SHARED SEMANTIC DISTINCTIONS AND MODEL-DEPENDENT EXPRESSION IN TEXT ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text embeddings support semantic decisions, yet downstream success does not explain which distinctions their vectors encode or how those distinctions are organized. Recent work suggests that independently trained models converge geometrically, raising the question of what meaning they encode in common. We study seven frozen text encoders using controlled sentences that vary six distinctions: success versus failure (outcome), affirmation versus negation (polarity), which event happens first (order), who acts on whom (role), what contains what (containment), and who has which attribute (binding). We compare recovery across constructions and vocabulary, correspondence in responses to semantic changes, and decision-rule transfer as a diagnostic of representational compatibility. Outcome and polarity remain readily recoverable, while relational distinctions show model- and expression-dependent generalization. Binding reveals a clear dissociation: nonlinear readouts improve recovery across all seven encoders on held-out constructions, yet corresponding response geometry coexists with a substantial transfer deficit through the tested linear maps relative to matched within-model readouts. Useful alignment constraints also differ by distinction: matching full responses benefits order, whereas matching their averages performs better for containment. These findings show how apparent semantic deficits can depend on the readout and why geometric correspondence alone is insufficient to establish compatible decision rules. They provide a controlled account of semantic content and organization, clarifying which conclusions about shared meaning persist across linguistic expressions and ways of accessing embeddings.

1 INTRODUCTION

Text embeddings support semantic search, sentence comparison and classification by representing language as vectors (Conneau et al., 2017; Reimers & Gurevych, 2019). Their usefulness, however, does not explain which semantic distinctions their vectors make accessible or how those distinctions are organized. Prior work provides evidence of common representational geometry and useful cross-model correspondences (Huh et al., 2024; Huang et al., 2025; Jha et al., 2025), raising the possibility that independently trained encoders share aspects of semantic structure. Understanding what this agreement means requires examining the distinctions themselves. We therefore ask: **What parts of meaning do independently trained text embedding models encode in common?**

Consider "The doctor examined the lawyer" and "The lawyer examined the doctor." The sentences share participants and action but reverse who acts on whom. Two encoders may each allow a classifier to recover this distinction, but that alone does not show that they express it in the same way. We ask whether recovery survives new constructions and vocabulary, whether reversing the roles produces corresponding changes in their embeddings, and whether one encoder's rule works on the other's vectors. These questions distinguish **joint recoverability**, separate recovery of the same distinction from each model; **common geometric expression**, corresponding responses to controlled semantic changes; and **transferable use**, source-model decisions that work through a learned cross-model map. Rule transfer thus serves as a diagnostic of representational compatibility. Figure 1 shows how these measurements agree or diverge across the six distinctions.

We investigate seven frozen text embedding models using controlled sentences spanning six distinctions: success versus failure (outcome), affirmation versus negation (polarity), event sequence

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

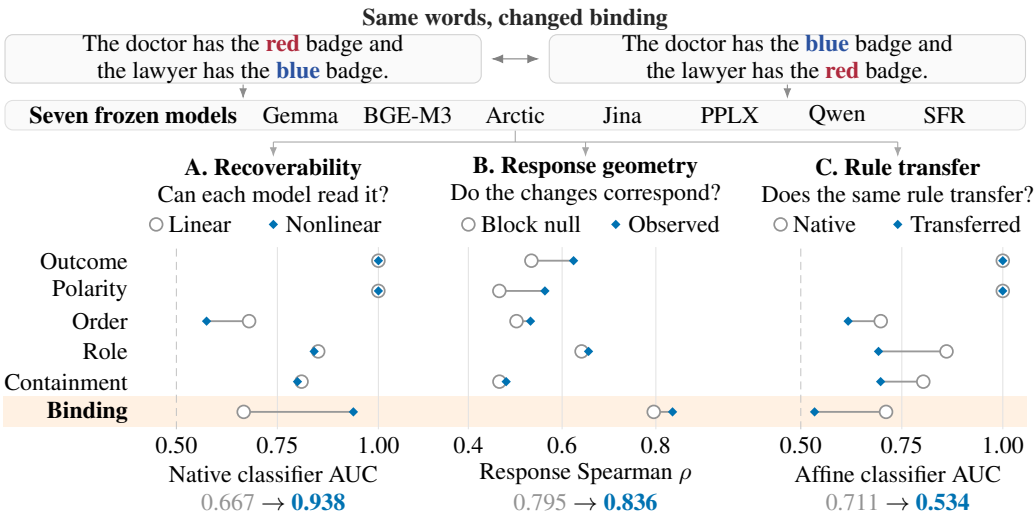


Figure 1: **Recoverability, corresponding geometry and rule transfer are distinct observations.** The binding example omits site phrases; highlighted values are repeated below. All panels use held constructions and known vocabulary. **A:** matched linear/nonlinear readouts (seven-model means). **B:** residual-response correlation versus filler-block null median (21-pair means). **C:** full-contrast linear-map transfer versus matched centered-affine native readouts (42-direction means). C's affine reference differs from A's readouts. Lines connect point estimates, not uncertainty intervals.

(order), who acts on whom (role), what contains what (containment), and who has which attribute (binding). Twenty construction-family holdouts and separate vocabulary shifts test generalization beyond the expressions used for fitting. We compare matched linear and nonlinear classifiers, measure relationships among models' responses to semantic changes, and evaluate cross-model decision-rule transfer against matched within-model references. This design connects the three observations on the same controlled distinctions and evaluation scenes.

Our findings reveal three aspects of semantic commonality:

1. **Recoverability depends on the distinction and its expression.** Outcome and polarity remain readily recoverable, whereas order, role, containment and binding exhibit different patterns across models, constructions and vocabulary shifts. Common success on familiar expressions therefore does not imply uniform generalization (Figure 2, Table 1).

2. **Apparent representational deficits can depend on the readout.** Nonlinear classifiers improve binding recovery over matched linear classifiers across all seven models and held-family cohort means. Role's nonlinear advantage, however, changes sign with the shift. These results show why conclusions about accessible semantic content must account for how embeddings are read (Figure 3).

3. **Corresponding geometry does not guarantee a portable decision rule.** Responses exhibit cross-model correspondence beyond the declared construction- and filler-preserving null, yet binding retains substantial transfer deficits through the tested linear interfaces. Useful alignment constraints also differ by distinction: order benefits from matching full responses, whereas matching average responses performs better for containment and for role under combined construction and vocabulary shifts (Figure 4, Figure 5, Section 5.1).

These results make semantic commonality testable at the level of individual distinctions. They show which findings persist across changes in expression and measurement, and where models differ.

2 BACKGROUND

2.1 PRELIMINARIES

Let t denote an input sentence and f_i a frozen text encoder. Encoder i produces a d_i-dimensional vector using fixed tokenization, prompting and encoding settings. We use its L2-normalized output $x_i(t) = \mathrm{unit}(f_i(t))$, where $\mathrm{unit}(v) = v/\left\|v\right\|_2$ for nonzero v.

Each binary semantic distinction c has a label $y_c(t) \in \{0, 1\}$. For example, binding can distinguish whether the doctor has a red or blue badge. Let D denote a labeled evaluation set under specified construction and vocabulary conditions. Construction-fold indices are suppressed below.

Recoverability. A readout $h_{i,c}$ maps encoder i's vector to a score for distinction c. Evaluating $h_{i,c}(x_i(t))$ on D measures recovery under that readout and linguistic condition. Training and validation data determine the readout before evaluation. Joint recoverability permits a different readout for each encoder.

Geometric expression. For distinction c, matched pair (t_k^-, t_k^+) changes the label from 0 to 1 while holding the background assignment fixed. Encoder i's response is

$$\Delta_{i,c,k} = x_i(t_k^+) - x_i(t_k^-).$$

We compare within-encoder cosine-similarity patterns among responses across models, also after subtracting the mean response from probe-training pairs. This comparison accommodates different coordinates and embedding dimensions.

Rule transfer. For transfer, let $z_i(t) = \mathrm{unit}(x_i(t) - \mu_i)$, where μ_i is a mean estimated from normalized training embeddings. Separate readouts $g_{i,c}$ operate on this centered view. Applying A's rule to B through an interface $T_{B\to A}$ gives

$$s_{A\to B,c}(t) = g_{A,c}(T_{B\to A}(z_B(t))). \tag{1}$$

The rule transfers from A to B, while vectors are mapped from B to A. B's own readout, fitted in the same representation view and with the same label budget, provides a matched reference. Performance depends jointly on the target representation, interface and source readout.

2.2 RELATED WORK

Sentence probing examines whether $h_{i,c}$ can recover a linguistic property from $x_i(t)$ (Conneau et al., 2018). Interpretation depends on the readout: Hewitt & Liang (2019) show why task accuracy must be considered alongside probe capacity. It also depends on D. PAWS and HANS expose failures concealed by lexical overlap or syntactic shortcuts (Zhang et al., 2019; McCoy et al., 2019), while SemAntoNeg documents model-dependent sensitivity to negation and antonymy (Vahtola et al., 2022). BERT-flow and Semantic Adapter show that transforming or reweighting embeddings can improve semantic utility (Li et al., 2020; Cao, 2025). Recoverability therefore concerns the representation, readout and evaluation condition together.

Complementing these readout-based tests, the Platonic Representation Hypothesis concerns correspondence among geometric relationships induced by representations $x_i(t)$ across models (Huh et al., 2024). The Linear Representation Hypothesis formalizes concept directions and connects their geometry to probing and steering (Park et al., 2024). Bangachev et al. (2026) relate alignment to linear object–attribute structure, distinguishing signal, bias and noise. These perspectives motivate examining responses $\Delta_{i,c,k}$ to specified semantic changes alongside whole-text agreement. Measurement matters: similarity statistics have different invariances (Kornblith et al., 2019), cosine similarities can depend on parameterization and regularization (Steck et al., 2024), and raw agreement can require calibration against a null (Gröger et al., 2026).

