WHO DID WHAT TO WHOM? RELATIONAL STRUCTURE IN EEG SEMANTIC DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering structured meaning from neural activity is a central goal of brain-to-language decoding and a key step toward natural communication through brain-computer interfaces. Recent semantic-unit decoding frameworks provide an interpretable route to recovering semantic content from EEG, yet content alone does not specify its relational organization. This leaves a fundamental boundary between recovering what is represented and identifying who did what to whom. To study this boundary, we develop a three-stage evaluation framework that separates role-neutral semantic content from relational organization. (1) For **functional relevance**, text-side content-controlled structural interventions show that correct predicate-argument binding improves fidelity to the intended proposition beyond semantic content alone. (2) For **structural readout**, contextual representations support reproducible role-related readout beyond retrieved semantic-unit identity, while larger relation-predictability effects are also reproduced by non-neural controls, limiting their interpretation as EEG-specific structure. (3) For **occurrence-specific attribution**, EEG conditioning yields small gains beyond semantic content, but matched-record specificity remains weak and heterogeneous. Together, these results identify relational structure as a distinct target for EEG semantic decoding and provide a principled basis for separating functional relevance, structural readout, and occurrence-specific neural attribution.

1 INTRODUCTION

Conditions such as aphasia and locked-in syndrome can severely impair communication through speech or writing, motivating EEG-based brain–computer interfaces that seek to recover language directly from neural activity. Existing EEG-to-language approaches broadly follow two paradigms: speech decoding, which targets phonetic, articulatory, or acoustic representations, and concept decoding, which seeks to recover higher-level semantic content from distributed neural activity (Dronkers et al., 2017; Moses et al., 2021; Silva et al., 2024; Zhang et al., 2024a). Within concept decoding, earlier approaches commonly formulated semantic recovery as prediction over predefined concepts or categories, whereas recent work has increasingly moved toward continuous semantic alignment and richer language reconstruction (Wang et al., 2011; Correia et al., 2015; Foster et al., 2021; Duan et al., 2023; Lu et al., 2025; Shams et al., 2025). Despite increasingly rich outputs, existing EEG semantic decoding studies have primarily emphasized concept identity, cross-modal semantic alignment, or sentence-level reconstruction, while the relational organization among decoded concepts has received comparatively little explicit, controlled evaluation. As concept decoding becomes increasingly expressive, the choice of intermediate representation between noisy, data-limited EEG and natural language becomes a central challenge: such a representation must preserve communicative meaning, remain learnable from limited neural data, and be sufficiently explicit and interpretable to support precise claims about what linguistic information has actually been decoded. Semantic Intent Decoding (SID) provides one such interface by representing communicative meaning as a flexible, variable-sized set of semantic units (Li et al., 2026). In its original formulation, these semantic units are decoded from EEG-derived representations, aligned with an open continuous semantic space, and used to support subsequent language reconstruction. This formulation moves EEG semantic decoding beyond assigning a neural response to a single predefined category toward an inspectable content-level representation of which semantic concepts are present in the intended message, providing the starting point for our study of richer proposition-level structure.

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.

This content-level interface nevertheless has a relational boundary: semantic-unit identity alone does not uniquely determine proposition-level organization. For example, the same semantic units $\{girl, chase, dog\}$ are compatible with both "The girl chases the dog" and "The dog chases the girl." The constituent concepts are unchanged; what differs is their predicate–argument binding: which participant fills which semantic role with respect to the predicate.

Prior linguistic and neuroimaging work has shown that sentence meaning depends on compositional role structure and that predicate–argument or thematic-role information can be decoded from controlled neural recordings (Friederici, 2011; Zhang & Pylkkänen, 2018; Frankland & Greene, 2015; 2020; Giglio et al., 2024; Morgan et al., 2025; Wang et al., 2024). Unlike these controlled paradigms, we ask whether such structure is identifiable within naturalistic non-invasive EEG semantic-decoding pipelines and whether successful readout depends on the matched neural record. Importantly, this boundary does not arise from permutation invariance itself. Relational structure can itself be represented in a permutation-invariant form. For example, the labeled bindings $\{\mathrm{ARG0} = girl, \mathrm{ARG1} = dog\}$ retain the same relational meaning regardless of their enumeration order. What is missing is therefore neither surface word order nor a full syntactic parse, but the proposition-level predicate–argument organization that binds the constituent semantic units to their roles. We therefore treat predicate–argument organization as the relational variable of interest, separating which semantic content is present from how that content is organized into a proposition.

A positive relational readout from an EEG-derived representation is not, by itself, sufficient evidence for occurrence-specific neural attribution (Hewitt & Liang, 2019; Kriegeskorte & Douglas, 2019). A probe may predict semantic roles or predicate–argument relations because lexical regularities, structure in the target distribution, or decoder-related priors make these labels predictable even when the correctly matched neural record is not required (Ravichander et al., 2021). This ambiguity is particularly acute in naturalistic EEG semantic decoding, where paired neural–linguistic data are limited and diagnostically useful same-content/different-binding counterfactuals are sparse, making it difficult to separate content-related predictability from occurrence-specific relational information using readout alone. Within the neural-attribution analysis, we organize the evidence into three progressively stricter requirements: representation-level relational readout asks whether relational structure is predictable from a representation; conditional EEG predictive capacity asks whether EEG conditioning adds predictive information beyond role-neutral semantic content; and matched-record specificity asks whether that additional predictive information depends on the correctly corresponding neural occurrence. Only converging evidence across these levels can support occurrence-specific attribution; accordingly, we distinguish what can be read out, what additional information EEG contributes, and whether that contribution depends on the matched neural record. Figure 1 summarizes this representational boundary and the three-stage evidence framework.

Together, these representational and evidential boundaries motivate three questions spanning functional relevance, structural readout, and neural contribution and specificity. **RQ1:** Is knowing which semantic units are present sufficient to determine who did what to whom, or does explicit predicate–argument binding contribute information beyond the content set itself? **RQ2:** Beyond retrieved semantic-unit identity, do EEG-derived contextual representations support additional role-related readout? **RQ3:** If EEG conditioning adds predictive information, does that advantage depend on the correctly matched neural occurrence? Our results reveal a corresponding gradient of evidence. For RQ1, with semantic content held fixed, gold binding attains a tie-adjusted Gold preference of 78.02% (95% CI, 73.49–82.33%) across 232 eligible comparisons, establishing the functional relevance of predicate–argument organization beyond semantic-unit identity. For RQ2, contextual representations show modest but reproducible gains in role-related readout beyond retrieved semantic-unit identity across both EEG corpora. However, the substantially larger relation-predictability effects are also reproduced by lexical and decoder-related controls, limiting their interpretation as uniquely EEG-derived structure. For RQ3, EEG conditioning provides small additional predictive capacity beyond content-only comparators, while matched-record specificity remains weak and heterogeneous across datasets and role endpoints. Taken together, the evidence is strongest for the functional relevance of relational structure and representation-level readout, and substantially weaker for general occurrence-specific neural attribution. Our contributions are threefold:

(1) We introduce an identifiability-oriented evidence framework that separates functional relevance, representation-level structural readout, and neural attribution, with the latter further distinguished into conditional EEG predictive capacity and matched-record specificity.

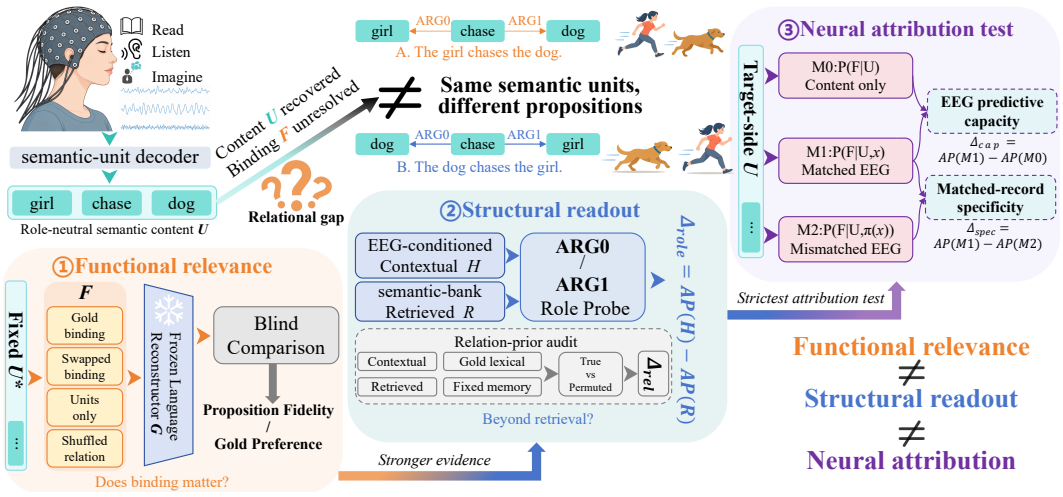

Figure 1: **Relational-structure problem and progressively stricter evidence framework.** The same role-neutral semantic content U can support different predicate–argument organizations and thus distinct "who did what to whom" propositions. **(1)** With U fixed, Gold binding is compared with swapped-binding, units-only, and shuffled-relation controls to test the functional relevance of explicit relational organization. **(2)** Contextual representations H and retrieved representations R are evaluated with the same post-hoc role probes; Δ_{role} measures readout beyond retrieval, while the relation-prior audit tests alternative feature sources under True-versus-Permuted labels. **(3)** Content-only (M_0), matched-EEG (M_1), and mismatched-EEG (M_2) conditions separate conditional EEG predictive capacity ($M_1 - M_0$) from matched-record specificity ($M_1 - M_2$). The stages impose increasingly stringent requirements from functional relevance to structural readout and neural attribution. Full semantic-role notation is provided in Table B1.

(2) We establish the relational boundary of role-neutral semantic-unit representations through content-controlled interventions, showing that explicit predicate–argument binding contributes proposition-level information beyond semantic-unit identity.

(3) We demonstrate an empirical dissociation between representation-level decodability and occurrence-specific neural attribution across two EEG corpora and multiple task conditions: contextual role-related readout is reproducible and conditional EEG predictive capacity is small, while matched-record specificity remains weak and heterogeneous.

2 Method

2.1 Relational Representation and Binding Relevance

For RQ1, we first define the relational quantity of interest by separating role-neutral semantic content from its proposition-level organization. Building on the semantic-unit interface of Semantic Intent Decoding (SID), we formalize this distinction as:

$$U = \{u_1, \ldots, u_n\}, \qquad I_{\text{struct}} = (U, F), \tag{1}$$

where U preserves semantic-unit identity without specifying contextual role assignments, whereas F specifies proposition-level frames and the bindings between semantic units and roles such as ARG0, ARG1, and ARG2 (Palmer et al., 2005). In the frozen target schema, U contains lexical semantic units and semantic operators, while F additionally includes directed inter-frame relations where annotated. Our primary analyses focus on predicate–argument organization; the complete frame, role, operator, and relation inventories are provided in Appendix B.

The representational boundary arises when relational organization is removed while semantic-unit identity is retained. Because labeled role assignments and frame relations can themselves be represented independently of enumeration order, this boundary is not a consequence of permutation

invariance or surface word order. This many-to-one representational boundary can be written as:

$$(U, F_1) \neq (U, F_2), \qquad \pi_U(U, F_1) = \pi_U(U, F_2) = U, \tag{2}$$

where π_U projects a structured representation onto its role-neutral semantic content. For example, the same $U = \{\textit{girl}, \textit{chase}, \textit{dog}\}$ can support either $F_A = \textit{chase}(\mathrm{ARG0} = \textit{girl}, \mathrm{ARG1} = \textit{dog})$ or $F_B = \textit{chase}(\mathrm{ARG0} = \textit{dog}, \mathrm{ARG1} = \textit{girl})$. Thus, F captures the proposition-level organization missing from U without requiring an ordered sentence representation, a full syntactic parse, or a comprehensive meaning representation.

Before testing whether such structure is recoverable from EEG, we establish its functional relevance by holding the gold semantic content U^* fixed and manipulating only the relational payload:

$$F_c \in \{\varnothing, F_{\mathrm{Gold}}, F_{\mathrm{Swap}}, F_{\mathrm{Shuffle}}\}, \qquad \hat{s}_c = G(U^*, F_c). \tag{3}$$

All conditions use the same reconstruction function G and the same normalized lexical units and semantic operators. F_{Gold} supplies the correct predicate–argument organization; F_{Swap} exchanges eligible argument assignments while preserving both semantic content and an explicit structural scaffold; the semantic-units-only condition sets $F_c = \varnothing$; and F_{Shuffle} substitutes a structurally compatible relation pattern from another record. Gold versus Swap is therefore the primary binding-specific counterfactual, while Gold versus semantic units only and Gold versus Shuffle test the broader contribution of explicit relational organization and whether the correct organization outperforms a structurally compatible mismatched alternative.

The comparison is designed to isolate relational effects from generation and evaluation artifacts. The reference sentence is withheld during generation, and candidates are evaluated by a frozen, condition-blind pairwise proposition-fidelity evaluator with order concealed. The primary endpoint is tie-adjusted Gold preference, with predicate-, ARG0-, ARG1-, ARG2-, and operator-level diagnostics reported separately. Eligibility rules, the formal preference definition, evaluator validation, and paired record-level bootstrap procedures are detailed in Appendix E. This text-side assay establishes whether correct binding contributes information beyond semantic-unit identity; it does not test whether that binding can be recovered from EEG or attributed to a specific neural occurrence.

2.2 Contextual Role Readout and Relation-Prior Audit

For RQ2, we test whether EEG-derived contextual representations retain role-related information beyond the semantic-unit identity preserved after retrieval, while separating such readout from information-source claims. All role and relation probes are fitted post hoc to frozen representations; the underlying EEG decoding models receive no explicit role or relation supervision during training. Given EEG input x, the frozen decoder produces contextual slot representations $H = \{h_i\}$, which are projected into the lexical space and matched against the semantic bank:

$$H = E_\theta(x), \qquad \hat{u}_i = \arg\max_{u \in \mathcal{B}} \mathrm{sim}(W_{\mathrm{lex}} h_i, e(u)), \qquad R = \{e(\hat{u}_i)\}. \tag{4}$$

Here, H denotes the pre-retrieval contextual representation and R the corresponding retrieved semantic-bank representation. ARG0 and ARG1 are the primary role endpoints. To quantify role-related readout beyond retrieved semantic-unit identity, we fit the same probe family to H and R and define the incremental role-readout contrast as:

$$\Delta^{(r)}_{\mathrm{role}} = \mathrm{AP}^{(r)}(H) - \mathrm{AP}^{(r)}(R), \qquad r \in \{\mathrm{ARG0}, \mathrm{ARG1}\}. \tag{5}$$

A positive $\Delta^{(r)}_{\mathrm{role}}$ indicates incremental role-related readout beyond semantic retrieval. Because H and R lie on opposite sides of the retrieval boundary within the same EEG-conditioned pipeline, this contrast is interpreted as incremental readout beyond retrieval rather than as a causal estimate of contextualization. This motivates a source audit of the larger relation-predictability effects.

We perform this audit by holding the eligible predicate–argument pairs, data partitions, probe family, and True-versus-Permuted construction fixed while varying only the feature source. Let $\mathcal{S} = \{\mathrm{Contextual}, \mathrm{Retrieved}, \mathrm{Gold\ lexical}, \mathrm{Fixed\ memory}\}$. For each $s \in \mathcal{S}$, we quantify relation predictability using the True-minus-Permuted AP difference:

$$\Delta^{(r)}_{\mathrm{rel},s} = \mathrm{AP}^{(r)}_{\mathrm{True},s} - \mathrm{AP}^{(r)}_{\mathrm{Permuted},s}, \qquad s \in \mathcal{S}. \tag{6}$$

In the Permuted condition, the joint argument-role labels are reassigned among eligible candidate rows within the same EEG sample, while predicate and candidate identities and their feature representations remain fixed. Contextual uses the current EEG-conditioned hidden states; Retrieved uses the corresponding semantic-bank representations; Gold lexical provides a non-neural lexical control from the gold semantic units; and Fixed memory preserves the frozen slot/query organization while removing current-record contextual information. Thus, a positive $\Delta^{(r)}_{\mathrm{rel},s}$ establishes relation predictability under source s, while a source-specific interpretation of the contextual effect requires comparison with these alternative sources rather than comparison with zero alone.

The same readout logic is evaluated across complementary representation and task conditions to test its robustness without changing its interpretation. We repeat the analysis with the primary Chisco Read model, a reference encoder, independently trained Chisco Imagine models, and ChineseEEG-2 Read representations. Dimension-matched lexical controls, exact pair construction, feature formation, permutation procedures, and representation-specific robustness analyses are provided in Appendix F. Accordingly, this stage distinguishes reproducible contextual role readout from relation predictability that may also arise from lexical or decoder-related priors; neither result alone establishes dependence on the correctly matched neural occurrence.

2.3 Conditional EEG Capacity and Matched-Record Specificity

For RQ3, we distinguish predictive information added by EEG conditioning from information that depends on the correctly matched neural occurrence. Here, U supplies target-side semantic content features rather than EEG-predicted or retrieved semantic units; with dataset-specific eligibility fixed, we compare a content-only model, a matched-EEG model, and a correspondence-breaking EEG control for target-aligned neural record x:

$$\begin{aligned} M_0 = P(F \mid U), \quad M_1 = P(F \mid U, x), \quad M_2 = P(F \mid U, \pi(x)), \\ \Delta^{(q)}_{\mathrm{cap}} = \mathrm{AP}^{(q)}(M_1) - \mathrm{AP}^{(q)}(M_0), \quad \Delta^{(q)}_{\mathrm{spec}} = \mathrm{AP}^{(q)}(M_1) - \mathrm{AP}^{(q)}(M_2). \end{aligned} \tag{7}$$

Here, $\pi(x)$ replaces the matched neural record while preserving the target, structural labels, candidate universe, and prespecified evaluation constraints. Dataset-specific correspondence-breaking procedures are detailed in Appendices D and G.

The two contrasts test different evidential requirements. $\Delta^{(q)}_{\mathrm{cap}}$ measures conditional EEG predictive capacity beyond role-neutral semantic content, whereas $\Delta^{(q)}_{\mathrm{spec}}$ asks whether that additional predictive information depends on the correctly corresponding neural record. The endpoint q follows the prespecified dataset-specific evaluation protocol: macro role AP for ChineseEEG-2 and ARG0/ARG1-specific AP for Chisco. A positive capacity contrast therefore does not imply positive matched-record specificity. We use matched-record specificity for the $M_1 - M_2$ endpoint and reserve occurrence-specific attribution for the broader cumulative claim.

