUERY AND DELIVERY AUDIT

Protocol and statistical unit. This separate evidence-access audit reuses S8's 300 MDD test people and two frozen predictors (unit-state and raw-FC logistic), producing 600 records. Relative contributions are recomputed from the actual predictor and training-control reference. Raw-FC edge terms use standardized upper-triangle connections; internal edges count once and cross-unit edges contribute half to each endpoint. Measurement and decision evidence retain separate source/model bindings. Arithmetic replay checks computations, not causal interpretation.

A frozen Qwen2.5-7B-Instruct maps each natural-language request to a five-field query: operation, unit, order, comparison and contribution sign. Only the question and schema enter the model. FP16 greedy decoding uses repetition penalty 1.1, 512 output tokens and a 16,384-token context ceiling. The unchanged executor rejects invalid combinations and checks evidence availability; offline gold never enters execution. The original direct-selection/query comparison used 40 people and 320 requests per predictor. E2E was 34.69/87.81% for state records and 37.50/89.38% for raw-FC records. Inputs averaged 13,062 versus 313 tokens; these are whole-route comparisons, not equal-context tests.

The follow-up freezes 96 in-scope and eight unsupported-operation questions before 104 new model calls. These agent-authored questions remain a development bank pending independent assessment. In-scope requests cover internal/external comparisons, ranking, conjunctions, provenance, clinical scope and missing decision evidence. Each unique parse is reused across all records, plus seven synthetic fixtures distinguishing signed/absolute ranking and strict comparisons. Five thousand bootstrap draws resample question-level means after record averaging; query–record executions are not independent language samples.

Table S26: Parsing accuracy and execution agreement. All-record agreement requires the expected result on all 600 real and seven synthetic records under the specified evidence visibility. A wrong plan can meet this criterion when missing evidence blocks execution. The old and new question banks are not difficulty-matched.

Wording bank	Exact	Invalid	Wrong rank	Other operation/scope	Replay agreement
Original, 112 texts	81	24	3	4	85
New in-scope, 96 texts	73	14	3	6	78

Of the new bank's six operation/scope errors, five request unavailable decision terms and receive correct executor refusals; one clinical question becomes an irrelevant provenance answer. This explains 78/96 replay agreements versus 73/96 exact parses. All eight unsupported-operation plans fail to request refusal: six violate applicability; two approximate OR by AND or top-two by top-one. Accidental empty sets do not establish scope recognition.

New-bank E2E is 82.60% for state records (query-bootstrap 95% CI [74.86,89.58]) and 83.66% for raw FC ([76.13,90.63]). Supported but wrong answers number 810/28,800 and 507/28,800. Uniform replay of the old ranking errors also finds 207 wrong raw-FC answers, despite zero in the original smaller allocation. The authored banks and record weightings remain separate.

Same-output checks. Structured checks validate schema, identifiers and path/value types. Provenance adds trusted person, source, model and cited-value bindings. Full dual-evidence checks additionally enforce claim type, permitted wording, visibility and exact evidence support. Every tier shares trusted reconstruction and executor evidence-availability guards; none sees the request or gold. This audit isolates added delivery checks while holding reconstruction and evidence-availability guards fixed.

Table S27: Identical-output delivery checks. S/P: structured and provenance give identical results; S/P/D: all three tiers agree. Accepted excludes refusals. Retained/lost refers to originally correct nonempty answers, not all answerable requests.

Route	Predictor	Policy	Accepted / requests	Wrong / accepted	Retained / lost
Direct	State	S/P	138/320	113/138	25/0
Direct	State	Dual	43/320	18/43	25/0
Direct	Raw FC	S/P	151/320	124/151	27/0
Direct	Raw FC	Dual	35/320	8/35	27/0
Query, original	State	S/P/D	199/320	5/199	194/0
Query, original	Raw FC	S/P/D	193/320	0/193	193/0
Query, new wording	State	S/P/D	20,064/28,800	810/20,064	19,254/0
Query, new wording	Raw FC	S/P/D	19,698/28,800	507/19,698	19,191/0

On actual executor outputs, all three policies make identical acceptance decisions. Direct-selection filtering retains the 25/27 correct nonempty state/raw-FC outputs but lowers coverage from 43.13/47.19% to 13.44/10.94%. Its gain is reduced selective risk at reduced coverage. Additional dual-evidence rejection is observed in deliberately injected type substitutions, measurement-to-decision misuse, unsupported causal wording and wrong support sets (600 cases each); provenance already rejects wrong-person, wrong-model and changed-value faults. All tiers accept clean and true-but-irrelevant statements. These constructed tests assess implemented checks, not natural failure prevalence or question relevance.

Artifacts. `scripts/run_query_audit.py` and `results/query_audit/` retain the frozen manifest, questions, expected/predicted plans, raw outputs, per-record execution decisions and same-output delivery replay. The original route remains under `results/predictor_evidence/`. Hashes bind the old evidence, responses, executor and new checker. This auxiliary audit is separate from the representation and language-matching experiments.