Such geometric correspondence also motivates cross-model interfaces that test whether representations support compatible operations. Relative representations construct anchor-based coordinates (Moschella et al., 2023), while embedding translation learns mappings between spaces (Jha et al., 2025). Related work transfers concept steering vectors or aligns models to reusable concept atlases

(Huang et al., 2025; Puri et al., 2026). Our transfer expression in (1) makes the representation, interface and source readout explicit. Compatibility is conditional: Hamideche et al. (2026) report model- and task-dependent translation failures, and Usama & Chang (2026) find representation–reasoning dissociations. Amnesic probing further distinguishes recoverable information from information that influences model behavior (Elazar et al., 2021). These studies concern different interfaces and behavioral tests, providing complementary evidence about functional correspondence.

Taken together, these studies establish both correspondence and important dissociations between measurements. What remains less clear is how recoverability, geometric expression and transfer relate for the same distinction c as the readout and evaluation condition change. A distinction may be recoverable through separate $h_{i,c}$, exhibit corresponding responses $\Delta_{i,c,k}$, and still support different generalization or transfer behavior. Establishing one observation leaves the others to be measured.

We investigate these relationships on common controlled distinctions and evaluation scenes. Varying readouts on fixed embeddings tests whether apparent deficits depend on access. Varying constructions and vocabulary tests whether findings persist across expressions. Comparing response correspondence with transfer through T, against matched within-model references, tests compatibility of that expression. Together, these comparisons characterize which conclusions about semantic commonality persist across measurements and which depend on them.

2.3 Experimental design

Encoders and semantic distinctions. We evaluate EmbeddingGemma, BGE-M3, Arctic Embed, Jina v5, PPLX-Embed, Qwen3-Embed and SFR-Embed (768–4,096 dimensions, approximately 300 million–8 billion parameters). The cohort spans different architectures and training objectives. Model and tokenizer versions, prompts, numerical precision and input-length limits are in Appendix A.

Controlled templates express the six distinctions using fixed focal terms such as the doctor, box, briefing and red/blue badges. Each base scene crosses two target labels with two background assignments, yielding four sentences. Swapping the doctor/lawyer badge assignments, for example, changes binding while preserving the word multiset. Independent text-to-fact checks reconstruct the labels. Definitions, examples and the checks' linguistic scope are in Appendix B.1.

Construction families and vocabulary shifts. Outcome/polarity/order/role/containment/binding have 3/2/3/4/4/4 construction families, respectively, totaling 50 templates and 20 leave-one-family-out folds. Each fold uses 288 training and 96 separate validation sentences from the other families. Sixty-four final scenes per distinction are excluded from all folds' training and validation sets. Known and held-out nuisance vocabularies give four test conditions:

Construction families	Known vocabulary	Held-out vocabulary
Used for fitting	Context	Lexical
Held-out family	Construction	Joint

Held-out vocabulary is excluded from supervised semantic fitting, with focal terms fixed. Each model embeds 35,860 unique texts. Q/A and clock-order are separate stress panels. Families group related grammatical realizations, not independent draws of English grammar.

Readouts and uncertainty. Each distinction, fold and model receives mean-direction, affine logistic, linear128 and GELU128 readouts. The latter pair uses 128 hidden units with matched initialization and batch ordering, differing only in activation. The primary analysis uses one prespecified seed. Role/binding neural comparisons repeat with two additional seeds. All readouts freeze before final evaluation. Map and geometry analyses are specified after inspecting native outcomes and frozen before their own final evaluation. Transfer uses means μ_i from 5,000 ordinary training texts, separately fitted affine $g_{i,c}$, and interfaces $T_{B\to A}(z) = \mathrm{unit}(W_{B\to A}z)$ with $W_{B\to A} \in \mathbb{R}^{d_A \times d_B}$. Map objectives and supervision are in Section 5. We report ROC-AUC. Accuracy uses validation-selected thresholds. AUC is computed per template and averaged equally within families and across folds. Different heads' scores are never pooled. Ten thousand paired bootstrap draws resample complete scenes together across templates, folds, models and vocabularies. Intervals condition on the fixed cohort, families and fits, without adjustment for adaptive design or multiple discoveries. Full recipes,

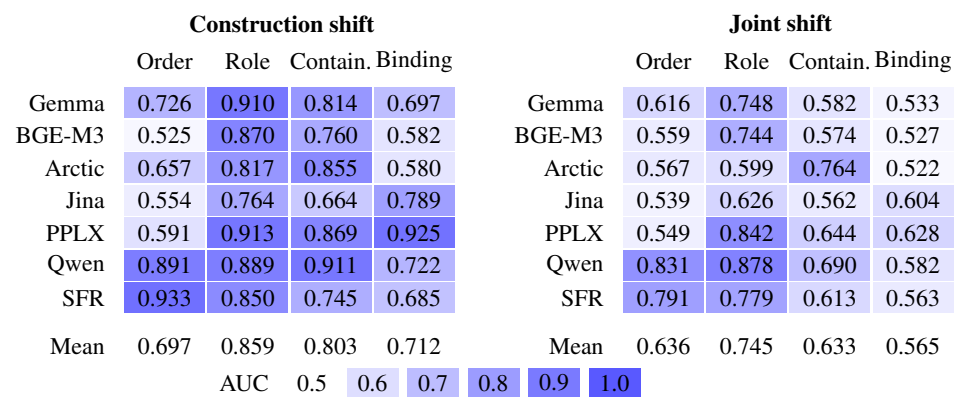


Figure 2: Affine recovery varies by distinction and encoder under construction and joint shifts. Each cell averages AUC across held-family folds for one encoder; the final scenes are paired across vocabulary variants. These are family averages, not worst-family scores. Table 1 includes outcome and polarity.

Table 1: Affine AUC, mean ± SD across seven fixed encoders. SD describes model heterogeneity after family averaging, not uncertainty or seed variation. Rounded 1.000 need not mean perfect discrimination.

Distinction	Context	Construction	Lexical	Joint
Outcome	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Polarity	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Order	0.999 ± 0.001	0.697 ± 0.150	0.871 ± 0.079	0.636 ± 0.113
Role	0.998 ± 0.002	0.859 ± 0.050	0.840 ± 0.082	0.745 ± 0.095
Containment	0.991 ± 0.011	0.803 ± 0.079	0.785 ± 0.054	0.633 ± 0.068
Binding	0.849 ± 0.136	0.712 ± 0.111	0.602 ± 0.054	0.565 ± 0.038

seeds and all 24 native, 168 transfer and 48 geometry intervals are in Appendix B, Appendix E and Appendix D.

3 Recoverability across expressions and readouts

Recoverability depends on both expression and readout. Following Section 2.1, we first vary the evaluation condition D with each fitted $h_{i,c}$ fixed, then compare readouts on the same embeddings $x_i(t)$ and supervised examples.

3.1 Affine recovery generalizes unevenly across distinctions

Outcome and polarity remain near ceiling across all four test conditions (Table 1). Relational distinctions generalize less consistently. Order shows the largest construction-shift loss: mean AUC falls from 0.999 on new scenes in familiar constructions to 0.697 on held-out constructions. Role, containment and binding also decline. All four paired 95% scene intervals exclude zero (Table 18), and adding held-out vocabulary lowers their means further.

The profiles vary across both encoders and grammatical realizations. SFR/Qwen lead affine order recovery, whereas PPLX leads binding on this inventory (Figure 2). For order, held-family AUC averages 0.818 for ordinal sequences, which explicitly identify the first and second events, but 0.540 for perfect-aspect constructions such as "When the survey began, the briefing had already finished." Recovery of the same distinction therefore depends on the held-out expression (Appendix B.5).

Character features also nearly saturate outcome, making its high recovery less diagnostic of semantic organization. Word and character controls instead score 0.5 on construction-shift binding (Appendix B.6). Does binding's weaker recovery persist when we change $h_{i,c}$?

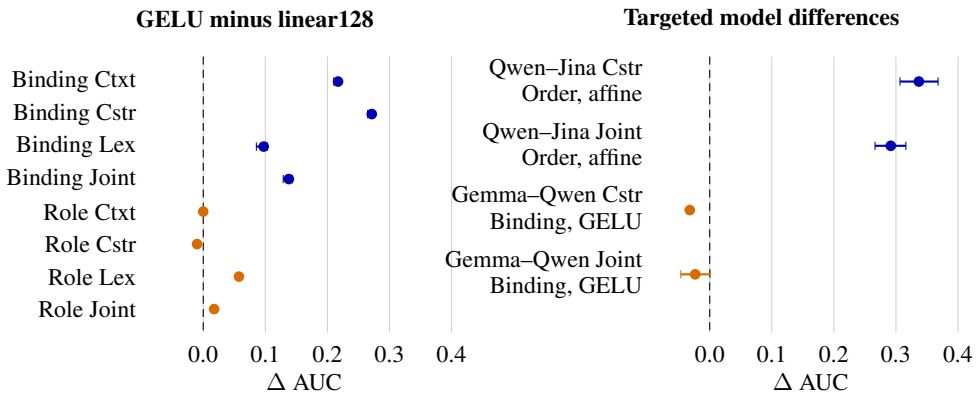


Figure 3: Readout and model comparisons, with conditional 95% scene intervals. Left: GELU improves binding recovery across test conditions, whereas role's effect changes sign. Right: targeted model differences for the labeled distinction and readout. Ctxt, Cstr, Lex and Joint denote context, construction, lexical and joint conditions. Pairs were selected from prior observations before final readout evaluation; intervals do not resample models or grammatical families.

3.2 Nonlinear readouts improve binding recovery across encoders

To test readout dependence, we compare two versions of $h_{i,c}$ with 128 hidden units and matched training: one uses GELU, the other an identity activation. The latter, linear128, controls for the activation change. It differs from the direct affine reference in Table 1 in optimization and regularization, despite sharing its class of score functions.