Supporting analyses test whether weaker record-level evidence survives stricter data separation or can be reproduced by non-neural nuisance information. For the secondary record-ranking assay, we evaluate the same participants on non-overlapping neural records and compare matched EEG with alternatives based on stimulus source, recording duration, and their combination. For ChineseEEG-2 Read, prespecified subject-, role-, nuisance-stratum-, donor-, and representation-scale diagnostics further characterize heterogeneity in the primary specificity contrast without changing its definition, eligible records, or correspondence-breaking map. These auxiliary controls apply to their designated endpoints and are not used to explain the role- or relation-level readout effects in Section 2.2.

Occurrence-specific attribution is therefore treated as a cumulative evidential claim rather than the consequence of any single positive contrast. Contextual readout does not imply matched-record specificity, and conditional EEG capacity does not by itself establish dependence on the matched neural occurrence. Conversely, failure of a stronger specificity test limits the supported attribution claim under the tested controls rather than establishing the absence of relational information from EEG. Split- and seed-level estimates are aggregated within participant before group-level uncertainty is computed for subject-level role, capacity, and specificity contrasts, while the secondary record-ranking analyses retain their prespecified paired record-level bootstrap procedures. All primary contrasts are reported with 95% confidence intervals; complete model definitions, donor construction, aggregation rules, and decision criteria are provided in Appendices D and G.

3 EXPERIMENTS

We organize the experiments around the three-stage evidence framework defined in Section 2: functional relevance of explicit binding (RQ1), contextual role readout and relation-prior audit (RQ2), and conditional EEG capacity with matched-record specificity (RQ3).

3.1 EXPERIMENTAL SETUP AND EVIDENTIAL ROLES

Datasets and evidential roles. We use ChineseEEG-2 (Chen et al., 2026) and Chisco (Zhang et al., 2024a) as complementary experimental resources rather than parallel performance benchmarks. ChineseEEG-2 Read supports all three stages: its stimulus corpus provides the text-side structural intervention for RQ1, while the aligned EEG supports contextual readout in RQ2 and conditional capacity and specificity analyses in RQ3. ChineseEEG-2 Listen provides a cross-task evaluation of the RQ3 capacity–specificity distinction. Chisco Read provides an independent setting for both RQ2 and RQ3, whereas independently trained Chisco Imagine models test cross-task robustness of the RQ2 readout pattern. The complete task-to-evidence mapping is provided in Appendix A.

Splits and evaluation. All neural analyses use prespecified frozen record-level partitions, with models trained and evaluated within subject. ChineseEEG-2 Read uses 2,168 records shared across four subjects, with 1,734 for training and 434 for testing; ChineseEEG-2 Listen uses three frozen subject-native splits for each of eight subjects. Chisco Read and Imagine each include five subjects and three frozen within-subject splits, with Imagine models trained independently within task rather than transferred from Read. Primary results follow the contrasts defined in Section 2 and are reported with 95% confidence intervals. Dataset construction, preprocessing, and frozen partitions are detailed in Appendix A; structured targets in Appendix B; representation models in Appendix C; and controls and statistical protocols in Appendix D. RQ-specific protocols and extended analyses are reported in Appendix E for RQ1, Appendix F for RQ2, Appendix G for RQ3, and Appendix H for reconstruction diagnostics.

3.2 RQ1: FUNCTIONAL RELEVANCE OF EXPLICIT ROLE BINDING

To test the functional relevance of predicate–argument binding beyond role-neutral semantic-unit identity, we hold the semantic content U fixed and intervene only on its relational organization F (Table 1A; Figure 1). The primary Gold-versus-Swap contrast preserves both the semantic units and an explicit relational scaffold while changing only the argument assignments, providing the most targeted binding-specific counterfactual. The tie-adjusted Gold preference is 78.02% (95% CI, 73.49–82.33%; $N = 232$). The broader controls show the same direction: Gold attains a preference of 66.99% over Semantic units only ($N = 418$) and 77.43% over Shuffled relation ($N = 401$). These controls respectively test the broader contribution of explicit relational organization and the advantage of the correct relational organization over a structurally plausible mismatched alternative.

Together, these interventions establish that correct predicate–argument binding contributes functionally relevant proposition-level information beyond semantic-unit identity. The Gold-versus-Swap result is particularly informative because the semantic content and the presence of an explicit structural representation are held constant, leaving the argument assignment as the manipulated component. This evidence remains text-side: it establishes the functional relevance of relational organization, but does not show that the corresponding binding has been recovered from EEG or attributed to a particular neural occurrence. Full win/loss/tie counts, predicate- and role-specific judgments, structural-complexity analyses, and exact-content collision analyses are reported in Appendix E.

3.3 RQ2: CONTEXTUAL ROLE READOUT AND RELATION-PRIOR AUDIT

Following the readout framework in Section 2.2, we first test whether EEG-derived contextual representations support role-related readout beyond the semantic-unit identity preserved after retrieval (Table 1B; Figure 2B). On Chisco Read, contextual representations improve role AP over their corresponding retrieved representations by $+0.0330$ for ARG0 (95% CI $[+0.0268, +0.0388]$) and $+0.0226$ for ARG1 ($[+0.0169, +0.0262]$). ChineseEEG-2 Read shows the same pattern, with gains of $+0.0448$ for ARG0 ($[+0.0389, +0.0507]$) and $+0.0158$ for ARG1 ($[+0.0132, +0.0202]$). These contrasts are directionally consistent across all five Chisco subjects and all 15 subject–split cells, and

Table 1: **Functional relevance and structural readout.** (A) Functional relevance of explicit binding. (B) Contextual-minus-retrieved role readout. (C) Relation predictability across feature sources. Extended RQ1 analyses are in Appendix E; absolute AP, source controls, and robustness analyses are in Appendix F. N varies by manipulation eligibility.

(A) Functional relevance of explicit binding

Contrast	N	**Tie-adjusted Gold preference (%) [95% CI]**
Gold vs Semantic units only	418	66.99 [63.52, 70.45]
Gold vs Swapped binding	**232**	$\mathbf{78.02}$ [**73.49**, **82.33**]
Gold vs Shuffled relation	401	77.43 [74.31, 80.55]

(B) Contextual role-related readout

Dataset	**ARG0 ΔAP [95% CI]**	**ARG1 ΔAP [95% CI]**
Chisco Read	$+0.0330$ [+0.0268, +0.0388]	$+0.0226$ [+0.0169, +0.0262]
ChineseEEG-2 Read	$+0.0448$ [+0.0389, +0.0507]	$+0.0158$ [+0.0132, +0.0202]

(C) Relation predictability across feature sources

Dataset	**Feature source**	**ARG0 Δ_{rel} [95% CI]**	**ARG1 Δ_{rel} [95% CI]**
Chisco Read	Contextual	$+0.3732$ [+0.3643, +0.3826]	$+0.2668$ [+0.2628, +0.2711]
	Retrieved	$+0.3473$ [+0.3418, +0.3533]	$+0.2476$ [+0.2447, +0.2502]
	Gold lexical	$+0.5807$ [$+\mathbf{0.5725}$, $+\mathbf{0.5888}$]	$+0.4490$ [$+\mathbf{0.4406}$, $+\mathbf{0.4590}$]
	Fixed memory	$+0.3405$ [+0.3333, +0.3465]	$+0.2499$ [+0.2455, +0.2554]
ChineseEEG-2 Read	Contextual	$+0.1106$ [+0.0690, +0.1375]	$+0.1916$ [+0.1684, +0.2197]
	Retrieved	$+0.1299$ [+0.0792, +0.1599]	$+0.1611$ [+0.1333, +0.1889]
	Gold lexical	$+0.4602$ [$+\mathbf{0.4515}$, $+\mathbf{0.4665}$]	$+0.3511$ [$+\mathbf{0.3424}$, $+\mathbf{0.3574}$]
	Fixed memory	$+0.0892$ [+0.0776, +0.1043]	$+0.1566$ [+0.1510, +0.1628]

across all four ChineseEEG-2 subjects and all 12 subject–probe-seed cells. Thus, across both Read datasets, contextual representations retain a modest but reproducible increment in role-related readout beyond retrieved semantic-unit identity. Because the contextual and retrieved representations lie at different stages of the same EEG-conditioned pipeline, this comparison localizes the increment relative to retrieval but does not by itself identify its information source.

The True-versus-Permuted relation assay produces larger predictability effects, but the feature-source audit limits their interpretation (Table 1C; Figure 2C). Contextual Δ_{rel} is +0.3732/+0.2668 for ARG0/ARG1 on Chisco and +0.1106/+0.1916 on ChineseEEG-2. Retrieved and Fixed-memory representations approach these effects, while non-neural Gold lexical is stronger, reaching +0.5807/+0.4490 on Chisco and +0.4602/+0.3511 on ChineseEEG-2; for ChineseEEG-2 ARG0, Retrieved even exceeds Contextual. Accordingly, the large True-versus-Permuted contrast establishes relation predictability but not specificity to the contextual EEG representation. Absolute AP, source-specific contrasts, dimension-matched controls, and encoder- and task-level robustness analyses are reported in Appendix F.

3.4 RQ3: Conditional EEG Capacity and Matched-Record Specificity

RQ2 shows that contextual representations retain a modest role-related readout, but readout alone does not establish whether the predictive information depends on the correctly matched neural occurrence. We therefore distinguish *conditional EEG predictive capacity*, measured by $M_1 - M_0$, from the stricter requirement of *matched-record specificity*, measured by $M_1 - M_2$ (Table 2; Figure 2D–E). Against the strong content-only baseline M_0, ChineseEEG-2 Read shows a small positive capacity gain of $+0.00501$ macro-AP (95% CI $[+0.00280, +0.00726]$; 4/4 subjects positive), whereas matched-record specificity is negative at -0.00104 ($[-0.00196, -0.00019]$; 1/4 positive). ChineseEEG-2 Listen shows the same capacity–specificity separation under a different task condition: capacity remains positive at $+0.000205$ ($[+0.000027, +0.000457]$; 7/8 positive), whereas specificity is effectively null at $+0.000014$ ($[-0.000076, +0.000106]$; 5/8 positive). Chisco Read provides a more localized positive result. For ARG0, both capacity ($+0.002541$, $[+0.000806, +0.005438]$) and specificity ($+0.003325$, $[+0.001023, +0.005730]$) are positive in all

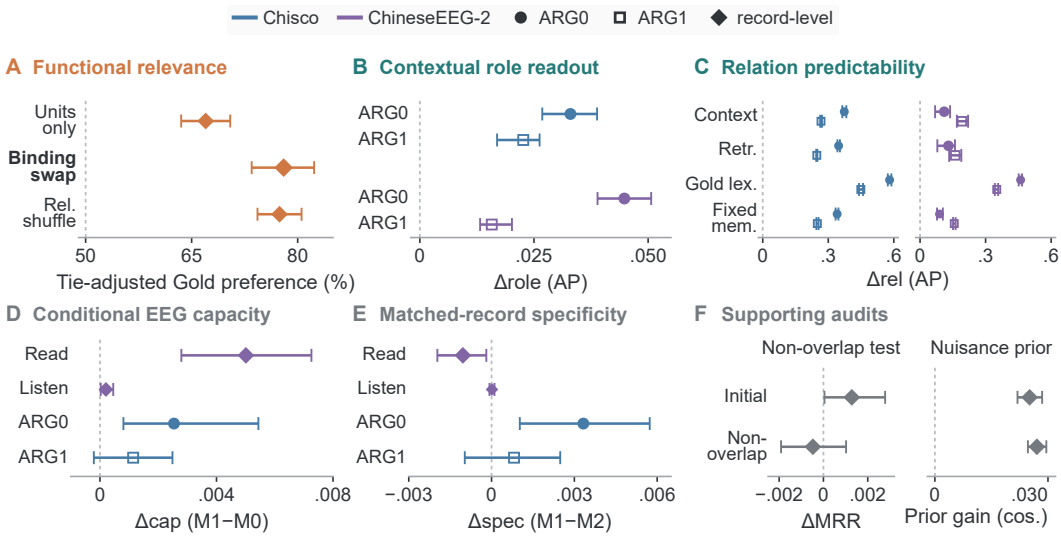

Figure 2: **Cumulative evidence for functional relevance, structural readout, and EEG specificity.** (A) Tie-adjusted preference for explicit role binding. (B) Contextual-minus-retrieved role readout. (C) Relation predictability across feature sources. (D) Conditional EEG predictive capacity. (E) Matched-record specificity. (F) Supporting non-overlap and nuisance-prior audits; the latter is restricted to the secondary record-ranking cosine endpoint. Context, Retr., Gold lex., and Fixed mem. denote contextual, retrieved, gold-lexical, and fixed-memory features. Colors denote datasets, marker shapes denote ARG0/ARG1 or record-level estimates, and error bars show 95% CIs. Dashed lines mark the 50% reference in A and zero otherwise.

Table 2: **Conditional EEG capacity and matched-record specificity.** (A) Absolute AP under content-only (M_0), matched-EEG (M_1), and correspondence-breaking EEG (M_2) conditions for ChineseEEG-2 Read. (B) Capacity and specificity contrasts across datasets and tasks. M. denotes macro AP over ARG0/ARG1. Endpoint-level results, specificity diagnostics, non-overlapping-record analyses, and nuisance controls are reported in Appendix G.

(A) Absolute AP under content-only and EEG-conditioned models

Model	**ARG0**	**ARG1**	**Macro ARG0/1**
Content-only M_0	0.4724	0.3188	0.3956
Matched EEG M_1	0.4789	0.3223	0.4006
Correspondence-breaking EEG M_2	0.4809	0.3224	0.4016

(B) Conditional EEG capacity and matched-record specificity across datasets and tasks

Setting	Δ_{cap} **[95% CI]**	Δ_{spec} **[95% CI]**	**Pos. Subj. (cap./spec.)**
ChineseEEG-2 Read (M.)	$+0.00501$ $[+0.00280, +0.00726]$	-0.00104 $[-0.00196, -0.00019]$	4/4 / 1/4
ChineseEEG-2 Listen (M.)	$+0.000205$ $[+0.000027, +0.000457]$	$+0.000014$ $[-0.000076, +0.000106]$	7/8 / 5/8
Chisco Read (ARG0)	$+0.002541$ $[+0.000806, +0.005438]$	$+0.003325$ $[+0.001023, +0.005730]$	5/5 / 5/5
Chisco Read (ARG1)	$+0.001140$ $[-0.000210, +0.002489]$	$+0.000814$ $[-0.000970, +0.002487]$	3/5 / 4/5

five subjects; for ARG1, however, both confidence intervals include zero. Across these primary assays, conditional EEG capacity is more consistently positive than matched-record specificity, whose evidence varies across datasets, tasks, and role endpoints.

Supporting diagnostics further constrain the specificity interpretation (Figure 2F). For ChineseEEG-2 Read, the negative aggregate specificity contrast is driven primarily by ARG0 and remains negative under leave-one-subject-out aggregation; donor-assignment and representation-scale diagnostics do not indicate an obvious correspondence-breaking artifact. We therefore interpret this result as a failure to establish matched-record specificity under the tested assay, rather than as evidence that correspondence-broken EEG is intrinsically more informative. The secondary record-ranking analysis provides no stronger support: its initially positive MRR contrast ($+0.001281$, 95% CI

$[+0.000046, +0.002786]$) changes sign on non-overlapping neural records from the same subjects (-0.000477, $[-0.001913, +0.001023]$), and a coarse source×duration prior exceeds matched EEG on the corresponding cosine endpoint. These secondary controls are specific to the record-ranking assay and do not explain the role- or relation-level AP results in RQ2. Detailed endpoint-, subject-, donor-, nuisance-stratum-, and non-overlapping-record analyses are reported in Appendix G. Taken together, ChineseEEG-2 shows small conditional EEG capacity without corresponding matched-record specificity, whereas Chisco provides a small ARG0-specific matched-record effect that remains unresolved for ARG1. Evidence for broader occurrence-specific attribution is therefore *weak and heterogeneous*, rather than uniformly absent or generally established.

3.5 Downstream Reconstruction Diagnostic

As a complementary diagnostic, we test whether downstream generation responds to predicate–argument changes. Under naturalistic payloads, 14 of 16 binding reversals leave the output unchanged, and only 1 of 16 is binding-faithful in both directions. A synthetic check shows that this is not an inability to use role structure: the frozen generator follows all 24 ARG0/ARG1 reversals, although their sentence embeddings remain highly similar (BGE cosine ≈ 0.985). Thus, reconstruction is not a reliable standalone indicator of relational recovery: naturalistic generation may ignore binding changes, while global similarity can remain high for correct reversals. We therefore use reconstruction only to diagnose generator and evaluation sensitivity, not as evidence of neural structural recovery; full results are reported in Appendix H.

4 Discussion

Our results identify a relational boundary of semantic-unit interfaces rather than a failure of semantic-unit decoding. Role-neutral semantic units remain an interpretable bridge between EEG and language, but unit identity alone does not determine proposition-level organization. RQ1 establishes the functional relevance of this missing organization: with semantic content fixed, correct predicate–argument binding improves proposition-level fidelity. RQ2 shows a modest but reproducible increment in contextual role-related readout beyond retrieved semantic-unit identity, whereas the larger True-versus-Permuted relation effect is substantially reproduced by retrieved, fixed-memory, and gold-lexical representations and is therefore better interpreted as relation predictability. RQ3 imposes a stricter source requirement: EEG conditioning provides small gains beyond role-neutral semantic content, while matched-record specificity remains weak and heterogeneous across datasets and endpoints. Because M_1 and M_2 are fitted separately, very small specificity contrasts may partly reflect optimization variability and are interpreted conservatively. Together, these results support a graded interpretation in which functional relevance and structural readout are better supported than conditional EEG contribution, matched-record specificity, and broader occurrence-specific attribution.

Downstream reconstruction should be interpreted within the same evidential hierarchy. The frozen generator responds to few naturalistic binding reversals but follows all ARG0/ARG1 reversals in controlled minimal pairs, showing that naturalistic insensitivity does not reflect an inability to use explicit role structure. Global sentence-embedding similarity nevertheless remains high even for correct reversals. Fluent or semantically similar reconstruction is therefore insufficient evidence of relational recovery from EEG; downstream generation is better treated as a diagnostic of structural use and metric sensitivity than as an independent test of neural structural recovery.