Changing the activation substantially improves binding recovery (Figure 3). Construction AUC rises from 0.667 to 0.938, with a paired difference of 0.271 [0.265,0.276]. Joint-shift AUC rises from 0.560 to 0.698, a difference of 0.138 [0.129,0.144]. Both point gains are positive in all seven encoders and every held-family cohort mean. The cohort-mean gains remain positive across all three seeds (Appendix B.4). The same $x_i(t)$ thus supports more binding recovery than the matched linear readout reveals.

The improvement does not remove generalization limits. Joint-shift GELU AUC ranges from 0.547 to 0.793 across encoders, and mean accuracy is 0.606, versus 0.877 for construction-only shifts. Nor does GELU improve every distinction: for role, it lowers construction AUC by 0.010 but raises lexical and joint AUC by 0.058 and 0.018, respectively, with all three intervals excluding zero. Readout benefits therefore depend on both the distinction and the test condition.

Targeted model comparisons further show that a consistent readout benefit need not imply a uniform model ordering. Qwen exceeds Jina on affine order under both construction and joint shifts. For GELU binding, however, Gemma's construction deficit relative to Qwen is concentrated in badge assignment. Their joint-shift differences change sign across families, and the aggregate interval crosses zero without establishing equivalence (Figure 3, Table 15, Table 18).

These controlled comparisons establish that recoverability depends on both $h_{i,c}$ and D.

3.3 Readout dependence on natural sentence pairs

Does readout dependence also appear in natural text? We compare matched linear128/GELU128 heads on MultiNLI (Williams et al., 2018), using all seven encoders. Premise and hypothesis embeddings p, h form $[p, h, |p - h|, p \odot h]$, so even the linear head uses features nonlinear in the individual sentence vectors. We train on 9,994 pairs and select heads using 2,000 validation pairs, freezing them before both public-dev splits are scored.

GELU improves mean macro-F1 by 0.0072 [0.0044,0.0100] on matched genres and 0.0054 [0.0028,0.0081] on mismatched genres, with negative encoder-specific effects retained (Appendix C). These modest, heterogeneous gains extend readout dependence to natural entailment, without independently validating the six controlled distinctions or a binding-specific mechanism. Returning to

those distinctions, we next ask whether their responses $\Delta_{i,c,k}$ exhibit common geometric expression despite differences in recovery.

4 Cross-model correspondence in semantic responses

Do models respond to a semantic change in corresponding ways, even when their readouts generalize differently? Using $\Delta_{i,c,k}$ from Section 2.1, we ask whether response pairs that are similar within one encoder are also similar within another. Spearman correlation compares the ranks of $\cos(\Delta_{i,c,k}, \Delta_{i,c,\ell})$, for $k < \ell$, across encoders. This requires neither shared coordinates nor a fitted cross-model map.

We also subtract each encoder's mean response on other-family training pairs, asking whether the remaining *residual responses* still correspond. Scores are computed per template and vocabulary, then averaged equally within families, across families and across the 21 model pairs. Zero responses or constant similarities leave some lexical-control correlations undefined rather than zero (Appendix D).

High correlation can reflect shared template or filler structure (Gröger et al., 2026). Our blocked null shuffles scene matches across encoders within each template and specified filler category, keeping each scene's two responses together. Its 1,000 draws break exact scene correspondence while retaining those groupings, which may preserve meaningful structure. Excess over this reference is therefore not a fraction of shared semantic information.

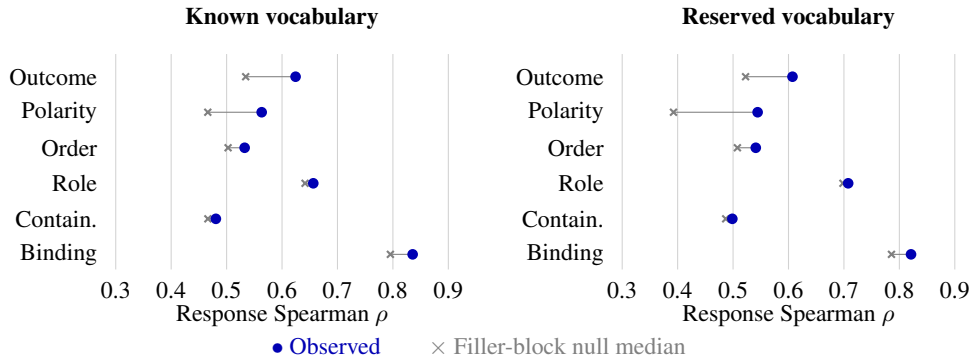


Figure 4: Residual-response geometry: observed correlation and blocked-null median, averaged over the fixed 21 encoder pairs. Lines connect the two values, not confidence limits. Every distinction has positive excess in both vocabulary conditions; binding has the highest observed correlation, but outcome and polarity have larger excesses. Full responses, both null regimes, pairwise results and conditional intervals are in Appendix D.

Mean residual correlation exceeds the blocked-null median for all six distinctions under both vocabulary conditions (Figure 4). Binding has the highest observed agreement: 0.8356 versus a null median of 0.7954 with known vocabulary. All 21 binding pairs exceed their own null's 97.5th percentile under both conditions, a dependent descriptive count rather than a global significance test. Outcome and polarity have larger excesses over their respective nulls, whereas role and containment have smaller excesses. The full-response results also exceed their null medians (Table 26).

Role illustrates why observed agreement and null excess must be distinguished. With held-out vocabulary, residual correlation rises from 0.6564 to 0.7076 while native affine AUC falls from 0.8591 to 0.7452. The null median also rises, leaving a smaller excess: 0.0148 to 0.0095. Thus higher observed correlation establishes neither stronger correspondence beyond this null nor more robust recovery. We next test whether the observed geometry supports a portable decision rule.

5 From recoverable distinctions to portable rules

Can a source rule operate on another encoder's vectors? Across all 42 directed pairs, we compare $s_{A\to B,c}(t)$ in (1) with the target's own $g_{B,c}(z_B(t))$. Both use frozen affine readouts in the centered view, with matched label budgets. Map fitting uses additional data.

We vary what the map preserves. *Ordinary* alignment (O) uses 5,000 paired Dolma texts (Soldaini et al., 2024). *Endpoint-only* alignment (E) replaces 384 of these with controlled sentences, testing whether semantic examples alone help. On the same pool, *full-response* matching (F) adds constraints on semantic changes; *mean* (M) and *variation* (V) matching constrain their average and deviations from it, respectively. The pool provides 192 labeled contrasts across six distinctions, excluding the focus distinction's held family. Semantic modes share the ordinary-validation-selected ridge coefficient. Mean matching retains the contrast-label budget. Pool overlaps and selection details are in Appendix E.1.

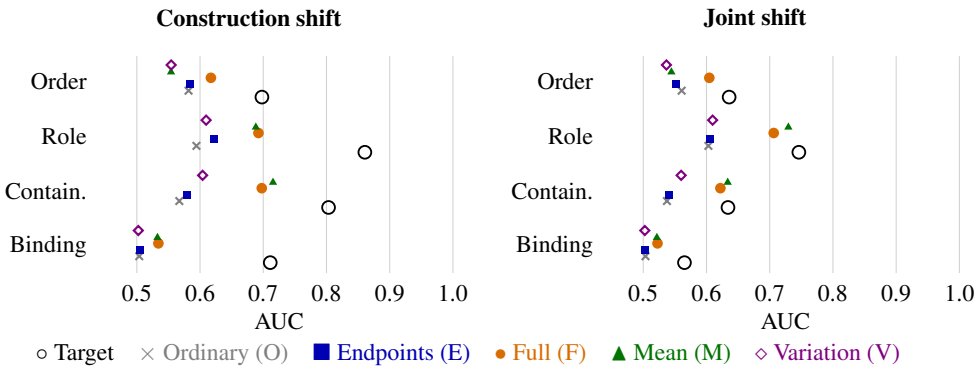


Figure 5: Mean transfer AUC over 42 directions. Target is the encoder's own centered-affine readout with the matched label budget. Full-response matching improves construction-shift transfer over endpoints alone, but target gaps remain. Complete conditions, accuracy and family/model profiles are in Appendix E.2.

Full-response matching improves construction-shift transfer over endpoints alone for all four relational distinctions, with positive conditional intervals (Figure 5). Binding nevertheless reaches only 0.5341 AUC against its matched target's 0.7111: a gap of -0.1769 [-0.1815,-0.1730], despite its response correspondence. This is an affine-transfer deficit; GELU recovery in Section 3.2 is a separate observation. Outcome remains saturated and polarity nearly so.

Mean matching approaches the joint-shift containment target: 0.6336 versus 0.6340, with difference -0.00043 [-0.00521,0.00395]. This is numerical proximity, not equivalence, and the target itself recovers less under joint shift.

5.1 Full and mean matching benefit different distinctions

Should a map match only the average semantic response or also its variation? In the centered view, write $\Delta z_{i,c,k} = z_i(t_k^+) - z_i(t_k^-)$ and let $e_{c,k} = \Delta z_{A,c,k} - W_{B\to A}\Delta z_{B,c,k}$ be the matching error. For $K_c = 32$ training contrasts of distinction c, let $\bar{e}_c$ be their mean error. Then

$$K_c^{-1}\sum_k \|e_{c,k}\|^2 = \|\bar{e}_c\|^2 + K_c^{-1}\sum_k \|e_{c,k} - \bar{e}_c\|^2 \tag{2}$$

separates mean-response and variation errors. Full matching adds both penalties; mean and variation matching add one each, with shared training-derived weights. This fitting-loss identity precedes output normalization and does not decompose AUC or shared/private information.