Two limitations define the next step. First, the structured semantic representation used here is deliberately minimal, capturing predicate–argument and frame-level organization rather than full syntax, surface order, or a comprehensive theory of meaning. Second, current naturalistic EEG corpora provide too few same-content/different-binding counterfactuals to support strict occurrence-specific identification. Future datasets should therefore counterbalance lexical content and role assignments while matching stimulus and recording nuisances, enabling relational hypotheses to be tested against genuinely diagnostic counterfactuals. The contribution of this study is not a claim of complete neural recovery of binding, but a framework for separating what is functionally relevant, what can be read out, what EEG adds beyond semantic content, and whether that contribution depends on the matched neural occurrence.

AI USE STATEMENT

Generative AI tools were used to assist the design and quality control of the structured semantic annotation protocol, provide feedback on research methodology and implementation, support downstream reconstruction and evaluation procedures, assist with result interpretation, and aid manuscript and figure preparation. The scientific uses of LLMs in structured-target construction and downstream reconstruction are described in Appendices B and H, respectively. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the final methods, claims, and reported results.

ETHICS STATEMENT

This study uses only publicly released EEG datasets and involves no new human-subject data collection. The original ChineseEEG-2 and Chisco studies report informed consent from participants and approval by their respective institutional ethics committees (Chen et al., 2026; Zhang et al., 2024a). All analyses in this work use the released data for research purposes, and no attempt was made to re-identify participants.

REPRODUCIBILITY STATEMENT

Detailed descriptions of the model architectures, training configurations, experimental protocols, controls, and evaluation procedures are provided in the main text and appendices. An anonymized repository for source code, experiment scripts, configuration files, and reproduction materials is provided at `https://anonymous.4open.science/r/RelStructEEG_ICLR2027-D5C0/`.

REFERENCES

Ina Bornkessel, Stefan Zysset, Angela D Friederici, D Yves Von Cramon, and Matthias Schlesewsky. Who did what to whom? the neural basis of argument hierarchies during language comprehension. *Neuroimage*, 26(1):221–233, 2005.

Sitong Chen, Beiqianyi Li, Cuilin He, Dongyang Li, Mingyang Wu, Xinke Shen, Song Wang, Xuetao Wei, Xindi Wang, Haiyan Wu, and Quanying Liu. An eeg dataset for multimodal semantic alignment and neural decoding during reading and listening. *Scientific Data*, 13 (1):148, 2026. doi: 10.1038/s41597-025-06466-8. URL `https://doi.org/10.1038/s41597-025-06466-8`.

João M Correia, Bernadette Jansma, Lars Hausfeld, Sanne Kikkert, and Milene Bonte. Eeg decoding of spoken words in bilingual listeners: from words to language invariant semantic-conceptual representations. *Frontiers in psychology*, 6:71, 2015.

Nina F Dronkers, Maria V Ivanova, and Juliana V Baldo. What do language disorders reveal about brain–language relationships? from classic models to network approaches. *Journal of the International Neuropsychological Society*, 23(9-10):741–754, 2017.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918, 2023.

Chris Foster, Chad C Williams, Olave E Krigolson, and Alona Fyshe. Using eeg to decode semantics during an artificial language learning task. *Brain and Behavior*, 11(8):e2234, 2021.

Steven M Frankland and Joshua D Greene. An architecture for encoding sentence meaning in left mid-superior temporal cortex. *Proceedings of the National Academy of Sciences*, 112(37):11732–11737, 2015.

Steven M Frankland and Joshua D Greene. Two ways to build a thought: distinct forms of compositional semantic representation across brain regions. *Cerebral Cortex*, 30(6):3838–3855, 2020.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011.

Laura Giglio, Peter Hagoort, and Markus Ostarek. Neural encoding of semantic structures during sentence production. *Cerebral Cortex*, 34(12):bhae482, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Carrie N Jackson, Heidi Lorimor, and Janet G van Hell. The erp correlates of thematic role assignment for passive versus active sentences. *Journal of Neurolinguistics*, 54:100886, 2020.

Nikolaus Kriegeskorte and Pamela K Douglas. Interpreting encoding and decoding models. *Current opinion in neurobiology*, 55:167–179, 2019.

Jiahe Li, Junru Chen, Fanqi Shen, Jialan Yang, Jada Li, Zhizhang Yuan, Baowen Cheng, Li Meng, and Yang Yang. Assembling the mind's mosaic: Towards eeg semantic intent decoding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8OgJ2uhiu8.

Xiaoqing Li, Haiyan Zhao, and Yong Lu. The relation between thematic role computing and semantic relatedness processing during on-line sentence comprehension. *PloS one*, 9(4):e95834, 2014.

Jacky Tai-Yu Lu, Jung Chiang, Chi-Sheng Chen, Anna Nai-Yun Tung, Hsiang Wei Hu, and Yuan Chiao Cheng. Eeg2text-cn: An exploratory study of open-vocabulary chinese text-eeg alignment via large language model and contrastive learning on chineseeeg. *arXiv preprint arXiv:2506.00854*, 2025.

Adam M Morgan, Orrin Devinsky, Werner K Doyle, Patricia Dugan, Daniel Friedman, and Adeen Flinker. Decoding words during sentence production with ecog reveals syntactic role encoding and structure-dependent temporal dynamics. *Communications Psychology*, 3(1):87, 2025.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

Sebastian Sauppe, Åshild Næss, Giovanni Roversi, Martin Meyer, Ina Bornkessel-Schlesewsky, and Balthasar Bickel. An agent-first preference in a patient-first language during sentence comprehension. *Cognitive Science*, 47(9):e13340, 2023.

Siavash Shams, Richard Antonello, Gavin Mischler, Stephan Bickel, Ashesh Mehta, and Nima Mesgarani. Neuro2semantic: A transfer learning framework for semantic reconstruction of continuous language from human intracranial eeg. *arXiv preprint arXiv:2506.00381*, 2025.

Alexander B Silva, Kaylo T Littlejohn, Jessie R Liu, David A Moses, and Edward F Chang. The speech neuroprosthesis. *Nature Reviews Neuroscience*, 25(7):473–492, 2024.

Jing Wang, Vladimir L Cherkassky, Ying Yang, Kai-min Kevin Chang, Robert Vargas, Nicholas Diana, and Marcel Adam Just. Identifying thematic roles from neural representations measured by functional magnetic resonance imaging. *Cognitive neuropsychology*, 33(3-4):257–264, 2016.

Wei Wang, Alan D Degenhart, Gustavo P Sudre, Dean A Pomerleau, and Elizabeth C Tyler-Kabara. Decoding semantic information from human electrocorticographic (ecog) signals. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6294–6298. IEEE, 2011.

Yipei Wang, Xiaoqian Wang, Ronnie Bring Wilbur, and Jeffrey Mark Siskind. Predicate-argument relations in the human brain. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WBkjxnYAgy.

Xiaoli Yang, Huiyuan Tian, Yurui Li, Jianyu Zhang, Shijian Li, and Gang Pan. Brain-CLIPLM: Semantic compression for EEG-to-text decoding, 2026. URL https://arxiv.org/abs/2604.16370.

Linmin Zhang and Liina Pylkkänen. Semantic composition of sentences word by word: Meg evidence for shared processing of conceptual and logical elements. *Neuropsychologia*, 119:392–404, 2018.

Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. Chisco: An eeg-based bci dataset for decoding of imagined speech. *Scientific Data*, 11(1): 1265, 2024a. doi: 10.1038/s41597-024-04114-1. URL https://doi.org/10.1038/s41597-024-04114-1.

Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. Chisco: Official dataset and code repository. GitHub repository, 2024b. URL https://github.com/zhangzihan-is-good/Chisco. Accessed 2026-09-02.

A Datasets, Preprocessing, Splits, and Corpus Statistics

A.1 Dataset and Task Overview

We analyze four task conditions from two public EEG resources: ChineseEEG-2 Read and Listen (Chen et al., 2026), and Chisco Read and Imagine (Zhang et al., 2024a;b). Throughout this appendix, a *record* denotes a unique linguistic record identity, whereas a *sample* or *occurrence* denotes an individual EEG trial associated with that record. This distinction is important for Chisco, where the same linguistic record may occur in multiple trials.

Table A1 summarizes the task-specific signal characteristics and analysis cohorts. ChineseEEG-2 provides four Read subjects and eight Listen subjects. Its structured semantic corpus contains 2,298 records; 2,168 of these have usable aligned EEG for all four Read subjects and form the common Read cohort used in the record-level analyses. The Listen recordings have subject-specific EEG coverage and are therefore analyzed using subject-native record inventories rather than a forced cross-subject intersection.

Chisco provides five subjects in both Read and Imagine conditions. The two tasks are stored and modeled separately. Across subjects, the Read recordings contain 5,011–6,289 unique records per subject, whereas Imagine contains 5,066–6,288.

Table A1: Dataset and task characteristics used in this work. Record counts denote unique linguistic record identities rather than EEG trial rows.

Dataset / task	Subjects	Channels	Native rate	Analysis record cohort
ChineseEEG-2 Read	4	128	250 Hz	2,168-record common cohort
ChineseEEG-2 Listen	8	128	1,000 Hz	465–536 records per subject
Chisco Read	5	124	500 Hz	5,011–6,289 records per subject
Chisco Imagine	5	124	500 Hz	5,066–6,288 records per subject

The four task conditions serve different evidential roles rather than constituting a symmetric benchmark suite. ChineseEEG-2 Read provides the stimulus-side intervention for RQ1 and, together with

Chisco Read, the primary EEG-based analyses for RQ2–RQ3; ChineseEEG-2 Listen and Chisco Imagine provide task-level robustness evaluations where applicable. Table A2 summarizes this assignment.

Table A2: **Evidential roles of datasets and task conditions.** The four task conditions are used as complementary experimental resources rather than parallel benchmarks.

Dataset / Task	RQ1: Functional relevance	RQ2: Readout / source audit	RQ3: Capacity / specificity
ChineseEEG-2 Read	Controlled text-side structural intervention	Contextual role readout + relation-prior audit	Conditional EEG capacity + matched-record specificity
ChineseEEG-2 Listen	—	—	Cross-task capacity/specificity
Chisco Read	—	Contextual role readout + relation-prior audit	Conditional EEG capacity + matched-record specificity
Chisco Imagine	—	Cross-task readout robustness	—

A.2 EEG Preprocessing and Record Alignment

ChineseEEG-2. We start from the released continuous BrainVision recordings and their associated event and linguistic-alignment metadata. The Read recordings contain 128 EEG channels sampled at 250 Hz, whereas the Listen recordings contain the same number of channels at 1,000 Hz. Linguistic records are linked to the corresponding temporal intervals in the continuous EEG through the released alignment information. Because the corpus is naturalistic, aligned record durations are variable rather than constrained to a single fixed window. Across usable aligned occurrences, the median durations are approximately 4.20 s for Read and 4.22 s for Listen.

The structured semantic release and EEG alignment are maintained separately: semantic targets are defined at the linguistic-record level, whereas EEG availability is determined independently for each subject and task. This separation prevents missing neural recordings from altering the semantic target inventory and enables the common-subject and subject-native cohorts described below. Acquisition-level and dataset-provided preprocessing details follow the original ChineseEEG-2 release (Chen et al., 2026); here we report the record alignment and analysis cohorts used in the present study.

Chisco. The Chisco analysis uses fixed-length `float32` EEG arrays generated from the task recordings. Read samples have shape 124×2500 and Imagine samples have shape 124×1650, corresponding to 5.0 s and 3.3 s windows at 500 Hz, respectively. The data-conversion procedure starts from 125-channel raw arrays, identifies the single marker channel, and removes it before storing the 124-channel EEG representation used by the models. The same channel contract is used for all five subjects and both tasks.

The resulting arrays are organized as task-specific runs and retain trial-to-record mappings, allowing repeated EEG occurrences of the same linguistic record to be tracked explicitly. Read and Imagine are treated as separate neural observations even when they correspond to the same linguistic stimulus, and task-specific coverage is handled by the frozen record manifests rather than assuming one-to-one availability across tasks.

A.3 Frozen Splits and Analysis Cohorts

All neural analyses use record-level partitioning. Samples sharing the same linguistic record identity are assigned to the same partition, so repeated EEG occurrences of a record cannot place that linguistic item in both training and test sets. This distinction is particularly important for Chisco, where the number of EEG trials is larger than the number of unique records.

ChineseEEG-2 Read. Among the 2,298 records in the structured corpus, 2,168 have usable aligned EEG in all four Read subjects. The primary analyses use the frozen RA-common seed0 partition, comprising 1,734 training and 434 test records, with the same linguistic partition applied to all four subjects. For the ChineseEEG-2 Read RQ2 analysis, the 12 repeated estimates arise from

Table A3: Corpus-level statistics of the frozen structured semantic targets. Lexical and operator counts denote occurrences in the semantic annotations, not EEG samples.

Quantity	ChineseEEG-2	Chisco
Structured records	2,298	6,567
Lexical occurrences	15,186	22,675
Operator occurrences	1,493	3,666
Single-frame records	819	5,000
Multi-frame records	1,479	1,567
Unit–role assignments	19,204	—
Frame relations	2,852	—
Role label inventory	—	14

four subjects crossed with three post-hoc probe seeds (42, 1, and 2); these seeds initialize probe fitting and do not redraw the record partition. Additional deterministic RA-common splits are retained as release assets but are not used by these paper-facing seed0 analyses.

ChineseEEG-2 Listen. EEG coverage differs across the eight Listen subjects, with 471, 484, 495, 498, 465, 480, 519, and 536 usable records, respectively. We therefore construct subject-specific record-level splits rather than restricting all subjects to the smaller common intersection. Across the three frozen splits, the corresponding train/test sizes are 376/95, 387/97, 396/99, 398/100, 372/93, 384/96, 415/104, and 428/108 records. Each subject is analyzed independently.

Chisco. Chisco is likewise partitioned by record identity within each subject and task. The five Read subjects contain 6,285, 6,289, 5,164, 6,251, and 5,011 unique records, respectively; their test partitions contain 1,257, 1,258, 1,033, 1,250, and 1,002 records. The corresponding Imagine inventories contain 6,140, 6,288, 5,066, 6,136, and 5,254 records, with 1,228, 1,258, 1,013, 1,227, and 1,051 test records. Three record-level splits are available for each subject–task combination. Across subjects, the Read EEG manifests cover all 6,567 records in the frozen structured target, whereas the Imagine manifests cover 6,566, reflecting a single task-specific missing record.

The number of EEG trial rows is not used interchangeably with these record counts. For example, under the first Read split, the five Chisco test sets contain 1,284, 1,281, 1,040, 1,268, and 1,040 EEG trials, but only 1,257, 1,258, 1,033, 1,250, and 1,002 unique linguistic records. All paper-facing dataset sizes therefore refer to record identities unless an analysis explicitly requires EEG occurrences as its sampling unit.

A.4 Corpus and Structural Statistics

The structured targets are defined once per linguistic record and are independent of the number of available EEG occurrences. Table A3 summarizes the resulting corpus-level structure. Detailed target definitions and annotation conventions are given in Appendix B.

ChineseEEG-2 contains 2,298 structured records with 15,186 lexical occurrences and 1,493 semantic-operator occurrences. The annotations contain 5,438 proposition-level frames, 19,204 unit–role assignments, and 2,852 directed frame relations. Of the 2,298 records, 819 contain a single frame and 1,479 contain multiple frames. The average record contains 6.61 lexical units, 0.65 semantic operators, 2.37 frames, and 8.36 unit–role assignments.

Chisco contains 6,567 structured records with 22,675 lexical occurrences and 3,666 operator occurrences. Its annotation follows the lighter frame-level contract used for this dataset: 5,000 records are single-frame and 1,567 are multi-frame, and lexical-role supervision is available for 14 contextual role labels. Inter-frame relation supervision is not included in the frozen Chisco target release. Consequently, analyses requiring explicit directed frame relations are based on the ChineseEEG-2 annotations, whereas Chisco provides an independent setting for role-sensitive analyses.

Within the 2,168-record ChineseEEG-2 Read cohort, 795 records are single-frame and 1,373 are multi-frame; 1,057 contain at least one semantic operator, and 1,336 contain at least one annotated

frame relation. The common EEG cohort therefore retains substantial variation in propositional complexity rather than being dominated by structurally simple utterances.

B Structured Semantic Targets and Annotation Quality

B.1 Core Notation and Structured Semantic Schema

The main text represents structured semantic intent as $I_{\text{struct}} = (U, F)$. Here we specify the annotation-level meaning of the symbols and role labels used throughout the paper. Table B1 provides a compact reference for the terminology that recurs in the figures, tables, and subsequent analyses.

Table B1: Core notation and semantic-role terminology used throughout the paper. Role labels are contextual assignments to lexical occurrences rather than intrinsic properties of lexical items.

Notation / term	Meaning in this work
I_{struct}	Structured semantic representation of a linguistic record, written as (U, F).
U	Role-neutral semantic content present in the record, comprising lexical units and semantic operators.
Lexical unit	A content-bearing semantic occurrence, such as an entity, predicate, state, attribute, location, or quantity.
Semantic operator	A proposition-level modifier such as negation, question, request, condition, obligation, intention, comparison, or uncertainty.
F	Relational organization of the semantic content, including proposition-level frames, contextual role assignments, and, where annotated, directed relations between frames.
Predicate	An event, state, or relation that licenses one or more contextual semantic arguments within a frame.
ARG0	Primary contextual participant, typically an initiator or experiencer.
ARG1	Contextual theme or affected participant.
ARG2	Secondary contextual argument, such as a goal, beneficiary, instrument, direction, or another relation-specific participant.
Frame	A proposition-level structure that groups predicate(s), participating semantic units, and their contextual roles.
Role assignment	The association between a lexical occurrence and its contextual role within a frame.
Frame relation	A directed semantic relation connecting two proposition-level frames.

The content component U is permutation-invariant and contains record-local semantic units. We distinguish lexical units from eight semantic operators: {`QUESTION`, `NEGATION`, `REQUEST`, `OBLIGATION`, `INTENTION`, `CONDITION`, `COMPARISON`, `UNCERTAINTY`}. Text spans and normalized glosses specify the identity of each occurrence; array position is not treated as semantic information. Importantly, lexical identity in U does not determine contextual role. The same lexical unit may therefore participate as ARG0, ARG1, or another role in different records or frames.

The relational component F organizes these units into proposition-level frames. The full schema distinguishes six frame types (*event*, *state*, *relation*, *topic*, *command*, and *question*) and assigns participating units to the role inventory {`ARG0`, `ARG1`, `ARG2`, `ATTRIBUTE`, `CAUSE`, `CONDITION`, `LOCATION`, `MANNER`, `POSSESSOR`, `PREDICATE`, `PURPOSE`, `QUANTITY`, `TIME`, `TOPIC`}. `PREDICATE` is included in this contextual role vocabulary, whereas semantic operators use the separate eight-class vocabulary above and their own query/output machinery; they are not role-probe outputs. A semantic occurrence may participate in more than one frame, allowing shared entities or embedded content to be represented without duplicating their semantic identity.

For records containing multiple propositions, the ChineseEEG-2 schema further uses nine directed inter-frame relations: {`COORDINATION`, `SEQUENCE`, `CAUSE`, `CONDITION`, `PURPOSE`, `CONTRAST`, `ELABORATION`, `CONTENT`, `TEMPORAL`}. Frame identities and explicit relation endpoints, rather than list order, define this graph. The representation is intentionally frame-lite: it preserves predicate–argument binding and proposition-level relations without attempting to reconstruct complete token order, a syntactic tree, or a full formal meaning representation. Operator

attachment is defined at the frame level; relative scope among multiple operators within the same frame is not represented.

B.2 Target Construction and Annotation Protocol

Structured semantic targets were constructed independently of EEG recordings and of all train–test splits. For ChineseEEG-2, the 2,298 linguistic records were first organized into a fixed record registry. The target sentence was the only source permitted to introduce semantic units or propositions. Preceding context could be consulted only when necessary to resolve an anaphor or an ambiguity already expressed in the target sentence; future context was not used. Each record was then decomposed into lexical units, semantic operators, contextual role assignments, proposition-level frames, and, where supported, directed relations between frames.

The annotation protocol was intentionally conservative. Lexical units and frames were required to be grounded in meaning expressed by the target sentence, although normalized glosses could be used to regularize surface forms. Context was not allowed to introduce an entity, event, or relation that was absent from the target. Likewise, omitted subjects or objects were not forced into the representation: when an argument was not explicitly recoverable from the target, the corresponding role could remain unfilled rather than being supplied by inference.

Frames were defined by semantic predication rather than punctuation. Independent events, states, conditions, causes, contents, purposes, and contrasts normally form separate frames, whereas attributes, manner expressions, result complements, and punctuation-delimited fragments without independent predication do not automatically introduce a new frame. Relation or state frames may therefore lack an explicit predicate when the linguistic realization omits a weak copular element.

Semantic operators were assigned according to their contribution to meaning rather than by surface triggers alone. For example, a negation marker is annotated as `NEGATION` only when it contributes semantic denial; `REQUEST` is distinguished from `OBLIGATION`; and an interrogative form is labeled `QUESTION` only when it denotes an unresolved question. Operators attach to the proposition represented by their frame and do not propagate across frames by default.

Within each frame, predicate–argument roles are assigned contextually. ARG0 generally denotes an initiator or experiencer, ARG1 a theme or affected participant, and ARG2 a goal, beneficiary, instrument, direction, or other secondary participant. These labels are properties of an occurrence in context rather than fixed attributes of lexical items. When multiple fillers share the same role, local role indices preserve their within-role ordering; they do not recover global token order.

The two corpora instantiate this protocol at different structural resolutions. ChineseEEG-2 retains the full schema used in this work, including lexical and operator occurrences, contextual roles, frames, and directed inter-frame relations. Chisco uses a reduced frame-lite target contract: it preserves lexical and operator content together with a 14-label contextual role vocabulary, but does not provide supervision for operator scope, frame type, or inter-frame relations. Consequently, the two target releases share the same content–role distinction but are not treated as interchangeable structural label spaces.

B.3 Quality Control and Frozen Target Representations

Quality control was performed at both the structural and semantic levels. Deterministic validation checked schema conformance, allowed label values, record identities, unit and frame references, and within-role indices. All 2,298 ChineseEEG-2 targets in the final release satisfy these mechanical constraints, with no dangling references, invalid enumerations, or role-index failures. Non-blocking ambiguities were retained explicitly rather than silently normalized; in particular, four records remain marked as known cases in which relative scope among multiple operators within a single frame is not represented.

Mechanical validity alone does not establish semantic fidelity. We therefore conducted a blinded diagnostic on 339 records, which identified 70 cases of decomposition-level information loss. Most were repairable within the existing schema (62/70), and consolidated review yielded 67 confirmed target replacements. This procedure served only to validate and refine the annotation release; reconstruction similarity from this diagnostic is not used as evidence of structural recovery.

LLMs assisted several stages of target construction and quality control, including decomposition, blinded reconstruction, semantic adjudication, and parts of the repair process. Accordingly, we do not describe the resulting targets as independently human-annotated gold labels. The final release is instead treated as a frozen, mechanically validated, and semantically reviewed annotation resource. A separate 48-sample evaluator-development packet that used ChatGPT-assisted review was excluded from the clean confirmatory analyses and does not contribute independent gold labels.

For ChineseEEG-2, the frozen text representation uses BGE-M3 embeddings with dimensionality 1,024. Sentence, lexical-occurrence, and operator-occurrence embeddings are stored separately and L2-normalized; no PCA or dimensional truncation is applied to these reference targets. The corresponding token bank is constructed at a cosine-similarity threshold of 0.98 and contains 7,936 lexical clusters together with the eight operator classes. Token-bank clustering affects retrieval identity but does not merge record-local occurrence nodes in the structured graph: contextual role and frame assignments therefore remain attached to individual occurrences. The frozen target capacities are 24 lexical and five operator slots, exceeding the observed maxima of 21 and five, respectively, so no ChineseEEG-2 target is truncated by the storage contract.

Chisco uses a separately frozen BGE-M3 representation and token bank. Its sentence, lexical-occurrence, and operator-occurrence embeddings are also 1,024-dimensional, while lexical retrieval uses a dataset-specific clustering threshold of 0.96. The resulting token bank contains 9,299 lexical clusters. The structured target stores up to 10 lexical and four operator units per record; the observed maxima are nine and four, respectively, so no Chisco target is truncated. As described in Appendix B.2, this release retains the frame-lite role supervision required by the Chisco analyses rather than introducing unsupported frame-type or inter-frame relation labels.

B.4 Annotation Examples and Boundary Cases

Table B2 illustrates how the structured schema separates role-neutral semantic content from relational organization. The examples are schematic and independent of EEG signals, model predictions, and experimental outcomes.

Table B2: Constructed examples of the structured semantic schema. The first two rows preserve the same semantic content U while reversing predicate–argument binding; the third illustrates explicit organization across multiple propositions.

Example	Role-neutral content U	Relational organization F
Binding A *The girl chases the dog.*	$\{girl, chase, dog\}$	*chase*(ARG0 = *girl*, ARG1 = *dog*)
Binding B *The dog chases the girl.*	$\{girl, chase, dog\}$	*chase*(ARG0 = *dog*, ARG1 = *girl*)
Multi-frame *If the girl arrives, the dog leaves.*	$\{girl, arrive,$ $dog, leave,$ $\texttt{CONDITION}\}$	$f_1 = arrive(\text{ARG0} = girl)$ $f_2 = leave(\text{ARG0} = dog)$ $\texttt{CONDITION}(f_1 \rightarrow f_2)$

The first two rows form a controlled minimal pair. Their semantic content U is identical, whereas the ARG0/ARG1 assignments are reversed. Thus, preserving which semantic units are present does not by itself preserve who participates in which predicate–argument role. The multi-frame example shows that relations between propositions are represented through explicit frame identities and directed endpoints rather than through clause order.

To complement these schematic examples, Table B3 shows compact serializations of representative frozen targets from the two corpora. The displayed records were selected from the training partitions for presentational clarity rather than scientific importance. We retain the semantic fields needed to illustrate each released annotation contract while omitting empty optional fields and implementation-only indexing or embedding metadata. The ChineseEEG-2 example illustrates the richer frame-level

Table B3: **Representative frozen structured-target serializations.** (A) A ChineseEEG-2 training record with two event frames, a shared NEGATION operator, contextual predicate–argument roles, and a directed COORDINATION relation. (B) A Chisco training record showing the frame-lite target contract with contextual role labels but without supervised frame type or inter-frame relations. Empty optional fields, redundant fields, and implementation-only metadata are omitted; roles in (A) are grouped under their corresponding frames for readability. No displayed semantic label is altered.

(A) ChineseEEG-2: frame-level structured target

```
{
  "record_id": "CE2_LP_REC_001266_F01",
  "sentence": "我既没看见人影，也没听见人声。",
  "lexical_occurrences": [
    {"unit_id":"u0", "text":"我",   "unit_type":"pronoun"},
    {"unit_id":"u1", "text":"看见", "unit_type":"predicate"},
    {"unit_id":"u2", "text":"人影", "unit_type":"entity"},
    {"unit_id":"u3", "text":"听见", "unit_type":"predicate"},
    {"unit_id":"u4", "text":"人声", "unit_type":"entity"}
  ],
  "operator_occurrences": [
    {"unit_id":"u5", "operator":"NEGATION", "trigger":"没"}
  ],
  "semantic_frames": [
    {"frame_id":0, "frame_type":"event",
     "roles":{"PREDICATE":"u1", "ARG0":"u0",
              "ARG1":"u2", "OPERATOR":"u5"}},
    {"frame_id":1, "frame_type":"event",
     "roles":{"PREDICATE":"u3", "ARG0":"u0",
              "ARG1":"u4", "OPERATOR":"u5"}}
  ],
  "frame_relations": [
    {"from_frame_id":0, "relation":"COORDINATION", "to_frame_id":1}
  ]
}
```

(B) Chisco: frame-lite contextual-role target

```
{
  "record_id": "chiso_000707",
  "sentence": "在美国，握手代表信任和尊重",
  "frame_count": 1,
  "single_frame": true,
  "role_supervision_available": true,
  "lexical_units": [
    {"unit_id":"u0", "text":"美国", "unit_type":"location",  "role_labels":["LOCATION"]},
    {"unit_id":"u1", "text":"握手", "unit_type":"event",     "role_labels":["ARG0"]},
    {"unit_id":"u2", "text":"代表", "unit_type":"predicate", "role_labels":["PREDICATE"]},
    {"unit_id":"u3", "text":"信任", "unit_type":"state",     "role_labels":["ARG1"]},
    {"unit_id":"u4", "text":"尊重", "unit_type":"state",     "role_labels":["ARG1"]}
  ]
}
```

schema, whereas the Chisco example shows the frame-lite contextual-role target used in the corresponding analyses.

The schema deliberately stops short of a full formal meaning representation. In particular, semantic operators are attached at the proposition or frame level, but relative scope among multiple operators attached to the same frame is not represented. Likewise, F preserves the relational distinctions required by the analyses in this work rather than complete surface word order or a full syntactic parse.

C Models and Representations

C.1 Neural Representation Encoders

The structural analyses operate on neural representations produced by an EEG encoder followed by a typed-query decoder. For the primary Chisco Read pipeline, the preprocessed EEG sequence is

first transformed by a temporal convolutional front end and convolutional tokenizer into a sequence of neural tokens. Two Transformer encoder layers integrate temporal context, after which the token sequence is projected to the common model width $d = 256$. The resulting sequence is retained as neural memory rather than collapsed into a single pooled representation, allowing the subsequent query decoder to condition each semantic-unit representation on the full EEG context.

The query decoder contains one intent query, ten lexical queries, and four operator queries, with query-group embeddings distinguishing their semantic functions. Lexical and operator queries act as unordered prediction slots. Each lexical slot predicts its existence and is projected from the 256-dimensional decoder state into the frozen 1024-dimensional semantic embedding space. Operator slots use the same semantic projection together with an operator-type prediction, while the intent query provides a separate global semantic representation. Lexical and operator targets are matched independently during representation learning. Most importantly for the analyses below, explicit semantic-role and frame-relation supervision is disabled when these representation models are trained. The contextual representations examined in Sections C.2 and C.3 are therefore learned without direct ARG0, ARG1, ARG2, or relation labels.

To assess whether the readout pattern depends on a particular EEG encoder, we additionally use an architecture-controlled reference pipeline for the Chisco analyses. This reference replaces the primary neural front end with a convolutional temporal encoder consisting of a patch projection followed by residual multi-kernel temporal blocks, while preserving the 256-dimensional neural-memory interface. The typed-query decoder, prediction heads, structured targets, matching objective, frozen data partitions, and downstream readout procedures are otherwise held fixed. The reference pipeline performs its own lexical projection and nearest-centroid retrieval; retrieved identities from the primary encoder are never reused. It therefore provides an encoder-level robustness control rather than an exact reproduction of a previously published model.

For the ChineseEEG-2 conditional analyses, the same lightweight neural encoding family is adapted to the dataset-specific channel configuration. Padding is propagated through tokenization, and the valid 256-dimensional neural-memory tokens are mask-averaged to obtain the record-level EEG representation used by the conditional residual model in Section C.4. Dataset-specific preprocessing and record construction are described in Appendix A; additional implementation details are provided with the released code.

C.2 Contextual and Retrieved Semantic Representations

For each valid lexical occurrence, we distinguish the EEG-conditioned representation available before lexical retrieval from the semantic representation selected after retrieval. Let $h_i \in \mathbb{R}^{256}$ denote the matched lexical decoder state for occurrence i. We refer to h_i as the *contextual representation*. The frozen lexical projection maps this state into the semantic embedding space,

$$z_i = W_{\text{lex}} h_i, \qquad \hat{u}_i = \arg\max_{u \in \mathcal{B}} \text{sim}(z_i, e(u)), \qquad r_i = e(\hat{u}_i), \tag{8}$$

where $\mathcal{B}$ is the frozen lexical semantic bank and $e(\hat{u}_i) \in \mathbb{R}^{1024}$ is the representation stored for the top-ranked retrieved semantic unit. We refer to r_i as the *retrieved representation*. Importantly, r_i is the frozen semantic-bank representation of the selected unit, not the projected neural vector z_i itself. Construction of the frozen semantic banks is described in Appendix B.

For the paired readout analyses, predicted lexical slots are aligned with valid gold lexical occurrences using the frozen lexical matching procedure. This alignment identifies the corresponding occurrence to be evaluated; the retrieved representation itself remains determined by nearest-neighbor search in the frozen semantic bank. Each eligible occurrence therefore provides a paired representation (h_i, r_i) with the same occurrence identity and downstream structural label, while the two members of the pair represent different stages of the frozen decoding pipeline.

The Contextual-versus-Retrieved comparison asks whether structural information is more accessible from the contextual representation than from the retrieved representation. A positive readout difference therefore indicates information available to the post-hoc probe from h_i beyond what is equally recoverable from r_i. Because both representations arise within the same EEG-conditioned decoding pipeline, this contrast does not identify the source of that information and does not establish dependence on the correctly matched neural occurrence. Gold lexical representations, which later serve as non-neural lexical controls, are kept distinct from the model-selected Retrieved representations

Algorithm C1 Frozen contextual and retrieved representation extraction

Require: EEG record x; frozen model θ; bank $(\mathcal{B}, e)$; frozen alignment $\mathcal{A}_x$
Ensure: Paired $H = \{h_i\}$ and $R = \{r_i\}$ on aligned occurrences

$\mathcal{M}_x \leftarrow \mathrm{EEGEncoder}_\theta(x)$
$Q_x \leftarrow \mathrm{TypedDecoder}_\theta(\mathcal{M}_x)$
for $(i, j) \in \mathcal{A}_x$ **do**
  $h_i \leftarrow Q^{\mathrm{lex}}_{x,j}$
  $z_i \leftarrow W_{\mathrm{lex}} h_i / \|W_{\mathrm{lex}} h_i\|_2$
  $\hat{u}_i \leftarrow \arg\max_{u \in \mathcal{B}} \mathrm{sim}(z_i, e(u))$
  $r_i \leftarrow e(\hat{u}_i)$
end for
return $\{(h_i, r_i) : (i, j) \in \mathcal{A}_x\}$

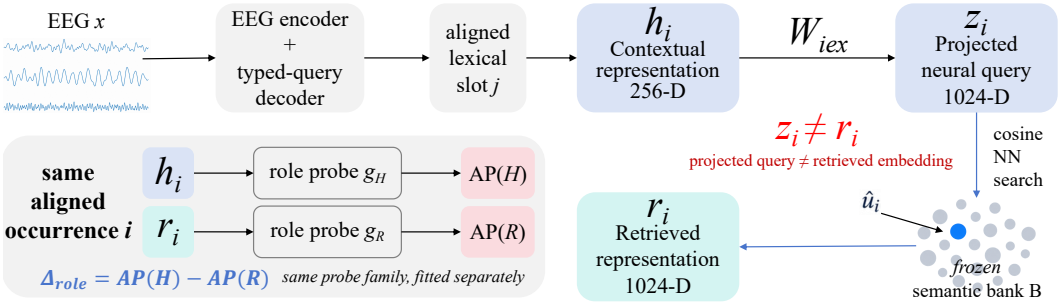

Figure C1: **Representation stages across the EEG-to-semantic retrieval boundary.** For an aligned lexical occurrence i, the frozen decoder provides the contextual representation h_i, which is projected by W_{lex} to the neural query z_i and used to retrieve semantic identity $\hat{u}_i$ from the frozen semantic bank $\mathcal{B}$. The Retrieved representation is $r_i = e(\hat{u}_i)$ and is therefore distinct from the projected neural query z_i. The lower panel illustrates the paired post-hoc role readout: Contextual and Retrieved representations from the same aligned occurrence are evaluated with the same probe family fitted separately, and their role-readout difference is summarized by $\Delta_{\mathrm{role}} = AP(H) - AP(R)$.

defined here. Figure C1 summarizes the representation stages across the retrieval boundary and the paired Contextual-versus-Retrieved role-readout comparison used below, while Algorithm C1 specifies the frozen extraction of H and R.

C.3 Structural Readout Models

We use lightweight post-hoc probes to quantify how much structural information is accessible from the frozen representations defined above. The representation encoders and retrieval modules remain fixed throughout these analyses, and probe gradients are never propagated back into the representation model. As described in Section C.1, the underlying representations are learned without explicit semantic-role or frame-relation supervision. The probes therefore measure downstream readout from an already learned representation rather than performance of an encoder trained directly for the structural labels.