Order benefits from combining the constraints. Full matching exceeds mean matching by 0.0631 [0.0611,0.0659] on construction and 0.0600 [0.0575,0.0634] under joint shift; either component alone falls below endpoint-only alignment on construction. Mean matching instead exceeds full matching for containment by 0.0178/0.0116 on construction/joint, and for joint role by 0.0232, with intervals excluding zero. Role's construction difference is uncertain. Binding's full-minus-mean gains are only 0.0015/0.0010 alongside weak transfer. Neither constraint is uniformly preferable across distinctions and shifts.

5.2 Endpoint weighting and pairing qualify the gains

Could extra weight on semantic examples explain the gains? Endpoint boost (EB) increases pressure to match individual sentence vectors. Random pairing (J0) forms contrasts within each distinction; mean-preserving pairing (J1) shuffles positive partners within each distinction and construction family while fixing background assignments and average signed responses. Unlike the null in Section 4, these controls retain cross-model text matches. Eight fixed directions use their own full-response reference. Pairing weights are fixed or rescaled by training contrast energy; metrics average three seeds (Appendix E.3).

Endpoint boost exceeds full matching for all four relational distinctions on construction, but loses for joint-shift order/role; the other joint intervals cross zero. Random pairing loses in all four joint cases under both scales. The mean-preserving binding control changes from +0.0086 relative to full matching on construction to -0.0260 after energy matching. Preserving the unscaled mean does not fix the weighted penalty. Useful constraints thus depend on distinction, shift and weighting; gains cannot be uniquely attributed to meaningful pairing.

6 Discussion and scope

Interpreting shared meaning requires distinguishing recoverable content from compatible use. Binding illustrates why: nonlinear readouts improve recovery on each encoder's fixed embeddings (Figure 3), and responses correspond beyond the declared null after training-mean subtraction (Figure 4). Yet affine rule transfer falls below matched within-model affine readouts (Figure 5). Apparent semantic deficits can therefore depend on the readout, while corresponding geometry alone does not establish compatibility with a source-model rule through the tested interfaces.

The useful readout and alignment constraint are also distinction-dependent. Role's nonlinear advantage changes with the evaluation condition (Figure 3), while order benefits from full-response constraints and containment from mean constraints (Section 5.1). These differences make a single model ranking or preferred interface an incomplete account of semantic organization. By comparing recoverability, geometry and transfer on the same distinctions and scenes, we complement studies of geometric convergence and translation failures (Huh et al., 2024; Hamideche et al., 2026). The contribution is a representation-level account of which findings persist across expressions and ways of accessing embeddings.

Limitations. Coverage is limited to the tested encoders and construction families. Fixed focal terms and lexical cues limit abstraction, and parser agreement does not certify unrestricted linguistic validity. Held-family exclusion applies to the focus distinction's supervised pools, not all grammatical devices or encoder pretraining. MultiNLI checks a separate entailment task.

The tested readouts, centered-affine interfaces and geometry null bound the measurement claims. Excess over the null does not eliminate every nuisance explanation, and finite fitting budgets do not certify optimal readouts or maps. These measurements concern output representations, not encoder-internal causal mechanisms.

Scene intervals condition on fixed models, families and fits, without adjustment for adaptive design or multiple discoveries. Native results were inspected before the map and geometry analyses were specified.

Conclusion. Across the tested encoders, outcome and polarity remain readily recoverable, while relational distinctions depend more on expression and readout. Shared recoverability and corresponding responses nevertheless do not ensure portable rules through the tested interfaces. Semantic commonality is therefore distinction-specific: what can be recovered, under which changes in expression, and through which readouts and interfaces.

AI USE STATEMENT

AI coding and language tools assisted with literature exploration, implementation and debugging, construction of synthetic text renderers, numerical-checking code, figure generation and initial manuscript drafting. Controlled-data labels follow symbolic records and restricted renderers; external task labels are publisher-supplied. The computation checks described in Appendix F use saved inputs/predictions and separately implemented numerical reconstructions; these checks also involved AI-assisted code and are not a substitute for author review. The authors retain responsibility for the final text, claims, code and artifacts.

REPRODUCIBILITY STATEMENT

Encoder recipes and effective revisions are in Appendix A; data roles, readout fitting and exhaustive construction holdouts are in Appendix B. Geometry and its nulls are in Appendix D; map objectives and controls are in Appendix E. Uncertainty procedures are in Appendix B.3, Appendix D.4 and Appendix C. The appendices report model/family/seed controls, all 24 native, 168 transfer and 48 geometry intervals, and natural-task comparisons. Checksum-bound source summaries and deterministic table/figure importers preserve the numerical lineage (Appendix F). A repository link will be provided after acceptance.

REFERENCES

Kiril Bangachev, Guy Bresler, and Yury Polyanskiy. Representation alignment rests on linear structure. *arXiv preprint arXiv:2605.28870*, 2026. URL https://arxiv.org/abs/2605.28870.

Hongliu Cao. Semantic adapter for universal text embeddings: Diagnosing and mitigating negation blindness to enhance universality. In *European Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2504.00584.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Conference on Empirical Methods in Natural Language Processing*, 2017. URL https://arxiv.org/abs/1705.02364.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Annual Meeting of the Association for Computational Linguistics*, 2018. URL https://arxiv.org/abs/1805.01070.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2006.00995.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.14486.

Sid Ali Hamideche, Louis-Adrien Dufrène, Quentin Lampin, and Guillaume Larue. How far do simple transformations translate across text embedding models? *arXiv preprint arXiv:2608.05980*, 2026. URL https://arxiv.org/abs/2608.05980v2. Version 2.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://arxiv.org/abs/2501.02009.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2405.07987.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025. URL https://arxiv.org/abs/2505.12540.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1905.00414.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://arxiv.org/abs/2011.05864.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://arxiv.org/abs/1902.01007.

Luca Moschella, Valentino Maiorca, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15430.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2311.03658.

Bruno Puri, Jim Berend, Sebastian Lapuschkin, and Wojciech Samek. Atlas-alignment: Making interpretability transferable across language models. *arXiv preprint arXiv:2510.27413*, 2026. URL https://arxiv.org/abs/2510.27413v2. Version 2.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, 2019. URL https://arxiv.org/abs/1908.10084.

Luca Soldaini et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2402.00159.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference*, 2024. doi: 10.1145/3589335.3651526. URL https://arxiv.org/abs/2403.05440.

Muhammad Usama and Dong Eui Chang. Convergence without understanding: When language models agree on representations but disagree on reasoning. *arXiv preprint arXiv:2605.23315*, 2026. URL https://arxiv.org/abs/2605.23315.

Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new SemAntoNeg benchmark. In *Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 249–262, 2022. doi: 10.18653/v1/2022.blackboxnlp-1.20. URL https://aclanthology.org/2022.blackboxnlp-1.20/.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics*, pp. 1112–1122, 2018. URL https://aclanthology.org/N18-1101/.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://arxiv.org/abs/1904.01130.

APPENDIX GUIDE

The appendices provide the experimental specifications and the complete supporting results. Encoder configurations are in Appendix A; controlled sentences, readouts and metric aggregation in Appendix B; and the separate MultiNLI check in Appendix C. Response geometry and its nulls are in Appendix D, followed by transfer objectives and controls in Appendix E. Appendix F records analysis chronology and computational verification. We use the notation of Section 2.1 throughout.

A ENCODERS AND ENCODING CONFIGURATION

Table 2: The fixed encoder cohort. Approximate sizes, architectures and training objectives vary jointly, so their individual contributions cannot be isolated.

Encoder	Organization	Width	Parameters
EmbeddingGemma	Google	768	300M
BGE-M3	BAAI	1,024	568M
Arctic Embed	Snowflake	1,024	568M
Jina v5	Jina AI	1,024	677M
PPLX-Embed	Perplexity	2,560	4B
Qwen3-Embed	Alibaba	4,096	8B
SFR-Embed	Salesforce	4,096	8B

Repository identities are `google/embeddinggemma-300m`, `BAAI/bge-m3`, `Snowflake/snowflake-arctic-embed-l-v2.0`, `jinaai/jina-embeddings-v5-text-small`, `perplexity-ai/pplx-embed-v1-4b`, `Qwen/Qwen3-Embedding-8B` and `Salesforce/SFR-Embedding-2_R`. We use the checkpoints' published SentenceTransformers pooling modules without fitting encoder weights or applying PCA. All texts use the same matching recipe within a model, rather than distinct retrieval query/document recipes.

Table 3: Encoding overrides. Unspecified settings follow the pinned checkpoint. The phrase-oriented instruction is also applied to sentences; its neutrality is not assumed.

Encoder	Matching recipe
Gemma	`prompt_name="STS"`, resolved through the pinned checkpoint's prompt dictionary.
Jina	`task="text-matching"`, selecting the corresponding adapter.
Qwen, SFR	Explicit prefix: `Instruct: Represent the meaning of the following word or phrase`, then a newline and `Query:` before the input.
BGE, Arctic, PPLX	No task or prompt override.

The encoder wrapper requests unnormalized outputs and stores float32 vectors. The analysis then forms $x_i(t)$ and, for transfer, $z_i(t)$ as defined in Section 2.1. Configured token limits are listed below; the effective limit is the minimum of that value and the checkpoint's limit. Token checks reject truncated controlled inputs. Effective prompt dictionaries, tokenizer identities, package versions and batching settings are recorded with the embedding artifacts.

Table 4 records upstream Git revisions. Experiment-output checksums are separate reproducibility metadata (Appendix F).

Jina's top-level custom code and configuration use the revision listed in the table. Its backbone weights, task adapters and tokenizer instead use `dd76d535f5447ca3897a9c893fb1e612ead98192`. The custom loader does not automatically propagate the top-level revision to every component. We explicitly bind the component versions actually loaded and check reproduction of cached embeddings. Gemma replay preserves the original batch composition because numerical outputs can depend on padding and batch shape. This replay checks the recorded implementation, rather than certifying that a top-level revision alone locks every dependency.

Table 4: Hugging Face repository revisions, configured input-length limits and numerical precision. Revisions identify the top-level model-loading versions.

Model	Repository commit	Limit	Precision
Gemma	`57c266a740f537b4dc058e1b0cda161fd15afa75`	2,048	bf16
BGE	`5617a9f61b028005a4858fdac845db406aefb181`	8,192	fp16
Arctic	`ac6544c8a46e00af67e330e85a9028c66b8cfd9a`	8,192	fp16
Jina	`d153fb80cb0b1410831a12f4d4795d39d348245c`	32,768	fp16
PPLX	`06456497a00540a582918fe8dcd3a5eabb207772`	32,768	bf16
Qwen	`1d8ad4ca9b3dd8059ad90a75d4983776a23d44af`	32,768	bf16
SFR	`f62d15f411ca97b66acc0f34da2a65f3420b55b0`	4,096	bf16

B Controlled benchmark and native readout evaluation

B.1 Semantic distinctions, scenes and holdouts

Each label identifies a specific binary distinction: success, affirmation that a device is active, an event occurring after the briefing, the doctor being the agent, the box being the inner object, or the doctor having a red badge. Focal terms remain fixed. Replacing site phrases or other fillers preserves the label, not the entire proposition. Table 5 shows two expressions of each positive label.

Table 5: Positive examples, shortened by omitting site phrases. The counterpart changes the target label; each scene also varies a background assignment.

Distinction	Expression A	Expression B
Outcome	The survey succeeded.	The outcome of the survey was a success.
Polarity	The sensor is active.	The statement that the sensor is active is true.
Order	The survey occurred after the briefing.	The event sequence was: first the briefing, then the survey.
Role	The doctor examined the lawyer.	The lawyer received an examination from the doctor.
Containment	The box is inside the crate.	The inner container is the box; the outer container is the crate.
Binding	The doctor has the red badge and the lawyer has the blue badge.	The red badge is assigned to the doctor and the blue badge to the lawyer.

A base scene fixes a site and filler category. Crossing the target label with two background assignments gives four sentences. For binding, the background assignment changes which participant is mentioned first, while the label specifies the doctor's badge color (Table 6). Comparing labels at a fixed background assignment gives two oriented responses per scene.

Table 6: One binding scene under the possession template. The common site phrase is omitted. All four sentences have the same word multiset.

Label	Background	Sentence
0	Doctor first	The doctor has the blue badge and the lawyer has the red badge.
1	Doctor first	The doctor has the red badge and the lawyer has the blue badge.
0	Lawyer first	The lawyer has the red badge and the doctor has the blue badge.
1	Lawyer first	The lawyer has the blue badge and the doctor has the red badge.

The inventory contains 50 primary templates in 20 construction families and seven separate Q/A or numeric stress templates. Families group related grammatical realizations: active/passive forms share one family, and wrappers or punctuation do not create independent families. All family names appear in Table 14; the template registry specifies every realization.

For example, when badge-assignment is held out, fitting uses possession, wearing and attribute families. Assignment templates are evaluated only on the final scenes. This is repeated for every

family. The four test conditions cross familiar versus held-out constructions with known versus held-out fillers, as in Section 2.3.

Table 7: Scene allocation for each distinction. Each scene supplies four sentences. The same 64 final scene identities are reused across templates and vocabulary variants.

Split	Scenes	Use within a fold
Training	72	288 sentences from nonheld families
Validation	24	96 sentences from nonheld families
Final evaluation	64	Every primary template, both vocabularies

Candidate scenes cross 48 site phrases with four filler categories. Training, validation and final scenes are globally disjoint across folds, so a final scene never becomes training data in another fold. A balanced assignment selects one eligible template per training/validation scene, with validation support for every fitting template. Texts are deduplicated by their emitted content. The inventory contains 35,860 unique texts per encoder, including stress tests.

Text-to-fact checks reconstruct labels and referents and verify split membership, family exclusion and text uniqueness. These checks establish consistency with the templates, not unrestricted linguistic validity. Held-out fillers are excluded from supervised semantic fitting, not necessarily ordinary alignment text or encoder pretraining. Fixed focal/function words prevent a token-disjointness claim.

B.2 Readout definitions and fitting

Native readouts use $x_i(t)$ from Section 2.1. For the mean-direction readout, let m_+ and m_- be the positive and negative training means. Its score is $\langle w, x_i(t)\rangle$, where $w = \mathrm{unit}(m_+ - m_-)$, with zero intercept. Directions of norm at most 10^{-12} are undefined. Only the reported log loss standardizes this scalar score by its training mean and standard deviation.

Affine and neural readouts standardize each input coordinate using training-only means and standard deviations; scales below 10^{-12} are replaced by one. The neural pair has two affine layers with 128 hidden units and differs only in the hidden activation. Within a seed, initialization and minibatch ordering match.

Table 8: Native fitting recipes. Validation log loss is weighted equally by family and then by template. Thresholds are selected separately for accuracy.

Readout	Fitting	Selection
Mean direction	Difference of training class means	No fitted direction hyperparameter
Affine	Logistic regression; L-BFGS, tolerance 10^{-8}, at most 10,000 iterations	$C \in \{.01, .1, 1, 10, 100\}$ by validation log loss; smaller C breaks ties
Linear128 / GELU128	Float64 AdamW; learning rate .001, weight decay .0001, batch size 64	Lowest validation log loss, including epoch 0; at most 100 epochs, patience 10

Validation balanced accuracy selects thresholds, breaking ties by proximity to zero and then by the smaller threshold. Seed 42 is primary; seeds 43 and 44 repeat the role/binding neural comparisons. No seed is selected using final results. Word and character controls use training-fitted TF-IDF (word 1–2grams or character 3–5grams, at most 5,000 features). Surface controls use length, case and punctuation.

Transfer fits separate affine $g_{i,c}$ on the centered $z_i(t)$ view, with the same readout label budget. These are distinct from native $h_{i,c}$ on $x_i(t)$. The interface maps B's vectors into A's coordinates without fitting an additional map intercept or calibrating on final target labels (Appendix E.1).

B.3 Metric aggregation and uncertainty

ROC-AUC uses all positive–negative comparisons within a template, with half credit for ties. It is not paired ordering within a scene or accuracy at a fixed threshold. An affine change of score

with positive slope preserves AUC but can change decisions at a previously selected threshold. Aggregation proceeds as follows:

1. Compute each template's metric on the 64 final scenes. Do not pool scores from separately fitted heads.
2. Average templates equally within each family. For construction/joint tests, use the held family; for context/lexical tests, average the nonheld families equally.
3. Average the family-holdout folds equally, then the seven fixed encoders (or 42 directed pairs for transfer).

Accuracy, balanced accuracy, paired ordering and loss use the same hierarchy. Table 1's SD is descriptive dispersion across the seven model-level means, using denominator seven (`ddof=0`); it is not a seed SD or confidence interval.

For each of 10,000 seed-42 bootstrap draws, resample the 64 complete scenes within each distinction. Share scene multiplicities across templates, folds, models, vocabularies and compared readouts, then recompute the entire metric hierarchy and paired difference. Intervals condition on fixed models, families and fitted objects. Neither models nor families are resampled, and the intervals are not adjusted for adaptive design or multiple discoveries. Geometry uses the same scene sampling unit with a different statistic (Appendix D.4).

Prior observations selected the two targeted model-pair contrasts before final readout evaluation. They are not a blind search for the largest gap. Per-family profiles and the complete interval inventory retain the other reported comparisons.

B.4 Native results by readout, encoder and seed

Table 10 compares all four readouts, including thresholded accuracy. The model tables expose heterogeneity behind the cohort means. Binding's matched linear/GELU comparison and the repeated-seed results test the breadth of the activation advantage. Paired intervals are collected once in Table 18.

Table 9: Native affine construction AUC for every encoder and distinction, averaging all held-family folds. These are not worst-family scores.

Encoder	Outcome	Polarity	Order	Role	Containment	Binding
Gemma	1.0000	1.0000	0.7259	0.9099	0.8143	0.6973
BGE-M3	1.0000	1.0000	0.5251	0.8702	0.7603	0.5822
Arctic	1.0000	1.0000	0.6565	0.8173	0.8553	0.5797
Jina	1.0000	1.0000	0.5537	0.7637	0.6636	0.7890
PPLX	1.0000	1.0000	0.5910	0.9134	0.8690	0.9252
Qwen	1.0000	1.0000	0.8907	0.8890	0.9107	0.7223
SFR	1.0000	1.0000	0.9327	0.8499	0.7449	0.6849

Table 10: Fixed-cohort native profiles, seed 42. Ct=context, C=construction, L=lexical, J=joint. Accuracy uses validation-selected thresholds.