Role readout. For an eligible lexical occurrence with representation $x_i \in \mathbb{R}^{d_x}$, the Chisco Frame-lite role probe predicts a 14-dimensional multilabel semantic-role vector over the frozen role vocabulary defined in Appendix B,

$$\hat{\mathbf{y}}_i^{\mathrm{role}} = g_{\mathrm{role}}(x_i), \tag{9}$$

where

$$g_{\mathrm{role}}(x) = \mathrm{Linear}_{256 \to 14}(\mathrm{Dropout}\,(\mathrm{GELU}\,(\mathrm{Linear}_{d_x \to 256}(\mathrm{LN}(x))))). \tag{10}$$

This vocabulary includes `PREDICATE` and excludes `OPERATOR`, while the primary readout analyses focus on `ARG0` and `ARG1`. This post-hoc role probe is distinct from the endpoint-specific ChineseEEG-2 conditional models described in Section C.4. Contextual and Retrieved representations are evaluated with the same probe family and the same occurrence-level targets so that their

Algorithm C2 Frozen post-hoc structural readout and feature-source audit

Require: Frozen model θ; aligned data $\mathcal{D}$ with H, R, structural labels, and fixed partitions
Ensure: $\Delta^{(r)}_{\text{role}}$ and $\{\Delta^{(r)}_{\text{rel},s}\}_{s\in\mathcal{S}}$

```
Freeze(θ)
for X ∈ {H, R} do
    g_X ← FitRoleProbe(X_train, y^role_train)
    A^(r)_X ← AP^(r)(g_X; X_eval, y^role_eval)
end for
Δ^(r)_role ← A^(r)_H − A^(r)_R, r ∈ {ARG0, ARG1}
S ← {Contextual, Retrieved, Gold lexical, Fixed memory}
P ← FrozenPairs(D)
(y_True, y_Permuted) ← RelationControls(P_fit)
for s ∈ S do
    X_s ← ConstructSource(s; P, θ)
    for c ∈ {True, Permuted} do
        g_{c,s} ← FitRelationProbe(X_{s,fit}, y_{c,fit})
        A^(r)_{c,s} ← AP^(r)(g_{c,s}; X_{s,eval}, y_eval)
    end for
    Δ^(r)_{rel,s} ← A^(r)_{True,s} − A^(r)_{Permuted,s}
end for
return Δ^(r)_role, {Δ^(r)_{rel,s}}_{s∈S}
```

difference reflects the accessibility of role-related information from the two representation stages rather than a change in classifier capacity.

Relation readout. Relation readout operates on a predicate–candidate pair. For predicate representation x_p and candidate representation x_a, their features are concatenated and passed through a probe of the same form:

$$\hat{\mathbf{y}}^{\text{rel}}_{p,a} = g_{\text{rel}}([x_p; x_a])\,, \tag{11}$$

with

$$g_{\text{rel}}([x_p; x_a]) = \text{Linear}_{256\to 3}(\text{Dropout}\,(\text{GELU}\,(\text{Linear}_{2d_x\to 256}(\text{LN}([x_p; x_a]))))\,. \tag{12}$$

The three outputs correspond to ARG0, ARG1, and ARG2 relation indicators; a pair carrying none of these relations is represented by the all-zero target rather than by an additional output class. The same probe architecture is used across representation sources and relation-control conditions, so changes in predictability are not confounded by changes in readout capacity.

Both readouts use a multilabel binary-cross-entropy objective, with class balancing determined from the corresponding training partition. Eligibility rules, construction of predicate–candidate pairs, relation permutations, frozen data partitions, and statistical aggregation are specified in Appendix D. These readouts quantify *structural accessibility*: successful prediction from a representation does not by itself identify whether the predictive information originates from the current EEG occurrence, lexical identity, or other regularities preserved by the decoding pipeline. Algorithm C2 summarizes the corresponding post-hoc readout and feature-source audit.

C.4 Conditional EEG Structural Model

To test whether EEG contributes structural predictive information beyond role-neutral semantic content, we use three conditional model classes:

$$M_0 = P(F \mid U), \qquad M_1 = P(F \mid U, x), \qquad M_2 = P(F \mid U, \pi(x)), \tag{13}$$

where U denotes the frozen target-side role-neutral semantic content used to construct the numerical content features; EEG-predicted or retrieved semantic units do not enter these features. Here, x is the EEG record aligned with the target occurrence and $\pi(x)$ breaks this correspondence. M_0 defines the content-only reference, M_1 adds matched EEG, and M_2 uses correspondence-broken EEG for

the same structural prediction problem. The dataset-specific construction of $\pi(\cdot)$ is described in Appendix D.2.

For ChineseEEG-2, M_0 uses the semantic-content feature construction described below. Read reuses the frozen content-only M_0 weights used for the ChineseEEG-2 Read analysis, whereas Listen refits M_0 within each subject–split cell and freezes it before conditional EEG fitting. For a predicate candidate i, the content feature is

$$\phi_i = [\, v_i; \bar{v}; v_i \odot \bar{v}; |v_i - \bar{v}|; n_{\text{lex}}; n_{\text{op}} \,], \tag{14}$$

where v_i is the candidate lexical representation, $\bar{v}$ is the record-level lexical mean, and n_{lex} and n_{op} denote the numbers of lexical and operator units in the record. For an ordered predicate–argument candidate (i, j), the corresponding feature is

$$\phi_{ij} = [\, v_i; v_j; v_i \odot v_j; |v_i - v_j|; \bar{v}; \bar{o}; n_{\text{lex}}; n_{\text{op}} \,], \tag{15}$$

where $\bar{o}$ is the record-level operator mean. Separate binary models are fitted for `PREDICATE`, `ARG0`, `ARG1`, and `ARG2`; ARG0/ARG1 macro AP is a prespecified summary endpoint rather than a separate model output. Sentence embeddings, raw text, token order or position, frame identifiers, and EEG are excluded from M_0. Thus M_0 is intended to absorb a comparatively strong semantic-content prior rather than serving as a simple prevalence baseline.

For M_1 and M_2, the M_0 parameters are frozen and an EEG-conditioned residual is added to the corresponding content logit. Let ϕ denote the relevant content feature and let $z \in \mathbb{R}^{256}$ denote the record-level neural representation obtained by mask-aware averaging of the valid neural-memory tokens. The conditional logit is

$$\ell(\phi, z) = \ell_{M_0}(\phi) + \frac{\langle W\phi, z \rangle}{\sqrt{256}}, \tag{16}$$

where $W\phi$ projects the content feature to the neural representation dimension. The residual branch is initialized at zero so that optimization starts from the content-only solution. M_1 uses the matched neural representation $z(x)$, whereas M_2 replaces only this neural source by $z(\pi(x))$; the semantic target, candidate identities, structural labels, and content features remain attached to the original target record. The two EEG-conditioned conditions use the same model family and optimization contract, but are fitted separately.

ChineseEEG-2 Read and Listen use their respective task-specific partitions, with the M_0 provenance described above. Chisco Read fits and freezes a content-only M_0 within each subject–split cell from target-side semantic features. Its eligible occurrence and predicate–candidate population, however, is inherited from the frozen EEG-to-target matching population used to define the Chisco Read evaluation set and is held fixed across M_0, M_1, and M_2. This branch evaluates `ARG0` and `ARG1` only and is not the Chisco 14-dimensional RQ2 role probe. Across assays, the matched-versus-correspondence-broken comparison preserves the target, candidate population, and structural labels while changing the EEG source. The three model classes therefore retain the same inferential meaning despite dataset-specific implementations of $\pi(\cdot)$. Algorithm C3 summarizes the fitting and evaluation sequence.

The resulting contrasts separate conditional predictive capacity from matched-record dependence.

D Controls and Statistical Protocols

D.1 Analysis Units, Eligibility, and Endpoints

The structural assays operate on different statistical units. Following the sample–record distinction defined in Appendix A, role readout is evaluated at the lexical-occurrence level, relation readout at the predicate–candidate level, and reconstruction or record-ranking analyses at the record level. Candidate rows derived from the same linguistic record are therefore not treated as independent observations.

For role readout, the analysis unit is an eligible matched lexical-node occurrence. For a role $q \in \{\text{ARG0}, \text{ARG1}\}$, an occurrence is positive when the corresponding node bears role q; all other eligible matched nodes are negative. Average precision is computed only when both positive

Algorithm C3 Conditional EEG capacity and matched-record specificity

Require: RQ3 cell $\mathcal{D}$; role-neutral content U; matched EEG x; map π; endpoints $\mathcal{Q}_d$
Ensure: $\{\Delta^{(q)}_{\mathrm{cap}}, \Delta^{(q)}_{\mathrm{spec}}\}_{q\in\mathcal{Q}_d}$
$\phi \leftarrow \phi_d(U)$
$M_0 \leftarrow \mathrm{ObtainContentModel}_d(\phi_{\mathrm{train}}, F_{\mathrm{train}})$
$\mathrm{Freeze}(M_0)$
$x_1 \leftarrow x$; $x_2 \leftarrow \pi(x)$
for $a \in \{1, 2\}$ **do**
    $z_a \leftarrow \mathrm{MaskedMean}(\mathcal{E}_{\theta_a}(x_a))$
    $W_a^{(q)} \leftarrow 0$ for $q \in \mathcal{Q}_d$
    $M_a \leftarrow \mathrm{FitConditional}(M_0, \phi, z_a; \theta_a, W_a)$
end for
for $q \in \mathcal{Q}_d$ **do**
    $A_k^{(q)} \leftarrow \mathrm{AP}^{(q)}(M_k; \mathcal{D}_{\mathrm{eval}})$, $k \in \{0, 1, 2\}$
    $\Delta^{(q)}_{\mathrm{cap}} \leftarrow A_1^{(q)} - A_0^{(q)}$
    $\Delta^{(q)}_{\mathrm{spec}} \leftarrow A_1^{(q)} - A_2^{(q)}$
end for
return $\{\Delta^{(q)}_{\mathrm{cap}}, \Delta^{(q)}_{\mathrm{spec}}\}_{q\in\mathcal{Q}_d}$

and negative classes are supported. In the ChineseEEG-2 Read contextual-role analysis, the frozen test cohort contains 434 linguistic records and 2,902 matched lexical-node occurrences per subject, illustrating why records and lexical occurrences must remain distinct.

For predicate–argument relation readout, an eligible record contributes one gold predicate p, which is paired with each eligible non-predicate matched node c_j,

$$(p, c_j), \qquad c_j \neq p.$$

Each predicate–candidate row carries binary targets for ARG0, ARG1, and, where available, ARG2. For a given endpoint, positive rows are those in which the candidate bears the corresponding argument role, while the remaining eligible rows are negatives. ARG0 and ARG1 are evaluated separately, and their arithmetic mean is used only in assays for which the macro endpoint is prespecified. In ChineseEEG-2 Read, the relation assay contains 132 eligible test records and 382 predicate–candidate rows per subject, reflecting a stricter eligibility criterion than the lexical-occurrence role assay above. These labels describe predicate–argument roles and should not be confused with inter-frame relation labels.

The record-level assays comprise functional reconstruction, evaluated on applicable paired judgments, and record-specificity or ranking analyses, evaluated on linguistic records. Candidate rows sharing one record are not independent replicates. Likewise, repeated probe seeds and split realizations provide within-subject replication and are not additional independent subjects. Aggregation order, clustered resampling, and confidence-interval construction are specified in Section D.4.

D.2 Correspondence-Breaking and Permutation Controls

The structural controls were designed to break one correspondence at a time while retaining the remaining components of the evaluated sample as far as permitted by the dataset. Specifically, we distinguish controls that disrupt (i) the association between lexical identity and semantic role, (ii) the assignment of predicate–argument relations to fixed candidate pairs, and (iii) the correspondence between the current EEG occurrence and its linguistic target. These manipulations address different possible sources of structural predictability and are therefore interpreted separately. Figure D1 summarizes the three control constructions and the correspondence altered by each one.

Within-lexeme role permutation. To test whether role readout can be explained by stable lexical–role regularities, role targets are reassigned only among occurrences belonging to the same lexical cluster. TRAIN and validation partitions are permuted independently, whereas TEST labels remain unchanged. For every lexical cluster containing at least two eligible occurrences, we solve a constrained zero-diagonal assignment over the complete multilabel role vectors. The procedure

therefore preserves lexical identity, the number of occurrences in each lexical cluster, and the within-cluster multiset of role vectors while breaking the original occurrence-to-role assignment. Because identical role vectors may occur more than once within a cluster, a zero-fixed-row assignment does not imply that the numerical role vector changes for every occurrence. This control removes the direct occurrence-level lexical–role correspondence, but does not remove higher-order lexical regularities such as predicate–argument co-occurrence priors.

True-versus-permuted relation assignment. For relation readout, the predicate and candidate identities and their feature representations are held fixed while the associated argument-role vectors are reassigned among candidate rows within the same EEG sample. The complete $\{\mathrm{ARG0}, \mathrm{ARG1}, \mathrm{ARG2}\}$ label vector is permuted jointly, preserving role-label marginals within each exchangeability group. TRAIN and validation targets are permuted independently, while both the True and Permuted models are evaluated against the same genuine TEST labels. The exchangeability unit is the EEG `sample_id`: this coincides with a record in the ChineseEEG-2 release used here, whereas Chisco may contain multiple EEG samples associated with the same linguistic record. A fixed outcome-blind seed and deterministic assignment are used across the frozen cells. Consequently, a positive True-versus-Permuted effect measures relation predictability relative to a marginally matched alternative; it does not by itself establish that the predictability is neural in origin.

Matched versus correspondence-broken EEG. For the conditional EEG analyses, M_2 keeps the target semantic content, structural labels, candidate identities, and content features fixed while replacing the current EEG with EEG from a donor occurrence. Relative to the matched condition M_1, the $M_1 - M_2$ contrast therefore measures dependence on the matched EEG–target correspondence under the preservation contract available for each dataset.

For ChineseEEG-2 Read, donor EEG is assigned within subject, official partition, and source strata using a bijective Hungarian assignment that minimizes absolute recording-duration mismatch and forbids self-assignment, yielding zero fixed points; inner validation reuses the TRAIN donor map. ChineseEEG-2 Listen uses the same minimum-duration-mismatch Hungarian construction within subject, split, official partition, and source, with deterministic tie handling and zero fixed points; inner validation again reuses the TRAIN donor map. Chisco Read instead uses a deterministic cyclic derangement within subject, split, and partition. The donor sample and donor linguistic record must both differ from the target, and source and duration are not matched. Thus, the three M_2 conditions share the same inferential role—breaking the current EEG–target correspondence while preserving the target structural sample—but do not impose an identical nuisance-preservation contract. In all cases, the donor map is fixed independently of the structural outcome being evaluated. The secondary record-ranking reproduction uses its own frozen correspondence-breaking map and is not interchangeable with structural M_2.

D.3 Alternative Information Sources and Non-Neural Controls

The correspondence-breaking controls above test whether a structural effect depends on a particular association between observations and targets. We additionally examine whether the same predictability can arise from information sources that do not use the current EEG record, including label prevalence, lexical identity, semantic-content representations, and coarse corpus metadata. These controls differ in the information they retain and are therefore interpreted as an information-source audit rather than as a single family of EEG-independent baselines.

Prevalence and lexical priors. For each binary endpoint q, we first report the empirical positive prevalence $\pi_q = N_q^+/N_q$, where N_q^+ and N_q denote the number of positive and total eligible rows, respectively. This quantity provides the expected reference level of average precision under uniformly random ranking for the corresponding endpoint. We further construct a train-only marginal lexical prior. For a lexical cluster ℓ, its smoothed endpoint probability is $\hat{p}(q \mid \ell) = (N_{\ell,q}^+ + 1)/(N_\ell + 2)$, where $N_{\ell,q}^+$ is the number of positive training rows for endpoint q and N_ℓ is the number of eligible training rows associated with that lexical identity. Lexical identities not observed during training fall back to the globally smoothed endpoint prevalence. Neither reference uses EEG features or test labels.

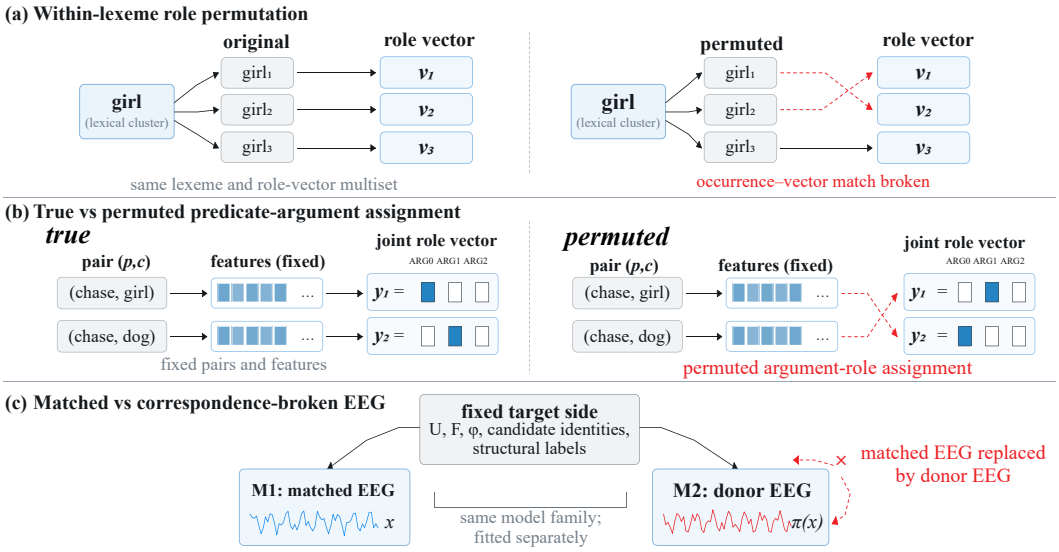

Figure D1: **Correspondence-breaking and permutation controls.** (a) Within-lexeme role permutation preserves lexical identity and the within-cluster multiset of multilabel role vectors while reassigning occurrence-to-role correspondences. (b) True-versus-Permuted predicate–argument assignment keeps predicate–candidate identities and feature representations fixed while jointly reassigning the $\{\mathrm{ARG0}, \mathrm{ARG1}, \mathrm{ARG2}\}$ target vectors among candidate rows within the same EEG sample. (c) Matched-versus-correspondence-broken EEG keeps the target-side semantic content, structural labels, candidate identities, and content features fixed while replacing the matched EEG x with donor EEG $\pi(x)$. M_1 and M_2 use the same model family but are fitted separately. Dataset-specific donor constraints are described in the text.

Content-only reference M_0. The conditional EEG analyses use M_0 as a stronger content-only reference. M_0 uses the lexical or semantic-content information permitted by the corresponding assay but excludes direct EEG input and the current EEG–record correspondence. The M_0 reference is obtained according to the assay-specific protocol and frozen before fitting the matched-EEG M_1 and correspondence-broken EEG M_2 arms. The M_0 provenance is assay-specific: ChineseEEG-2 Read reuses its frozen content-only M_0 weights, whereas ChineseEEG-2 Listen and Chisco Read fit and freeze their corresponding content-only M_0 models within each subject–split cell. Thus, M_0 denotes a common information boundary rather than a checkpoint reused from the RQ2 representation readout.