Distinction	Readout	AUC Ct	AUC C	AUC L	AUC J	Acc. C	Acc. J
Outcome	Mean	1.0000	1.0000	1.0000	1.0000	0.9724	0.9646
Outcome	Affine	1.0000	1.0000	1.0000	1.0000	0.9364	0.9047
Outcome	Linear128	1.0000	1.0000	1.0000	1.0000	0.9628	0.9379
Outcome	GELU128	1.0000	1.0000	1.0000	1.0000	0.9322	0.9031
Polarity	Mean	1.0000	0.9999	1.0000	1.0000	0.8356	0.8326
Polarity	Affine	1.0000	1.0000	1.0000	1.0000	0.8539	0.8509
Polarity	Linear128	1.0000	0.9999	1.0000	0.9998	0.8783	0.8619
Polarity	GELU128	1.0000	1.0000	1.0000	1.0000	0.8438	0.8298
Order	Mean	0.8304	0.4982	0.8113	0.5084	0.5044	0.5058
Order	Affine	0.9995	0.6965	0.8707	0.6361	0.5656	0.5654
Order	Linear128	0.9992	0.6800	0.8768	0.6296	0.5521	0.5576
Order	GELU128	0.9986	0.5747	0.9433	0.5793	0.5260	0.5303
Role	Mean	0.8002	0.6723	0.8531	0.7654	0.5912	0.6308
Role	Affine	0.9980	0.8591	0.8398	0.7452	0.7153	0.6453
Role	Linear128	0.9948	0.8510	0.8524	0.7665	0.7236	0.6603
Role	GELU128	0.9948	0.8409	0.9100	0.7840	0.7096	0.6660
Containment	Mean	0.8236	0.7203	0.7584	0.6544	0.6106	0.5720
Containment	Affine	0.9908	0.8026	0.7846	0.6329	0.6584	0.5754
Containment	Linear128	0.9833	0.8098	0.7974	0.6561	0.6482	0.5818
Containment	GELU128	0.9903	0.7997	0.8734	0.6850	0.6686	0.6036
Binding	Mean	0.5632	0.5420	0.5680	0.5431	0.5295	0.5303
Binding	Affine	0.8489	0.7115	0.6018	0.5655	0.5727	0.5310
Binding	Linear128	0.7791	0.6666	0.6011	0.5602	0.5606	0.5323
Binding	GELU128	0.9963	0.9380	0.6985	0.6981	0.8771	0.6057

Table 11: Joint-shift affine AUC by encoder. Rounded near-saturation is not always exact perfect discrimination.

Encoder	Outcome	Polarity	Order	Role	Containment	Binding
Gemma	1.0000	1.0000	0.6164	0.7477	0.5817	0.5332
BGE-M3	1.0000	1.0000	0.5592	0.7441	0.5741	0.5271
Arctic	1.0000	0.9999	0.5674	0.5994	0.7645	0.5216
Jina	1.0000	1.0000	0.5389	0.6256	0.5619	0.6039
PPLX	1.0000	1.0000	0.5495	0.8420	0.6444	0.6276
Qwen	1.0000	1.0000	0.8306	0.8784	0.6899	0.5823
SFR	1.0000	1.0000	0.7907	0.7790	0.6134	0.5627

Table 12: Binding by encoder: matched linear/GELU AUC and GELU accuracy, under construction (C) and joint (J) shifts.

Encoder	Linear C	GELU C	Acc. C	Linear J	GELU J	Acc. J
Gemma	0.6283	0.9670	0.9175	0.5310	0.7699	0.6216
BGE-M3	0.5444	0.8128	0.7300	0.5244	0.6402	0.5894
Arctic	0.5244	0.9190	0.8643	0.5115	0.5470	0.5142
Jina	0.7523	0.9881	0.9126	0.5998	0.6849	0.5728
PPLX	0.8794	0.9418	0.8647	0.6241	0.7760	0.6875
Qwen	0.6774	0.9992	0.9771	0.5694	0.7933	0.6636
SFR	0.6598	0.9384	0.8735	0.5614	0.6753	0.5908

Table 13: GELU-minus-linear128 AUC by seed, fixed-cohort means. Replication covers role and binding, not all six distinctions.

Distinction	Seed	Context	Construction	Lexical	Joint
Binding	42	+0.2172	+0.2715	+0.0974	+0.1379
Binding	43	+0.2203	+0.2739	+0.0973	+0.1367
Binding	44	+0.2192	+0.2741	+0.0927	+0.1320
Role	42	-0.0001	-0.0100	+0.0576	+0.0175
Role	43	+0.0003	-0.0068	+0.0553	+0.0167
Role	44	+0.0004	-0.0017	+0.0543	+0.0195

B.5 Construction-family dependence

The held-family profiles show which expressions contribute to each cohort result. Binding's activation gain remains positive in every held-family cohort mean, while the Gemma–Qwen construction difference is concentrated in badge assignment. Their joint-shift differences change sign across families. Narrow scene intervals therefore do not establish an invariant model ordering across grammars. The four aggregate model-pair intervals are in Table 18.

Table 14: Held-family affine and GELU AUC, fixed seven-model means, seed 42. C/J denote construction/joint shifts. Each row uses the head trained without that family; evaluations share final scenes.

Distinction	Held family	Affine C	Affine J	GELU C	GELU J
Outcome	finite-result	1.0000	1.0000	1.0000	1.0000
Outcome	adjectival-result	1.0000	1.0000	1.0000	1.0000
Outcome	nominal-result	1.0000	1.0000	1.0000	1.0000
Polarity	finite-negation	1.0000	1.0000	1.0000	1.0000
Polarity	truth-predication	1.0000	1.0000	1.0000	0.9999
Order	temporal-comparison	0.7320	0.6940	0.6451	0.6561
Order	ordinal-sequence	0.8181	0.6717	0.6952	0.6529
Order	perfect-aspect	0.5395	0.5426	0.3837	0.4290
Role	transitive-voice	0.9130	0.7871	0.8174	0.7770
Role	event-nominal	0.9781	0.7964	0.9606	0.8831
Role	event-receipt	0.8905	0.7695	0.9159	0.8297
Role	participant-nouns	0.6547	0.6276	0.6699	0.6464
Containment	locative-nesting	0.9855	0.7333	0.9778	0.8382
Containment	transitive-containment	0.9165	0.6889	0.9186	0.7506
Containment	container-roles	0.5667	0.5161	0.5814	0.5197
Containment	contents-membership	0.7417	0.5930	0.7209	0.6314
Binding	badge-possession	0.7789	0.5720	0.9983	0.6803
Binding	badge-wearing	0.7376	0.5775	0.9433	0.7092
Binding	badge-assignment	0.6839	0.5592	0.8721	0.7198
Binding	badge-attribute	0.6457	0.5532	0.9384	0.6832

Table 15: Binding family-level point differences. Activation deltas average seven encoders; Gemma–Qwen compares GELU heads. C/J denote construction/joint shifts.

Held binding family	Δ act. C	Δ act. J	Gemma–Qwen C	Gemma–Qwen J
badge-possession	+0.2682	+0.1106	+0.0000	-0.0874
badge-wearing	+0.2621	+0.1402	+0.0000	-0.1458
badge-assignment	+0.2349	+0.1684	-0.1287	+0.0989
badge-attribute	+0.3207	+0.1322	+0.0000	+0.0407

B.6 LEXICAL CONTROLS AND SEPARATE STRESS TESTS

Character features nearly saturate outcome, while word/character construction-shift binding AUC is 0.5. These controls qualify semantic interpretations of high scores without exhausting possible shortcuts. The constant majority score has AUC 0.5 in every condition and is omitted from the repeated table columns.

Table 16: Lexical-control AUC. Surface denotes length/case/punctuation features; word and character transforms are training-fitted TF-IDF. Scores below 0.5 are not reoriented using final labels.

Distinction	Condition	Surface	Word	Character
Outcome	context	0.5900	1.0000	1.0000
Outcome	construction	0.3469	0.5000	0.9999
Outcome	lexical	0.5904	1.0000	1.0000
Outcome	joint	0.3463	0.5000	0.9855
Polarity	context	0.8239	1.0000	1.0000
Polarity	construction	0.3246	0.2620	0.7311
Polarity	lexical	0.8093	1.0000	1.0000
Polarity	joint	0.3132	0.2500	0.6663
Order	context	0.5000	0.8323	0.8335
Order	construction	0.5000	0.5000	0.5420
Order	lexical	0.5000	0.8333	0.8345
Order	joint	0.5000	0.5000	0.4886
Role	context	0.5000	0.9366	0.9316
Role	construction	0.5000	0.4408	0.4353
Role	lexical	0.5000	0.8458	0.8788
Role	joint	0.5000	0.5125	0.3686
Containment	context	0.5000	0.9958	0.9997
Containment	construction	0.5000	0.5552	0.4165
Containment	lexical	0.5000	1.0000	0.9968
Containment	joint	0.5000	0.5417	0.4269
Binding	context	0.5000	0.5625	0.5625
Binding	construction	0.5000	0.5000	0.5000
Binding	lexical	0.5000	0.5625	0.5566
Binding	joint	0.5000	0.5000	0.5000

Stress tests. Q/A and clock expressions introduce additional demands and do not enter the primary family means or native contrasts. Q/A can expose answer-token cues, while clock order requires operand comparison. Affine clock AUC is 0.4467/0.4606 for known/held-out vocabulary, and containment Q/A is also below chance. These failures remain reported separately rather than being counted as independent primary families.

Table 17: Separate Q/A and clock tests. Encoder scores average the fixed fold heads and then encoders; lexical controls average their fixed fold heads.

Panel	Vocabulary	Affine	Linear128	GELU128	Word	Character
Outcome Q/A	known	1.0000	0.9990	1.0000	0.5000	0.5000
Outcome Q/A	reserved	0.9981	0.9958	0.9967	0.5000	0.5000
Polarity Q/A	known	0.9961	0.9821	0.9952	0.5000	0.5000
Polarity Q/A	reserved	0.9841	0.9772	0.9830	0.5000	0.5000
Order Q/A	known	0.8003	0.7762	0.7221	0.6677	0.9794
Order Q/A	reserved	0.6795	0.6744	0.6756	0.6667	0.9221
Order clock	known	0.4467	0.4458	0.4538	0.5000	0.5000
Order clock	reserved	0.4606	0.4597	0.4647	0.5000	0.5000
Role Q/A	known	0.9370	0.9558	0.9250	0.0896	0.3133
Role Q/A	reserved	0.8105	0.8392	0.8534	0.0000	0.5017
Containment Q/A	known	0.3289	0.3100	0.2999	0.1227	0.2510
Containment Q/A	reserved	0.3998	0.3720	0.3546	0.2500	0.2259
Binding Q/A	known	0.6481	0.6120	0.8480	0.5000	0.5000
Binding Q/A	reserved	0.5678	0.5597	0.7659	0.5000	0.5000

B.7 Complete native-readout contrasts

The 24 contrasts cover 12 affine shift penalties, eight activation effects and four targeted model comparisons. They are grouped by question below. Rounding can conceal very small effects; full-precision values are retained in the source manifest. No interval is recomputed for presentation.

Table 18: All 24 native AUC contrasts, with conditional 95% scene intervals.

Distinction	Measurement or contrast	Estimate	Lower	Upper
Expression shifts				
Outcome	Affine construction minus context	+0.00000	+0.00000	+0.00000
Outcome	Affine joint minus context	+0.00000	+0.00000	+0.00000
Polarity	Affine construction minus context	+0.00000	+0.00000	+0.00000
Polarity	Affine joint minus context	-0.00002	-0.00006	+0.00000
Order	Affine construction minus context	-0.30296	-0.30927	-0.29540
Order	Affine joint minus context	-0.36341	-0.36900	-0.35619
Role	Affine construction minus context	-0.13894	-0.14506	-0.13181
Role	Affine joint minus context	-0.25283	-0.25861	-0.24319
Containment	Affine construction minus context	-0.18824	-0.19413	-0.18220
Containment	Affine joint minus context	-0.35797	-0.36487	-0.34829
Binding	Affine construction minus context	-0.13733	-0.14104	-0.13317
Binding	Affine joint minus context	-0.28339	-0.28920	-0.27399
Activation effects				
Role	GELU minus linear: context	-0.00010	-0.00073	+0.00053
Role	GELU minus linear: construction	-0.01003	-0.01364	-0.00646
Role	GELU minus linear: lexical	+0.05756	+0.05366	+0.06063
Role	GELU minus linear: joint	+0.01754	+0.01280	+0.02149
Binding	GELU minus linear: context	+0.21715	+0.21033	+0.22090
Binding	GELU minus linear: construction	+0.27148	+0.26483	+0.27588
Binding	GELU minus linear: lexical	+0.09740	+0.08554	+0.10323
Binding	GELU minus linear: joint	+0.13787	+0.12886	+0.14373
Targeted model comparisons				
Order	Qwen–Jina affine: construction	+0.33696	+0.30658	+0.36793
Order	Qwen–Jina affine: joint	+0.29170	+0.26632	+0.31609
Binding	Gemma–Qwen GELU: construction	-0.03218	-0.03368	-0.03074
Binding	Gemma–Qwen GELU: joint	-0.02340	-0.04653	+0.00086

C MULTINLI READOUT EVALUATION

C.1 DATASET SAMPLING AND OVERLAP CONTROLS

We use MultiNLI1.0's training set and both labeled public-dev splits. The latter contain 10,000 rows each before excluding 185 matched and 168 mismatched rows with unknown labels, leaving 9,815 and 9,832 evaluation pairs. Training and validation sampling follows these steps:

1. Normalize sentence text by NFKC, case folding and whitespace, and group training rows by normalized premise.
2. Exclude any such group containing a premise or hypothesis found in either complete raw dev split, including unknown-label rows.
3. Select validation premise groups first in deterministic seed-42 order, with a cap of 400 pairs per genre. Remove all remaining groups touching a selected validation sentence, then select training groups with separate deterministic ordering and a cap of 2,000 pairs per genre.

The resulting sets contain 9,994 training and 2,000 validation pairs with no direct normalized sentence overlap with evaluation. This exclusion is propagated within premise groups, not recursively through every shared hypothesis; it does not certify underlying-document independence. The 42,191 unique sentences per encoder use the recipes in Appendix A, without encoder fitting or PCA. Dataset notices and exact row identities are retained with the data records.

C.2 FEATURES, FITTING AND SUPPORTING RESULTS

Let p, h denote the normalized premise and hypothesis embeddings. Feature families test the information supplied to the readout, not a causal intervention on reasoning:

Feature family	Input
Premise / hypothesis only	p / h
Additive	$[p, h]$
Symmetric pair	$[p + h, \lvert p - h \rvert, p \odot h]$
Full pair	$[p, h, \lvert p - h \rvert, p \odot h]$

Absolute differences and products are nonlinear in the individual sentence vectors, even when the following classifier is affine. The neural comparison uses the full features in both conditions.

Table 19: MultiNLI fitting recipes. All feature standardization uses training data; validation selects the readout before either dev split is scored.

Readout	Optimization	Selection
Affine softmax	Float64 L-BFGS; at most 3,000 iterations; gradient tolerance 10^{-7}, function tolerance 10^{-12}	$C \in \{.01, .1, 1, 10\}$ by validation macro-F1; smaller C breaks ties
Linear128 / GELU128	Input–128–3; Adam, learning rate .001, weight decay .0001, batch size 128	Validation macro-F1; at most 50 epochs, patience 8; epoch 0 eligible, earlier epochs break ties

The affine objective is mean cross entropy plus $\|W\|_F^2 / (2CN)$, with unpenalized intercept. Matched neural heads share seed-specific initialization and minibatch-order prefixes. Seed 42 is primary and 43/44 are fixed repetitions. Their fitting procedure differs from directly fitted affine softmax. Word unigram/ bigram TF-IDF controls use a training vocabulary capped at 20,000 features and the same C grid and label budget, with premise, hypothesis or role-tagged pair text.

Table 20: Fixed-cohort MultiNLI macro-F1. Feature capacities differ, so the single-input and symmetric references do not isolate a causal directionality effect.

Readout / features	Matched	Mismatched
Premise affine	0.3338	0.3291
Hypothesis affine	0.5175	0.5266
Directed additive affine	0.5303	0.5372
Symmetric pair affine	0.5897	0.5994
Full directed affine	0.6245	0.6334
Linear128 (seed 42)	0.6582	0.6695
GELU128 (seed 42)	0.6654	0.6748

Table 21: Lexical-control macro-F1. Each row is one TF-IDF system, not a seven-model mean.

TF-IDF features	Matched	Mismatched
Premise	0.3301	0.3285
Hypothesis	0.4814	0.4815
Role-tagged pair	0.4879	0.4854

Table 22: Per-encoder neural macro-F1 at the primary seed. M/MM denote matched/ mismatched genres. Negative and near-zero activation effects remain visible.

Encoder	Linear M	GELU M	Linear MM	GELU MM
Gemma	0.6730	0.6880	0.6898	0.7015
BGE-M3	0.6804	0.6768	0.6849	0.6810
Arctic	0.6049	0.6192	0.6205	0.6346
Jina	0.6679	0.6771	0.6825	0.6794
PPLX	0.6117	0.6161	0.6196	0.6250
Qwen	0.6916	0.6989	0.6960	0.7095
SFR	0.6781	0.6819	0.6930	0.6930

Table 23: All fixed neural seeds, equal seven-model means. No seed is selected using final performance.

Seed	Linear M	GELU M	Linear MM	GELU MM
42	0.6582	0.6654	0.6695	0.6748
43	0.6550	0.6655	0.6666	0.6727
44	0.6575	0.6646	0.6689	0.6755

C.3 Paired uncertainty and complete cohort contrasts

For uncertainty estimation, evaluation pairs sharing a sentence or publisher premise identity form 6,655 connected groups, including two spanning both dev splits. The 10,000 bootstrap draws resample complete groups with shared multiplicities across models, conditions and splits. This preserves the observed pair dependencies; it is distinct from sampling the controlled benchmark's scenes. The cross-split gain difference uses the joint draws.

Table 24: Complete fixed-cohort macro-F1 contrasts with conditional 95% intervals. Neural contrasts use seed 42. A zero-crossing cross-split interval does not establish equivalent gains.

Contrast	Split	Delta	Lower	Upper
GELU minus linear128	Matched	+0.0072	+0.0044	+0.0100
Full minus premise affine	Matched	+0.2907	+0.2844	+0.2969
Full minus hypothesis affine	Matched	+0.1069	+0.1008	+0.1132
Full minus additive affine	Matched	+0.0942	+0.0886	+0.1001
Full minus symmetric affine	Matched	+0.0347	+0.0314	+0.0380
GELU minus linear128	Mismatched	+0.0054	+0.0028	+0.0081
Full minus premise affine	Mismatched	+0.3043	+0.2984	+0.3103
Full minus hypothesis affine	Mismatched	+0.1068	+0.1007	+0.1128
Full minus additive affine	Mismatched	+0.0962	+0.0905	+0.1017
Full minus symmetric affine	Mismatched	+0.0340	+0.0307	+0.0374
GELU gain: MM minus M	Joint	-0.0018	-0.0057	+0.0020

The mean activation gain is modest and varies by encoder, while all three seeds retain positive cohort effects. This supports readout dependence on a separate natural task. MultiNLI does not label the six controlled distinctions or establish a binding-specific mechanism, and AUC/F1 differences are not a matched cross-task effect-size comparison. Computational checks are summarized in Appendix F.

D Response geometry and permutation references

D.1 Responses and cross-model comparison

For each template and vocabulary, the 64 final scenes supply 128 oriented responses $\Delta_{i,c,k}$ from Section 2.1. Full responses use normalized native embeddings x_i. Residual responses subtract the encoder's unnormalized mean over the 144 paired responses in the other-family readout-training set. This subtraction uses neither the unit mean-direction classifier weight nor the mean of the 32 centered contrasts used to fit maps (Appendix E.1). It does not identify a unique shared semantic component.

We normalize each nonzero response and form its within-encoder cosine matrix in float64. We symmetrize the matrix, clip to $[-1, 1]$, and set valid diagonals to one. Spearman correlation compares corresponding strict upper triangles, preserving the exact serialized ties without post-hoc rounding. Correlations are calculated per template before equal averaging within families, across families, and across the fixed 21 encoder pairs. No cross-model map is fitted for this measurement.

Full native responses do not depend on a fitted readout or its training fold. Exhaustive equal-family averages over familiar/held constructions would therefore coincide for known vocabulary, and likewise for held-out vocabulary. These aliases are not separate generalization confirmations.