Alternative representation sources. To determine which information sources can support structural readout, the source audit evaluates the same fixed candidate rows and target labels using four alternative representations. The *Contextual representation* is the frozen EEG-derived representation used by the structural decoder. The *Retrieved representation* replaces each Contextual representation with the embedding of the semantic unit selected by the original retrieval pipeline; it therefore contains no direct EEG feature vector, although EEG still influences which semantic identity is selected. The *Gold lexical representation* instead uses the embedding associated with the ground-truth lexical identity and provides a purely target-side lexical reference.

The fourth source is the *Fixed-memory representation*, which removes the current-record EEG while retaining more of the original decoding organization. It replaces the record-specific encoder memory with a position-wise average constructed from TRAIN EEG and propagates this fixed memory through the frozen decoder, while preserving the original contextual slot/query assignment. It therefore serves as a control for dependence on the current EEG occurrence, not as a fully EEG-independent representation: it remains informed by training EEG and by the query organization of the original decoding pipeline.

Because the lexical representations are higher-dimensional than the contextual ones, we additionally repeat the source comparison after applying a fixed subsampled randomized Hadamard transform (SRHT). Retrieved and Gold lexical node representations are projected $1024 \to 256$, and predicate–

candidate pair representations $2048 \to 512$, matching the corresponding Contextual and Fixed-memory widths. The projection is fixed across TRAIN, validation, and TEST, contains no trainable parameters, and is followed by the same downstream probe. This control separates representation source from probe-input dimensionality.

Coarse nuisance priors for record ranking. Finally, for the secondary record-ranking analyses, we construct three train-only text-side priors to quantify how much retrieval performance can be explained by coarse corpus structure. We consider C_0 (global), C_1 (source), and C_2 (source$\times$duration). C_0 is the normalized global mean sentence embedding computed from the training set. C_1 replaces this with a source-specific training centroid. C_2 further conditions the source centroid on recording-duration quartiles. Duration bins are defined separately within each subject and source using the training-set 25%, 50%, and 75% quantiles of the enclosing recording duration. A source–duration cell is used only when it contains at least 20 training records; unseen cells are not assigned a back-off estimate. All centroids and bin boundaries are therefore determined without access to test outcomes. These nuisance priors are specific to the secondary record-ranking endpoint and are not treated as alternative explanations for the role- or predicate–argument readout assays.

D.4 Metrics, Aggregation, and Statistical Interpretation

Structural metrics. Binary role and relation endpoints are evaluated using average precision (AP), implemented with `sklearn.metrics.average_precision_score`. AP is computed separately for each endpoint, with the positive class defined by the corresponding structural label. Endpoints without support for both positive and negative classes are treated as unsupported rather than assigned a zero score. When a macro ARG0/ARG1 endpoint is prespecified, we first compute the two endpoint-specific values and then take their arithmetic mean, $\mathrm{AP}_{01} = (\mathrm{AP}_{\mathrm{ARG0}} + \mathrm{AP}_{\mathrm{ARG1}})/2$. Candidate counts are therefore not used to reweight the two roles.

The record-level assays use metrics appropriate to their corresponding questions. Functional reconstruction preference is $(N_{\mathrm{win}} + 0.5N_{\mathrm{tie}})/N_{\mathrm{applicable}}$, with non-applicable judgments excluded from the denominator. The initial record-ranking assay uses paired differences in cosine similarity between unit-normalized sentence representations, with mean reciprocal rank (MRR) reported as a secondary endpoint. For the non-overlapping-record replication, duplicate-aware MRR assigns an item the best rank among candidates with an exactly duplicated sentence representation, while retaining the original record identity for strict ranking.

Contrast estimands. For a prespecified structural endpoint q, we define four complementary contrasts:

$$\begin{aligned}
\Delta^{(q)}_{\mathrm{role}} &= \mathrm{AP}(H_{\mathrm{ctx}}, q) - \mathrm{AP}(H_{\mathrm{ret}}, q),\\
\Delta^{(q)}_{\mathrm{rel}} &= \mathrm{AP}^{(q)}_{\mathrm{True}} - \mathrm{AP}^{(q)}_{\mathrm{Perm}},\\
\Delta^{(q)}_{\mathrm{cap}} &= \mathrm{AP}^{(q)}_{M_1} - \mathrm{AP}^{(q)}_{M_0},\\
\Delta^{(q)}_{\mathrm{spec}} &= \mathrm{AP}^{(q)}_{M_1} - \mathrm{AP}^{(q)}_{M_2}.
\end{aligned}$$

These estimands deliberately address different evidential questions. Δ_{role} measures additional structural readout available in the Contextual representation relative to the Retrieved representation; Δ_{rel} measures sensitivity to the genuine predicate–argument assignment relative to a permuted training target; Δ_{cap} measures the incremental predictive capacity of the EEG-conditioned model over the content-only reference; and Δ_{spec} measures dependence on the matched EEG–record correspondence under the dataset-specific M_2 construction. A positive value for one contrast is therefore not substituted for evidence required by another.

Aggregation and uncertainty. Replicated optimization or probe seeds and split realizations are treated as within-subject replication rather than additional independent subjects. Where multiple probe seeds are used, contrasts are first summarized within the corresponding subject or subject–split cell. Split-level estimates are then aggregated within subject, and subject-level summaries are combined with equal subject weighting unless a frozen assay protocol specifies otherwise. For example, the ChineseEEG-2 Read contextual-role analysis averages the three probe-seed contrasts within each of the four subjects before computing the group mean, whereas Chisco retains its three split cells within each subject before the subject-level summary.

When several lexical or predicate–candidate rows originate from the same linguistic record, uncertainty is estimated by resampling the record cluster rather than individual candidate rows. All rows associated with a sampled record inherit the same bootstrap multiplicity. Paired conditions are resampled synchronously whenever their comparison requires identical record support. In particular, the ChineseEEG-2 Read $M_0/M_1/M_2$ analysis uses the same record draws across all three arms and across subjects before computing endpoint AP, the ARG0/ARG1 macro, and the equal-subject mean. Subject-level resampling is used for assays whose frozen protocol treats subjects, rather than records, as the replication unit. Confidence intervals are two-sided bootstrap intervals under the corresponding frozen assay protocol; they are descriptive unless the assay explicitly uses the lower bound as part of a confirmatory criterion.

Statistical interpretation and reporting. We distinguish descriptive uncertainty, technical completion checks, and prespecified confirmatory criteria. The ChineseEEG-2 analyses linking contextual role readout and conditional EEG capacity, together with the Chisco $M_0/M_1/M_2$ conditional analysis, use their bootstrap summaries descriptively; their technical completion status is not interpreted as an inferential "pass." The confirmatory Chisco role and relation analyses additionally require directional consistency across subjects, subject–split cells, and leave-one-subject-out summaries. For ChineseEEG-2 Read, the occurrence-specific conditional analysis requires both the capacity and matched-record specificity contrasts to satisfy their prespecified uncertainty criterion. For ChineseEEG-2 Listen, capacity and specificity are evaluated as separate contrasts rather than allowing one to compensate for the other. Likewise, secondary ranking metrics or diagnostic decompositions do not override a failed primary endpoint.

E RQ1: Functional Relevance of Explicit Role Binding

E.1 Functional Intervention and Evaluation Protocol

RQ1 is evaluated through a text-side controlled intervention on the ChineseEEG-2 Read stimulus corpus, independently of EEG signals. As defined in Section 2.1, all reconstruction conditions receive the same frozen gold lexical units and semantic operators, while only the relational payload is manipulated. The four conditions provide semantic units without explicit structure, the gold predicate–argument structure, swapped argument bindings, or a type- and cardinality-compatible shuffled relation structure. Gold structure and semantic units only are available for all 434 test records; valid swapped-binding and shuffled-relation interventions are available for 248 and 417 records, respectively. The reduced coverage of the latter conditions reflects manipulation eligibility rather than post hoc selection.

All candidates are generated with the same reconstruction procedure and compared using an arm-blind pairwise proposition-fidelity evaluator, with condition identity and candidate order hidden during evaluation. In addition to overall proposition fidelity, the evaluator records separate judgments for the predicate, `ARG0`, `ARG1`, `ARG2`, and operator dimensions. The 48 examples used during evaluator development and validation are excluded from the confirmatory analysis. After this exclusion, the clean populations contain 418 comparisons for Gold versus Semantic units only, 232 for Gold versus Swapped binding, and 401 for Gold versus Shuffled relation, for 1,051 eligible contrast-level judgments across the three comparisons in total. Tie-adjusted preference and its confidence interval follow the statistical definitions in Appendix D; confidence intervals use 2,000 paired record-level bootstrap draws with frozen seed 20260816.

E.2 Functional Preference and Binding-Dimension Results

The most targeted RQ1 comparison is Gold versus Swapped binding, because semantic content and the presence of an explicit relational representation are matched while the argument assignment is selectively altered. Gold structure was preferred in 0.7802 of the $N = 232$ eligible comparisons, with a 95% confidence interval of $[0.7349, 0.8233]$ (Table E1). The two broader controls yielded consistent functional advantages: Gold was preferred to Semantic units only at 0.6699 ($N = 418$, $[0.6352, 0.7045]$) and to Shuffled relation at 0.7743 ($N = 401$, $[0.7431, 0.8055]$). The corresponding win/loss/tie counts are reported in Table E1, including ties rather than discarding them from the comparison.

Dimension-level judgments further localize the Gold-versus-Swap effect. Preference remained close to the null value for predicate identity (0.5496), whereas it increased to 0.6875 for ARG0 and 0.7586 for ARG1. This separation is consistent with the intervention design: swapping changes which semantic units fill the argument roles while leaving the predicate content itself unchanged. The broader Shuffled-relation control showed the same direction of separation, with preferences of 0.7295 for ARG0, 0.7683 for ARG1, and 0.6400 for the predicate. Because shuffling perturbs relational organization more broadly than a within-record argument swap, these dimension-level shuffle results are supporting controls rather than the primary binding-specific evidence.

Results for ARG2 and operators are included for completeness. Their applicable populations differ from those of the core roles—only 55 swapped-binding comparisons contain an applicable ARG2 judgment—and operator content itself is held fixed across conditions. Accordingly, these dimensions are not interpreted as independent evidence for the binding effect. All dimension scores are correlated judgments of the same reconstructed candidate pairs and are therefore not treated as independent replications.

Table E1: **Functional preference and binding-dimension results.** The Gold-versus-Swapped-binding comparison is the primary binding-specific contrast; Semantic-units-only and Shuffled-relation comparisons provide broader functional controls.

Panel A: Overall functional contrasts

Functional contrast	N	**Gold W/L/T**	**Preference**	**95% CI**
Gold vs Swapped binding	232	158 / 28 / 46	0.7802	$[0.7349, 0.8233]$
Gold vs Semantic units only	418	206 / 64 / 148	0.6699	$[0.6352, 0.7045]$
Gold vs Shuffled relation	401	251 / 31 / 119	0.7743	$[0.7431, 0.8055]$

Panel B: Binding-dimension analysis

Dimension	**Swap** N	**Swap preference**	**Shuffle** N	**Shuffle preference**
Predicate	232	0.5496	400	0.6400
ARG0	232	0.6875	353	0.7295
ARG1	232	0.7586	328	0.7683
ARG2	55	0.6455	72	0.7917
Operator	178	0.5815	315	0.6603

E.3 Structural Complexity Analysis

Structural complexity strata were defined from the training partition before examining the functional-comparison outcomes. Frame complexity separates single-frame records ($n_{\text{frames}} = 1$) from multi-frame records ($n_{\text{frames}} \geq 2$). Binding load is the number of explicit predicate–role–argument edges. Its training-partition median was 4 (mean $= 5.363$, interquartile range $=$ 2–7), yielding the frozen split of low binding load (< 5) and high binding load (≥ 5). This produced 880/854 low/high-binding records and 646/1,088 single/multi-frame records in TRAIN. Applying the same definitions to TEST yielded 217/217 low/high-binding records and 149/285 single/multi-frame records; TEST was not used to select either threshold.

The primary complexity hypotheses concern the binding-specific Gold-versus-Swapped-binding contrast. Gold preference was similar across low- and high-binding-load records (0.7945 versus 0.7736), corresponding to a high-minus-low modulation of -0.0209 with 95% CI $[-0.1098, 0.0676]$. Likewise, preference was 0.7959 for single-frame records and 0.7760 for multi-frame records, giving a multi-minus-single modulation of -0.0200 with 95% CI $[-0.1132, 0.0728]$ (Table E2). Thus, the functional advantage of correct argument assignment is already present in structurally simpler records, with no evidence that it increases monotonically with either binding load or frame count.

A complementary pattern appears for the broader Gold-versus-Shuffled-relation control. Gold preference increased from 0.6289 in single-frame records to 0.8425 in multi-frame records, a raw modulation of $+0.2136$ (95% CI $[+0.1528, +0.2711]$). After adjustment for lexical-unit count and sentence length, the multi-frame coefficient remained positive ($+0.1011$, 95% CI $[+0.0194, +0.1809]$).

This secondary result indicates that multi-frame propositions are more sensitive to broadly incorrect relational organization, rather than that binding-specific utility follows a single monotonic complexity scale.

Other raw modulations involving Semantic-units-only, binding load, or operator presence were not stable after adjustment for lexical and surface complexity and are therefore not treated as independent complexity effects. Because role-binding load is itself strongly correlated with lexical and surface load, these analyses do not identify one source as explaining the other; rather, their contributions cannot be cleanly separated in the available corpus. No alternative threshold sweep or post hoc complexity partition was used.

Table E2: **Structural-complexity modulation of RQ1 functional contrasts.** Δ denotes complex minus simple preference.

Functional contrast	Complexity axis	Simple	Complex	Δ	95% CI
Gold vs Swapped binding	Binding load (< 5 / ≥ 5)	0.7945	0.7736	-0.0209	$[-0.1098, +0.0676]$
Gold vs Swapped binding	Frames (1 / ≥ 2)	0.7959	0.7760	-0.0200	$[-0.1132, +0.0728]$
Gold vs Shuffled relation	Frames (1 / ≥ 2)	0.6289	0.8425	$+0.2136$	$[+0.1528, +0.2711]$

E.4 Exact-Content Collision Analysis

As a complementary corpus-level check, we searched for records that share exactly the same unordered lexical and operator content but differ in their modeled predicate–argument structure. Such collisions provide a natural existence check of whether semantic-unit identity uniquely determines relational organization, without constructing the controlled counterfactual relations used in the intervention above.

Exact-content collisions were limited, and structure-varying collisions were particularly sparse (Table E3). In TRAIN, 29 collision groups comprising 70 records shared identical unordered content, but only one group, containing three records, exhibited different modeled relational structures. TEST contained four collision groups comprising eight records, none of which varied in modeled structure. Approximate lexical matches and embedding-nearest neighbors were not counted as collisions.

These observations show that identical unordered semantic content does not logically imply a unique predicate–argument organization. At the same time, the scarcity of naturally occurring structure-varying collisions prevents a reliable corpus-level estimate of the functional effect of binding. The controlled Gold-versus-Swapped-binding intervention therefore remains the primary RQ1 test, while the collision analysis serves only as supporting existence evidence.

Table E3: **Exact-content collisions in ChineseEEG-2 Read.**

Partition	Exact-content groups	Records	Structure-varying groups	Varying records
TRAIN	29	70	1	3
TEST	4	8	0	0

F Full Analyses for RQ2: Contextual Role Readout and Relation Predictability

F.1 Contextual Role Readout Across Datasets

We compare Contextual representations with their corresponding Retrieved representations under matched role-readout conditions. On Chisco Read, contextual representations improve role-readout AP from 0.404280 to 0.437266 for `ARG0`, corresponding to a paired gain of $+0.032986$ (95% CI $[+0.026815, +0.038816]$). For `ARG1`, AP increases from 0.366047 to 0.388692, a gain of

$+0.022645$ ($[+0.016921, +0.026224]$). Both contrasts are positive for all five subjects and all 15 subject–split cells.

ChineseEEG-2 Read shows the same contextual advantage for both core roles. The contextual-minus-retrieved gain is $+0.044791$ for ARG0 (95% CI $[+0.038913, +0.050668]$) and $+0.015752$ for ARG1 ($[+0.013203, +0.020174]$). Both effects are positive for all four subjects and all 12 subject–probe-seed cells. Across the two Read datasets, Contextual representations therefore retain a modest but reproducible role-related increment beyond Retrieved representations. This comparison localizes the increment to the contextual stage relative to retrieval, but does not by itself identify its neural source.

Table F1: **Contextual role readout across datasets and robustness conditions.** Absolute AP and paired contextual-minus-retrieved gains are reported for ARG0 and ARG1. Subject and cell counts indicate positive paired differences over the corresponding totals.

Panel A: Primary Read analyses

Setting	Role	Retrieved	Contextual	ΔAP [95% CI]	Subjects	Cells
Chisco Read	ARG0	0.404280	0.437266	+0.032986 [+0.026815, +0.038816]	5/5	15/15
	ARG1	0.366047	0.388692	+0.022645 [+0.016921, +0.026224]	5/5	15/15
ChineseEEG-2 Read	ARG0	0.335544	0.380335	+0.044791 [+0.038913, +0.050668]	4/4	12/12
	ARG1	0.272913	0.288665	+0.015752 [+0.013203, +0.020174]	4/4	12/12

Panel B: Robustness analyses

Setting	Role	Retrieved	Contextual	ΔAP [95% CI]	Subjects	Cells
Chisco Read (reference encoder)	ARG0	0.388566	0.413880	+0.025314 [+0.013515, +0.035723]	5/5	14/15
	ARG1	0.329801	0.345615	+0.015813 [+0.013502, +0.018125]	5/5	15/15
Chisco Imagine	ARG0	0.404193	0.420439	+0.016247 [+0.002147, +0.029347]	4/5	12/15
	ARG1	0.368820	0.395473	+0.026653 [+0.023869, +0.029686]	5/5	15/15

F.2 Relation Predictability Under True and Permuted Assignments

The True-versus-Permuted relation assay yields substantially larger effects than the contextual-minus-retrieved role contrast. On Chisco Read, contextual representations give $\Delta_{\text{rel}} = +0.373234$ for ARG0 (95% CI $[+0.364342, +0.382590]$) and $+0.266752$ for ARG1 ($[+0.262840, +0.271077]$), with positive direction in all five subjects and all 15 subject–split cells. ChineseEEG-2 Read shows the same qualitative pattern, with $\Delta_{\text{rel}} = +0.110568$ for ARG0 ($[+0.068961, +0.137480]$) and $+0.191604$ for ARG1 ($[+0.168434, +0.219734]$), again positive for all four subjects and all 12 probe cells. True predicate–argument assignments are therefore substantially more predictable than their frozen permutation controls under the contextual pair representations, motivating the source audit below.

F.3 Feature-Source Audit

To test whether the large relation-predictability effect is specific to contextual EEG-derived representations, we repeat the same assay while changing only the feature source under matched eligibility, splits, True-versus-Permuted assignments, and probe procedures. We compare the Contextual, Retrieved, Gold lexical, and Fixed-memory representations. The Gold lexical representation provides a non-neural lexical source, whereas the Fixed-memory representation preserves the frozen slot/query organization while removing current-record EEG.

The feature-source audit shows that the large relation effect is not specific to the Contextual representation. On Chisco, contextual Δ_{rel} exceeds the Retrieved and Fixed-memory representations only modestly, by $+0.025953/+0.019192$ and $+0.032695/+0.016865$ for ARG0/ARG1, respectively. In contrast, the Gold lexical representation is substantially stronger, giving contextual-minus-gold contrasts of -0.207436 for ARG0 and -0.182285 for ARG1. ChineseEEG-2 provides an even stricter case for ARG0: the retrieved representation exceeds the contextual representation by 0.019345 AP (95% CI for Contextual−Retrieved, $[-0.034607, -0.004083]$). For ARG1, Contextual retains smaller advantages over Retrieved ($+0.030505$, $[+0.017339, +0.047472]$) and Fixed-

memory (+0.035054, [+0.016270, +0.055372]), while remaining below the Gold lexical representation.

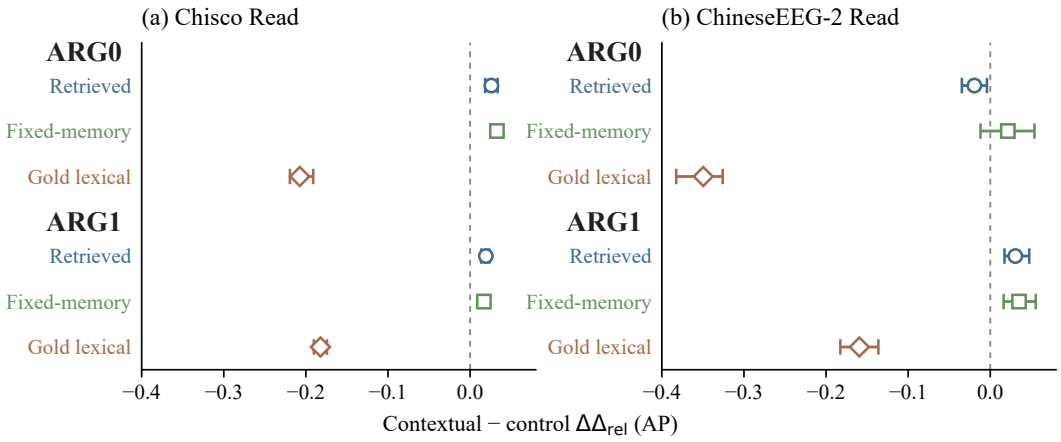

Figure F1: **Source audit of relation predictability.** Points show $\Delta\Delta_{\text{rel}} = \Delta_{\text{rel}}^{\text{Contextual}} - \Delta_{\text{rel}}^{\text{Control}}$ with 95% confidence intervals. Positive values favor the contextual representation, whereas negative values indicate stronger relation predictability in the control source. Both panels use the same horizontal scale.

Because the Retrieved and Gold lexical representations are 1024-dimensional in their native form, compared with 256 dimensions for the Contextual and Fixed-memory representations, we additionally repeat the lexical-source assays after a fixed 1024-to-256-dimensional SRHT projection. The resulting Δ_{rel} values remain close to their native counterparts (Table F2, Panel B), including strong gold-lexical effects on both datasets. The feature-source pattern is therefore not explained by the larger native dimensionality of the lexical representation sources.

Table F2: **Extended relation-predictability and dimension-matched controls.** Panel A reports absolute True and Permuted AP and their difference. Panel B compares native and SRHT-256 relation effects for the higher-dimensional Retrieved and Gold lexical representations.

Panel A: Absolute relation-predictability statistics

		ARG0			ARG1		
Dataset	**Feature source**	**True**	**Perm.**	**Δ_{rel} [95% CI]**	**True**	**Perm.**	**Δ_{rel} [95% CI]**
Chisco	Contextual	0.568051	0.194817	+0.373234 [+0.364342, +0.382590]	0.545013	0.278262	+0.266752 [+0.262840, +0.271077]
	Retrieved	0.551448	0.204167	+0.347281 [+0.341828, +0.353273]	0.528674	0.281114	+0.247560 [+0.244693, +0.250198]
	Gold lexical	0.777612	0.196942	+0.580670 [+0.572495, +0.588844]	0.728209	0.279172	+0.449036 [+0.440570, +0.459017]
	Fixed-memory	0.546623	0.206085	+0.340539 [+0.333314, +0.346518]	0.530569	0.280683	+0.249887 [+0.245536, +0.255389]
ChineseEEG-2	Contextual	0.437325	0.326757	+0.110568 [+0.068961, +0.137480]	0.387208	0.195604	+0.191604 [+0.168434, +0.219734]
	Retrieved	0.411936	0.282023	+0.129913 [+0.079152, +0.159871]	0.378685	0.217586	+0.161099 [+0.133300, +0.188898]
	Gold lexical	0.717999	0.257815	+0.460184 [+0.451510, +0.466488]	0.576354	0.225296	+0.351057 [+0.342381, +0.357443]
	Fixed-memory	0.388312	0.299152	+0.089159 [+0.077645, +0.104273]	0.378456	0.221905	+0.156550 [+0.150980, +0.162783]

Panel B: Dimension-matched sensitivity check

		ARG0 Δ_{rel}		ARG1 Δ_{rel}	
Dataset	**Feature source**	**Native**	**SRHT-256**	**Native**	**SRHT-256**
Chisco	Retrieved	+0.347281	+0.346621	+0.247560	+0.249899
	Gold lexical	+0.580670	+0.570509	+0.449036	+0.415612
ChineseEEG-2	Retrieved	+0.129913	+0.134027	+0.161099	+0.157390
	Gold lexical	+0.460184	+0.465917	+0.351057	+0.318852

Taken together, the much larger True-versus-Permuted relation effect is not specific to the Contextual representation and is substantially reproducible from alternative feature sources, whereas the smaller contextual role-readout increment remains reproducible across datasets.

F.4 Robustness Across Representation and Task Conditions

To test whether the observed readout pattern depends on the primary Chisco encoder or on the Read task, we repeat the analysis with a reference encoder and independently trained Chisco Imagine models. Under the reference encoder, the contextual-minus-retrieved role gain remains positive at $+0.025314$ for `ARG0` (95% CI $[+0.013515, +0.035723]$; 5/5 subjects, 14/15 cells) and $+0.015813$ for `ARG1` ($[+0.013502, +0.018125]$; 5/5 subjects, 15/15 cells). The corresponding True-versus-Permuted relation effects are $+0.350206$ and $+0.233250$, respectively, with positive direction in all subjects and cells.

The independently trained Chisco Imagine models show the same qualitative role-readout pattern. Contextual-minus-retrieved gains are $+0.016247$ for `ARG0` (95% CI $[+0.002147, +0.029347]$; 4/5 subjects, 12/15 cells) and $+0.026653$ for `ARG1` ($[+0.023869, +0.029686]$; 5/5 subjects, 15/15 cells). Their True-versus-Permuted relation effects are $+0.372185$ and $+0.269862$, again positive in all subjects and cells.

These analyses show that the qualitative readout pattern is not confined to the primary encoder or to the Read task. They do not alter the source interpretation established by the feature-source audit: robust relation predictability across model and task conditions is still compatible with substantial lexical and decoder/query contributions.

F.5 Consistency and Heterogeneity

The contextual-minus-retrieved role increment is small in magnitude but directionally consistent in the primary Read analyses. Both Chisco primary role contrasts are positive in all 15 subject–split cells, and both ChineseEEG-2 Read contrasts are positive in all 12 subject–probe-seed cells. The reference-encoder `ARG0` contrast is positive in 14/15 cells, while Chisco Imagine `ARG0` is positive in 12/15 cells and 4/5 subjects; the corresponding `ARG1` contrasts remain positive in all cells. The True-versus-Permuted relation effects show still stronger directional consistency across the primary, reference-encoder, and Imagine analyses.

This consistency characterizes the reproducibility of the readout assays rather than the specificity of their information source. Table F1 summarizes directional consistency across subjects and repeated-measure cells, while complete cell-level results are retained in the released analysis artifacts. Overall, RQ2 supports a modest and reproducible contextual role-related readout, whereas the larger relation-predictability effect is not specific to the Contextual representation and is substantially reproducible from alternative feature sources.

G RQ3: Conditional EEG Capacity and Matched-Record Specificity

G.1 ChineseEEG-2 Read: Conditional Capacity and Matched-Record Specificity

ChineseEEG-2 Read provides the primary conditional assay for separating EEG-conditioned predictive capacity from dependence on the correctly matched neural occurrence. Because the content-only model already provides a strong semantic baseline, EEG contribution is evaluated as a residual increment beyond U. Relative to M_0, matched EEG under M_1 yields a small but consistent increase in macro-AP, $\Delta_{\mathrm{cap}} = +0.00501$ (95% CI $[+0.00280, +0.00726]$), with the contrast positive in all four subjects. The stricter matched-record contrast, $M_1 - M_2$, has the opposite direction: $\Delta_{\mathrm{spec}} = -0.00104$ (95% CI $[-0.00196, -0.00019]$), with only one of four subjects showing a positive difference. Thus, additional predictive capacity and matched-record specificity separate under the same frozen assay: adding an EEG-conditioned branch improves prediction relative to content alone, but the improvement is not preferentially associated with the correctly matched EEG record.

The negative specificity contrast is not interpreted as evidence that correspondence-broken EEG is beneficial. Its implication is narrower: under the correspondence-breaking control, the additional predictive capacity cannot be attributed to the correct target–record correspondence. Section G.2 therefore examines the subject-, endpoint-, and nuisance-stratum structure of this contrast together with donor- and representation-scale diagnostics.

Table G1: **ChineseEEG-2 Read conditional capacity and matched-record specificity.** Panel A reports equal-subject mean AP for the content-only (M_0), matched-EEG (M_1), and correspondence-broken EEG (M_2) models, together with paired capacity and specificity differences. Panel B reports the frozen bootstrap intervals for the prespecified macro-AP (ARG0/ARG1) contrasts.

Panel A: Endpoint-level absolute AP and paired effects

Endpoint	M_0	M_1	M_2	Δ_{cap}	Δ_{spec}	**Pos. subj. (cap./spec.)**
ARG0	0.472408	0.478872	0.480884	+0.006464	−0.002012	4/4 / 1/4
ARG1	0.318761	0.322316	0.322383	+0.003556	−0.000067	4/4 / 2/4
ARG2	0.104869	0.106041	0.106428	+0.001173	−0.000386	4/4 / 0/4
macro-AP (ARG0/1)	0.395584	0.400594	0.401634	+0.005010	−0.001039	4/4 / 1/4

Panel B: Prespecified macro-AP (ARG0/ARG1) contrasts

Contrast	ΔAP	**95% CI**
Conditional capacity ($M_1 - M_0$)	+0.005010	[+0.002795, +0.007259]
Matched-record specificity ($M_1 - M_2$)	−0.001039	[−0.001964, −0.000190]

G.2 CHINESEEEG-2 READ: MATCHED-RECORD SPECIFICITY DIAGNOSTICS

The negative ChineseEEG-2 Read specificity contrast is concentrated primarily in ARG0 ($\Delta_{\mathrm{spec}} = -0.002012$), whereas ARG1 is close to zero (-0.000067) and ARG2 shows a smaller negative difference (-0.000386). At the macro-AP (ARG0/ARG1) level, three of four subjects show a negative contrast. Removing any single subject leaves the aggregate difference negative (range: -0.001421 to -0.000708), indicating that the reversal is not driven by a single participant.

The correspondence-breaking assignment itself shows no evident mapping failure: all 1,736 donors preserve source, with no self-donors, fixed points, invalid mappings, or repeated donor use. Duration mismatch is also small for most records (median 0.004 s; 75th percentile 0.008 s). The specificity contrast is nevertheless heterogeneous across source and duration strata, with stronger negative ARG0 differences in some strata and weak or positive differences in others. Thus, the aggregate reversal cannot be reduced to a malformed donor permutation, but neither is it uniform across nuisance strata.

Independent training of M_1 and M_2 also produces non-negligible residual-scale differences in several subjects. These diagnostics do not identify a single mechanism for the AP reversal. They do not support interpreting the negative contrast as evidence that correspondence-broken EEG is intrinsically more informative. Matched-record attribution therefore remains unestablished for ChineseEEG-2 Read.

Table G2: Diagnostics of negative matched-record specificity in ChineseEEG-2 Read.

Panel A: Localization and leave-one-subject-out stability

Endpoint	Δ_{spec}	**Positive subjects**
ARG0	-0.002012	1/4
ARG1	-0.000067	2/4
ARG2	-0.000386	0/4
macro-AP (ARG0/1)	-0.001039	1/4

LOSO macro-AP: $\Delta_{\mathrm{spec}} \in [-0.001421,\ -0.000708]$; all four omissions remain negative.

Panel B: Correspondence-breaking donor integrity

Diagnostic	**Observed value**
Eligible donor assignments	1,736
Same-source assignments	1,736 / 1,736
Same-record / self-donor / fixed-point assignments	0 / 0 / 0
Invalid mappings	0
Maximum donor reuse	1
Duration mismatch (median / 75th percentile)	0.004 / 0.008 s
Duration mismatch (mean / maximum)	0.026594 / 1.240 s

Panel C: Descriptive heterogeneity across nuisance strata

Endpoint	**Source-stratum Δ_{spec} range**	**Duration-stratum Δ_{spec} range**
ARG0	$[-0.002907,\ -0.000663]$	$[-0.005206,\ +0.001316]$
ARG1	$[-0.001093,\ +0.000207]$	$[-0.000939,\ +0.001334]$
ARG2	$[-0.000615,\ +0.000567]$	$[-0.000796,\ +0.000451]$

G.3 ChineseEEG-2 Listen: Cross-Task Capacity–Specificity Separation

ChineseEEG-2 Listen tests the same capacity–specificity distinction under a distinct task condition. Unlike Read, where the matched-record contrast is negative, this analysis asks whether additional EEG-conditioned capacity again separates from matched-record specificity without requiring the same effect magnitude or sign.

Under Listen, matched EEG again provides a small positive increment over the content-only model, with $\Delta_{\mathrm{cap}} = +0.000205$ (95% CI $[+0.000027, +0.000457]$). The contrast is positive in seven of eight subjects. The effect is substantially smaller than in the Read condition, so we do not interpret it as a replication of effect magnitude. Rather, it provides a distinct task condition in which adding the EEG-conditioned branch yields a small positive increment beyond semantic content alone.

The stricter matched-record contrast remains near zero: $\Delta_{\mathrm{spec}} = +0.000014$ (95% CI $[-0.000076, +0.000106]$), with five of eight subjects showing a positive direction. Thus, the Listen condition does not establish matched-record specificity. Read and Listen differ in the numerical direction of the specificity contrast—negative in Read and approximately null in Listen—but they support the same inferential distinction: a positive conditional-capacity effect does not by itself establish that the increment is tied to the correctly matched neural occurrence.

Across the two ChineseEEG-2 task conditions, EEG conditioning provides a small increment beyond content alone without consistent evidence that the increment depends on the matched neural occurrence.

G.4 Chisco Read: Endpoint-Dependent Matched-Record Specificity

Chisco Read provides a complementary within-dataset test linking the contextual role-readout result in RQ2 to the stricter $M_0/M_1/M_2$ conditional assay. The content-only baseline is already strong, with mean AP of 0.7943 for ARG0 and 0.7424 for ARG1, so the EEG-conditioned effects should be interpreted as residual increments on the order of 10^{-3}.

For ARG0, both conditional contrasts are positive. Matched EEG improves over the content-only baseline by $\Delta_{\mathrm{cap}} = +0.002541$ (95% CI $[+0.000806, +0.005438]$), with the effect positive in

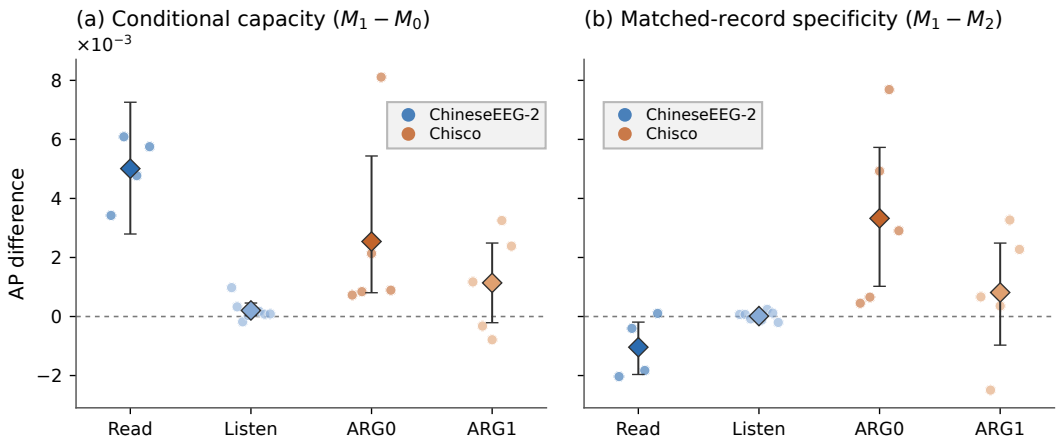

Figure G1: **Subject-level heterogeneity of conditional EEG capacity and matched-record specificity.** Each point represents a subject-level aggregate under the frozen assay, with group summaries and 95% confidence intervals shown for ChineseEEG-2 Read, ChineseEEG-2 Listen, and Chisco Read ARG0/ARG1. Positive conditional capacity does not imply positive matched-record specificity, and the latter varies across datasets and role endpoints.

all five subjects. The matched-record contrast is likewise positive, $\Delta_{\text{spec}} = +0.003325$ (95% CI $[+0.001023, +0.005730]$), again with the same direction in all five subjects. Thus, ARG0 provides the clearest case in the current analyses where the additional predictive increment is preferentially associated with the correctly matched EEG record.

The corresponding evidence is weaker for ARG1. The conditional-capacity contrast is $\Delta_{\text{cap}} = +0.001140$ (95% CI $[-0.000210, +0.002489]$), with three of five subjects showing a positive direction, while matched-record specificity is $\Delta_{\text{spec}} = +0.000814$ (95% CI $[-0.000970, +0.002487]$), with four of five subjects positive. Both confidence intervals include zero. We therefore regard ARG1 as unresolved rather than as a second positive specificity result.

Chisco therefore qualifies the ChineseEEG-2 result rather than overturning it. Matched-record specificity is detectable for ARG0 but unresolved for ARG1, while the ChineseEEG-2 contrasts are negative or near zero. The present evidence therefore supports endpoint- and dataset-level heterogeneity without identifying the mechanism responsible for that variation.

G.5 Secondary Record-Ranking and Nuisance Audits

As a secondary stress test, we asked whether the matched neural record is preferentially ranked over correspondence-breaking alternatives under a semantic-similarity criterion. This endpoint is distinct from the role-prediction assays above and is used only as supporting evidence for record specificity.

On the initial record set, matched and shuffled records produced nearly identical semantic-similarity scores. The cosine difference was $\Delta = +0.000479$ (95% CI $[-0.000128, +0.001190]$), whereas the corresponding ranking metric showed a weak positive MRR difference, $\Delta\text{MRR} = +0.001281$ (95% CI $[+0.000046, +0.002786]$). As summarized in Table G3, the non-overlapping-record analysis did not preserve the initial weak ranking advantage. The cosine difference remained very small, $\Delta = +0.000193$ (95% CI $[-0.000365, +0.000756]$), while the MRR contrast changed sign, $\Delta\text{MRR} = -0.000477$ (95% CI $[-0.001913, +0.001023]$). This is a same-subject, non-overlapping-record reproduction rather than an independent-cohort replication. Accordingly, the initial positive MRR contrast is not treated as stable evidence that the matched neural occurrence is preferentially identifiable under this secondary ranking assay.

A complementary nuisance audit further shows that this secondary endpoint is sensitive to coarse record context. In particular, conditioning the record prior jointly on source and duration improves the cosine-based ranking score by approximately $+0.025136$ on the initial set (95% CI $[+0.021919, +0.028452]$) and $+0.027061$ on the non-overlapping set (95% CI $[+0.024678, +0.029563]$), substantially larger than the matched-versus-shuffled EEG differences

on the same endpoint. Table G4 reports the full comparison among global, source-conditioned, source–duration-conditioned, shuffled-EEG, and matched-EEG alternatives.

The source–duration prior therefore provides a strong competing explanation for the secondary cosine-based ranking endpoint. This result does not explain the role- or relation-level AP effects in RQ2 or the primary $M_0/M_1/M_2$ conditional effects above, but it prevents the record-ranking assay from providing independent support for occurrence-specific attribution.

Table G3: **Secondary record-ranking results on initial and non-overlapping record sets.** Differences are matched minus correspondence-broken EEG.

Record set	Metric	Matched	Shuffled	Δ [95% CI]
Initial	Cosine	0.6796	0.6791	$+0.000479$ $[-0.000128, +0.001190]$
Initial	MRR	0.0294	0.0281	$+0.001281$ $[+0.000046, +0.002786]$
Non-overlapping	Cosine	0.6843	0.6841	$+0.000193$ $[-0.000365, +0.000756]$
Non-overlapping	MRR	0.0310	0.0315	-0.000477 $[-0.001913, +0.001023]$

Table G4: **Nuisance-conditioned alternatives for the secondary cosine-ranking endpoint.**

Condition	Initial	Non-overlapping
Global prior	0.6616	0.6639
Source prior	0.6804	0.6850
Source$\times$duration prior	0.6868	0.6910
Shuffled EEG	0.6791	0.6841
Matched EEG	0.6796	0.6843

Taken together, conditional EEG capacity is more consistent than matched-record specificity across the evaluated settings. ChineseEEG-2 Read and Listen show small capacity increments without corresponding matched-record specificity, whereas Chisco provides a positive ARG0-specific effect but unresolved ARG1 evidence. The secondary record-ranking assay provides no stronger support because its initial MRR advantage does not reproduce and its cosine endpoint admits a strong source–duration alternative. Taken together, the evidence relevant to broader occurrence-specific attribution remains weak and heterogeneous across datasets and endpoints.

H Reconstruction and Evaluation Diagnostics

H.1 Reconstruction Protocol and Diagnostic Endpoints

Following the structure-conditioned reconstruction interface defined in Section 2.1, we use the frozen language generator to compare role-neutral semantic content U with otherwise matched inputs augmented by frame-lite structure F. For Chisco, the structural payload refers to the available frame-lite predicate–argument organization—role assignments and directed predicate–argument edges—rather than the inter-frame relations defined for ChineseEEG-2.

The primary reconstruction assay uses a frozen Chisco Read reconstruction cohort from the `sub-01/read/seed0` test split, comprising 1,284 sample-level EEG items corresponding to 1,257 unique linguistic records. For each reconstruction item, U and the generation procedure are held fixed while only the supplied structural information is varied. We compare five inputs: U alone; U with predicted role assignments; U with predicted roles and directed predicate–argument edges; U with shuffled frame-lite structure; and U with gold frame-lite structure. The shuffled and gold conditions serve as diagnostic controls for the use of frame-lite structural information rather than deployable inference settings. All conditions use the same frozen DeepSeek-v4-flash generator with deterministic decoding (temperature 0).

The U-only condition is conceptually analogous to semantic-unit-guided reconstruction in BRAINMOSAIC (Li et al., 2026); the project-local BRAINMOSAIC-style reference uses a different decoder and evaluation protocol and is therefore not compared numerically here.

As described in Section 3.5, reconstruction is used only as a downstream diagnostic. For the structure-augmentation assay, sentence reconstruction similarity (SRS) is the cosine similarity between the L2-normalized BGE-M3 embeddings of the generated and reference sentences. The complementary structural scores are computed from the output of the condition-blind parser: *Role F1* is micro-F1 over $(\text{node}, \text{role})$ facts, *Edge F1* is micro-F1 over directed $(\text{predicate}, \text{argument}, \text{ARG}k)$ facts, and *Frame-lite F1* is the mean item-level F1 over the union of role and edge facts. These parser-derived scores quantify structure expressed in the generated sentence, rather than agreement with the structural fields supplied to the generator.

For the binding-reversal assay, *original-binding fidelity* and *reversed-binding fidelity* indicate whether the parsed generation matches the target binding in the corresponding arm, and *bidirectional binding fidelity* requires both to hold for the same pair. *Exact output identity* records whether the two generated strings are identical. We separately report a *parsed edge-set response*, which indicates a change in the parser-derived directed ARG0/ARG1 edge set. This response label does not by itself imply a correct reversal of the target binding.

The manipulation assay also reports an original–reversed *pair cosine*, computed between the BGE-M3 embeddings of the two generated sentences for a content-matched pair. This quantity is distinct from SRS: SRS compares a generated sentence with its reference sentence, whereas pair cosine compares the generations produced under the original and reversed bindings. The condition-blind parser uses the same frozen model backend with a separate prompt and receives the generated sentence and lexical nodes, but not the reconstruction condition or the structural fields supplied to the generator. All binding-fidelity labels are therefore parser-assessed rather than independently human-validated linguistic judgments.

H.2 Naturalistic Structure Augmentation

Table H1 reports the complete same-protocol reconstruction matrix underlying the summary in Section 3.5. Across all five conditions, the role-neutral semantic content U, evaluation cohort, and generation procedure are fixed; only the supplied structural information is changed. In addition to SRS, we report Role, Edge, and Frame-lite F1 computed from the generated sentences.

Table H1: **Downstream reconstruction under frame-lite structure augmentation.** All conditions use the same role-neutral semantic content U and generation protocol; only the supplied frame-lite structure varies. SRS is reference–generation BGE-M3 cosine, and the remaining columns are parser-derived structural F1 scores.

Reconstruction input	SRS ↑	Role F1 ↑	Edge F1 ↑	Frame-lite F1 ↑
Role-neutral semantic units (U)	0.543774	0.329521	0.173115	0.237968
U + predicted roles	0.543774	0.329076	0.172629	0.237767
U + predicted roles & edges ($\hat{F}$)	0.543826	0.329412	0.173600	0.238026
U + shuffled structure (F_{shuffle})	0.543751	0.329189	0.173115	0.237838
U + gold structure (F_{gold})	0.543700	0.329408	0.174113	0.238220

Across the five inputs, the downstream scores remain nearly unchanged. The full ranges are only 1.26×10^{-4} for SRS, 4.45×10^{-4} for Role F1, 1.48×10^{-3} for Edge F1, and 4.53×10^{-4} for Frame-lite F1. Notably, supplying gold frame-lite structure does not produce a consistent improvement over the role-neutral content condition on these endpoints. This pattern indicates naturalistic reconstruction insensitivity and should not be interpreted as evidence that relational structure is functionally irrelevant.

H.3 Controlled Binding-Reversal Diagnostics

To separate generator responsiveness from metric sensitivity, we perform controlled ARG0/ARG1 binding reversals while holding the role-neutral content U fixed. For each pair, the lexical content

is unchanged and only the ARG0/ARG1 assignment is swapped. The naturalistic condition uses complete payloads retained from the frozen Chisco Read reconstruction set, whereas the synthetic condition uses reversible minimal semantic triples with unambiguous argument assignments. Eligibility requires a complete single-predicate payload with both ARG0 and ARG1; collapsing repeated identical payloads leaves 16 unique naturalistic groups as the analysis units.

Table H2: **Controlled binding-reversal diagnostics.** Naturalistic results use 16 unique payload groups and synthetic results 24 reversible minimal pairs. Naturalistic intervals are 95% group-bootstrap CIs. Pair cosine compares generations under the original and reversed bindings.

Diagnostic	Naturalistic	95% CI	Synthetic
Original-binding fidelity	9/16 (56.3%)	[0.313, 0.813]	24/24 (100%)
Reversed-binding fidelity	3/16 (18.8%)	[0.000, 0.375]	24/24 (100%)
Bidirectional binding fidelity	1/16 (6.3%)	[0.000, 0.188]	24/24 (100%)
Parsed edge-set response	2/16 (12.5%)	[0.000, 0.313]	24/24 (100%)
Exact output identity	14/16 (87.5%)	[0.688, 1.000]	0/24 (0%)
Pair cosine, mean	0.9887	–	0.9851
Pair cosine, median	1.0000	–	0.9868
Pair cosine, range	[0.8940, 1.0000]	–	[0.9537, 0.9948]

Under naturalistic payloads, the generated text is largely insensitive to the supplied binding reversal: 14 of 16 pairs produce exactly identical outputs, and only 1 of 16 is binding-faithful in both the original and reversed conditions. The parsed edge-set response is 2/16, but, as defined in Section H.1, this endpoint only records a change in the parser-derived directed edge set and does not necessarily indicate a correct reversal. The stricter bidirectional-fidelity criterion therefore provides the more direct measure of successful binding use in this assay.

In contrast, all 24 synthetic pairs are binding-faithful in both directions, and none produces identical outputs across the two conditions. Thus, the generator is capable of following explicit predicate–argument assignments under controlled reversible inputs. Despite these verified reversals, the original–reversed BGE-M3 pair cosine remains high (mean $= 0.985$, median $= 0.987$), showing that a substantial change in argument binding can correspond to only a small change in global sentence-embedding similarity.

The following examples were selected by frozen pair-ID ordering within their prespecified diagnostic strata. A naturalistic example makes the first pattern concrete. With $U = \{$需要, 我, 工作$\}$, reversing ARG0 = 我, ARG1 = 工作 to ARG0 = 工作, ARG1 = 我 leaves the generated sentence unchanged in both conditions: “我需要工作。” (pair cosine $= 1.000$). For a synthetic reversible example with $U = \{$支持, 刘伟, 陈浩$\}$, the generator correctly changes “刘伟支持陈浩。” to “陈浩支持刘伟。” after the binding reversal, while the pair cosine remains 0.988.

H.4 Interpretation and Evidential Boundary

Taken together, the reconstruction assays do not provide an independent basis for attributing relational structure to the neural representation. Under naturalistic inputs, predicted, shuffled, and gold structure produce nearly identical downstream reconstruction scores, whereas the synthetic reversal control shows that the generator can follow explicit ARG0/ARG1 bindings. Even correctly realized role reversals can remain close in global sentence-embedding space. Flat naturalistic reconstruction should therefore be interpreted neither as evidence that relational structure is functionally irrelevant nor as evidence that the generator is intrinsically incapable of using structure.

Reconstruction is therefore most informative here as an evaluation diagnostic of naturalistic generation and sentence-embedding sensitivity. Conclusions about relational readout, conditional EEG capacity, and occurrence-specific attribution are based on the dedicated analyses in Sections 3.3–3.4 and Appendices F–G, rather than on reconstruction performance.

I Related Work and Positioning

I.1 Neural Language Decoding and Intermediate Semantic Representations

From concept decoding to richer neural language reconstruction. Neural language decoding has progressively expanded the form of meaning that is recovered from brain activity. Alongside speech-decoding work targeting phonetic, articulatory, or acoustic representations (Moses et al., 2021; Silva et al., 2024), early concept-decoding studies predicted predefined semantic concepts or categories from ECoG or EEG responses (Wang et al., 2011; Correia et al., 2015; Foster et al., 2021). More recent work has pursued continuous or open-vocabulary semantic alignment and richer neural-to-language reconstruction, including EEG-to-text translation, open-vocabulary alignment, and continuous-language reconstruction (Duan et al., 2023; Lu et al., 2025; Shams et al., 2025). This progression shifts the central question from whether coarse semantic distinctions are decodable to what linguistic information is preserved by increasingly expressive neural–language interfaces.

Intermediate semantic interfaces. Semantic bottlenecks make this question more inspectable by placing an explicit representation between noisy neural signals and natural language. Semantic Intent Decoding in BRAINMOSAIC represents communicative content as a flexible, variable-sized set of continuous semantic units that can support subsequent reconstruction (Li et al., 2026). Brain-CLIPLM instead recovers an ordered sequence of semantic anchors before retrieval-grounded language-model reconstruction (Yang et al., 2026). These approaches provide useful, interpretable content-level interfaces: they make decoded semantic content explicit without requiring the neural signal to determine every surface-form choice directly. These interfaces primarily make semantic content explicit, but they do not explicitly parameterize predicate-argument binding as a decoding target. We therefore study the relational boundary of such intermediate semantic representations rather than proposing a competing semantic decoding architecture.

I.2 Relational and Compositional Structure in Neural Language Representations

Sentence meaning depends not only on which concepts are present, but also on how they are compositionally and relationally organized. Neurocognitive accounts distinguish lexical-semantic content from the combinatorial operations that construct structured interpretations (Friederici, 2011), while MEG responses track the incremental composition of conceptual and logical elements (Zhang & Pylkkänen, 2018).

Controlled neuroimaging and neural-recording studies identify complementary forms of role- and structure-sensitive organization. During comprehension, neural responses reflect argument hierarchies involved in assigning participants to actor- and undergoer-like roles (Bornkessel et al., 2005). Multivariate fMRI studies distinguish event participants by their roles in sentence meaning and directly identify thematic roles such as agent and patient (Frankland & Greene, 2015; Wang et al., 2016); related work further distinguishes forms of compositional semantic representation across brain regions (Frankland & Greene, 2020). Most directly, Wang et al. (2024) study predicate-argument relations with counterbalanced subject-verb-object stimuli in fMRI, testing decoding across subjects and across linguistic and visual modalities. Our setting differs in both regime and evidential target: we study naturalistic non-invasive EEG semantic decoding and ask not only whether relational information is readable, but whether its predictability survives content and decoder-related controls and depends on the matched neural record. During sentence production, neural patterns separate lexical or content information from relational event structure, and ECoG reveals syntactic-role encoding and structure-dependent temporal dynamics (Giglio et al., 2024; Morgan et al., 2025). These targets are related but not identical: thematic roles characterize semantic participation, predicate–argument organization binds participants to a predicate, and subject/object roles characterize grammatical structure. Together, these studies establish compositional and role-sensitive organization as a well-motivated target for neural language analysis.

EEG and ERP studies likewise show sensitivity to thematic-role processing in controlled sentence comprehension. This work links thematic-role computation to semantic-relatedness processing (Li et al., 2014), identifies ERP correlates of role assignment across active and passive sentences (Jackson et al., 2020), and examines agent-patient assignment in a patient-first language, providing a

setting in which thematic-role assignment is dissociated from a simple agent-first linear heuristic (Sauppe et al., 2023).

Taken together, prior work establishes compositional and thematic-role structure as a measurable neural target, including in EEG. We address a complementary identifiability question within naturalistic non-invasive EEG semantic decoding: whether contextual role-related readout remains accessible beyond semantic-unit identity, whether relation predictability is specific to contextual EEG-derived representations rather than alternative feature sources, and whether any EEG-conditioned advantage depends on the matched neural occurrence.

I.3 Probing, Controls, and Neural Identifiability

A successful probe establishes that a target is predictable from a representation, but predictability alone does not establish its information source, its functional relevance to the underlying computation, or its dependence on the matched neural occurrence. Work on control tasks emphasizes that probe performance must be interpreted relative to probe capacity and appropriate comparators (Hewitt & Liang, 2019); analyses of neural encoding and decoding likewise separate predictive success from claims about the represented mechanism (Kriegeskorte & Douglas, 2019). Moreover, probe accuracy need not imply relevance to the behavior or computation of interest (Ravichander et al., 2021).

These distinctions are especially consequential when lexical regularities, target structure, decoder or query priors, stimulus properties, and other nuisance variables can support the same prediction. Controls are therefore needed to distinguish general predictability from a feature source, conditional EEG predictive information beyond semantic content, and dependence on the correctly corresponding neural record.

We apply this methodological distinction to relational structure in EEG semantic decoding. The framework separates the *functional relevance* of predicate–argument organization from structural or representation-level *readout*, then distinguishes *conditional EEG predictive capacity* from *matched-record specificity*. Occurrence-specific attribution is the broader cumulative claim supported only when these forms of evidence converge. This positioning connects established content-level and compositional findings to a stricter attribution question in naturalistic EEG decoding.