D.2 Permutation nulls and lexical references

We compare observed correspondence with two scene-permutation references, each using 1,000 seed-42-derived draws per distinction:

1. The within-template null permutes all 64 scenes. Both responses belonging to a scene move together.
2. The blocked null permits exchanges only within each of the four filler categories, with 16 final scenes per category. Site phrases are not blocking keys.

Gemma anchors the scene ordering. Every other encoder or lexical view has its own permutation, reused across pairs, templates, vocabularies and response types. Thus each cohort-null draw represents one coherent set of view permutations, rather than independent randomizations for each model pair. The blocked null preserves filler groupings that may contain meaningful structure. Its excess tests correspondence beyond that grouping, not a semantic-information fraction.

Lexical references use training-fitted word 1–2gram and character 3–5gram TF-IDF, or five length/case/punctuation features. Feature rows are L2-normalized; classifier weights and affine-readout standardization do not enter the geometry. Residuals subtract each lexical view's own readout-training mean.

Table 25: Availability of template-level geometry measurements across both response types and vocabularies. The zero-norm threshold is 10^{-12}.

Reference	Defined	Zero endpoint/response	Constant correlation input
Encoder pairs	4,200	0	0
Lexical–encoder pairs	1,666	1,610	924

A fixed-family aggregate is unavailable if any required member is undefined. Consequently, an unavailable aggregate does not mean every underlying measurement is undefined. Available-member descriptions have different denominators and remain separate in the artifacts. Undefined correlations are neither zero nor evidence of neural superiority over lexical features.

D.3 Complete geometry profiles

The next tables report both references, family dependence and every native model pair. Pair exceedance counts compare each pair with its own null's 97.5th percentile; the pairs share encoders and are not independent discoveries. The lexical table lists defined fixed-family aggregates and consolidates repeated unavailable panels, while the source manifest retains every individual cell.

Table 26: Full and residual geometry against the template-plus-filler-block null. Excess is observed minus cohort-null median; pair counts are dependent descriptions.

Distinction	Vocabulary	Response	Observed	Null median	Excess	Pairs /21
Outcome	Known	Full	0.6206	0.5284	+0.0922	21
Outcome	Known	Residual	0.6244	0.5344	+0.0900	21
Outcome	Held-out	Full	0.5519	0.4678	+0.0841	21
Outcome	Held-out	Residual	0.6072	0.5226	+0.0846	21
Polarity	Known	Full	0.5680	0.4695	+0.0985	21
Polarity	Known	Residual	0.5633	0.4661	+0.0972	21
Polarity	Held-out	Full	0.5462	0.3923	+0.1538	21
Polarity	Held-out	Residual	0.5443	0.3925	+0.1518	21
Order	Known	Full	0.5513	0.5216	+0.0297	21
Order	Known	Residual	0.5325	0.5027	+0.0299	21
Order	Held-out	Full	0.5595	0.5259	+0.0335	21
Order	Held-out	Residual	0.5410	0.5078	+0.0331	21
Role	Known	Full	0.6769	0.6642	+0.0127	21
Role	Known	Residual	0.6564	0.6415	+0.0148	19
Role	Held-out	Full	0.7175	0.7073	+0.0103	20
Role	Held-out	Residual	0.7076	0.6982	+0.0095	21
Containment	Known	Full	0.4872	0.4739	+0.0133	17
Containment	Known	Residual	0.4806	0.4662	+0.0144	18
Containment	Held-out	Full	0.5250	0.5145	+0.0105	19
Containment	Held-out	Residual	0.4987	0.4867	+0.0120	17
Binding	Known	Full	0.8367	0.7971	+0.0397	21
Binding	Known	Residual	0.8356	0.7954	+0.0402	21
Binding	Held-out	Full	0.8210	0.7863	+0.0347	21
Binding	Held-out	Residual	0.8210	0.7858	+0.0353	21

Table 27: The within-template scene null without filler blocking, using the same observed geometry and aggregation.

Distinction	Vocabulary	Response	Observed	Null median	Excess	Pairs /21
Outcome	Known	Full	0.6206	0.0294	+0.5912	21
Outcome	Known	Residual	0.6244	0.0249	+0.5996	21
Outcome	Held-out	Full	0.5519	0.0229	+0.5289	21
Outcome	Held-out	Residual	0.6072	0.0227	+0.5845	21
Polarity	Known	Full	0.5680	0.0751	+0.4929	21
Polarity	Known	Residual	0.5633	0.0779	+0.4854	21
Polarity	Held-out	Full	0.5462	0.0933	+0.4529	21
Polarity	Held-out	Residual	0.5443	0.0865	+0.4578	21
Order	Known	Full	0.5513	0.0252	+0.5261	21
Order	Known	Residual	0.5325	0.0257	+0.5068	21
Order	Held-out	Full	0.5595	0.0338	+0.5257	21
Order	Held-out	Residual	0.5410	0.0289	+0.5121	21
Role	Known	Full	0.6769	0.1153	+0.5616	21
Role	Known	Residual	0.6564	0.1214	+0.5350	21
Role	Held-out	Full	0.7175	0.0959	+0.6217	21
Role	Held-out	Residual	0.7076	0.0998	+0.6078	21
Containment	Known	Full	0.4872	0.0377	+0.4495	21
Containment	Known	Residual	0.4806	0.0387	+0.4419	21
Containment	Held-out	Full	0.5250	0.0265	+0.4985	21
Containment	Held-out	Residual	0.4987	0.0273	+0.4713	21
Binding	Known	Full	0.8367	0.7549	+0.0818	21
Binding	Known	Residual	0.8356	0.7540	+0.0816	21
Binding	Held-out	Full	0.8210	0.7547	+0.0663	21
Binding	Held-out	Residual	0.8210	0.7539	+0.0671	21

Table 28: Geometry by construction family, fixed 21-pair means. Known/held denote known/held-out vocabulary; full and residual responses remain separate.

Distinction	Held family	Known Full	Known Residual	Held-out Full	Held-out Residual
Outcome	finite-result	0.5871	0.6353	0.5269	0.6246
Outcome	adjectival-result	0.6860	0.6854	0.6015	0.6426
Outcome	nominal-result	0.5888	0.5526	0.5272	0.5544
Polarity	finite-negation	0.5537	0.5582	0.5465	0.5589
Polarity	truth-predication	0.5822	0.5684	0.5459	0.5297
Order	temporal-comparison	0.5178	0.5216	0.6060	0.6077
Order	ordinal-sequence	0.5611	0.4982	0.4801	0.4458
Order	perfect-aspect	0.5750	0.5779	0.5923	0.5694
Role	transitive-voice	0.7446	0.7068	0.7378	0.7337
Role	event-nominal	0.7052	0.6607	0.7747	0.7893
Role	event-receipt	0.6525	0.6575	0.6439	0.6353
Role	participant-nouns	0.6052	0.6004	0.7138	0.6723
Containment	locative-nesting	0.5301	0.5307	0.5753	0.5705
Containment	transitive-containment	0.5201	0.5060	0.5582	0.5415
Containment	container-roles	0.4214	0.4118	0.4181	0.3810
Containment	contents-membership	0.4771	0.4740	0.5484	0.5017
Binding	badge-possession	0.8327	0.8335	0.8169	0.8193
Binding	badge-wearing	0.8347	0.8335	0.8201	0.8187
Binding	badge-assignment	0.8430	0.8419	0.8157	0.8160
Binding	badge-attribute	0.8365	0.8335	0.8313	0.8301

Table 29: All native model-pair response correlations, grouped by vocabulary and response type. Every pair is retained in each panel.

Pair	Panel	Outcome	Polarity	Order	Role	Containment	Binding
Known / full							
Gemma–BGE-M3	Known / full	0.7328	0.6499	0.6962	0.6666	0.5857	0.8427
Gemma–Arctic	Known / full	0.6140	0.3754	0.5216	0.6524	0.5371	0.8395
Gemma–Jina	Known / full	0.6965	0.6180	0.5650	0.6976	0.5195	0.8250
Gemma–PPLX	Known / full	0.4979	0.3566	0.5476	0.7155	0.4529	0.8043
Gemma–Qwen	Known / full	0.7912	0.6931	0.6930	0.6488	0.4708	0.8223
Gemma–SFR	Known / full	0.7157	0.6624	0.6060	0.5716	0.4877	0.8338
BGE-M3–Arctic	Known / full	0.5710	0.5600	0.6553	0.7439	0.6230	0.8840
BGE-M3–Jina	Known / full	0.6723	0.6910	0.5735	0.7448	0.4198	0.8496
BGE-M3–PPLX	Known / full	0.4956	0.4349	0.5155	0.7534	0.4537	0.8177
BGE-M3–Qwen	Known / full	0.6462	0.7379	0.6572	0.7047	0.3928	0.8543
BGE-M3–SFR	Known / full	0.6126	0.7310	0.5151	0.6101	0.5209	0.8545
Arctic–Jina	Known / full	0.5966	0.4457	0.4681	0.6577	0.3907	0.8540
Arctic–PPLX	Known / full	0.5224	0.4219	0.4433	0.7070	0.4665	0.8073
Arctic–Qwen	Known / full	0.5755	0.4514	0.5250	0.6431	0.4108	0.8354
Arctic–SFR	Known / full	0.5715	0.4586	0.4108	0.5877	0.4832	0.8540
Jina–PPLX	Known / full	0.5105	0.4274	0.4175	0.7249	0.4302	0.8115
Jina–Qwen	Known / full	0.7049	0.7743	0.5675	0.7175	0.5542	0.8467
Jina–SFR	Known / full	0.6836	0.7211	0.4920	0.6166	0.5425	0.8665
PPLX–Qwen	Known / full	0.5484	0.4829	0.5334	0.7161	0.4307	0.7963
PPLX–SFR	Known / full	0.4994	0.4518	0.5445	0.6416	0.4779	0.7976
Qwen–SFR	Known / full	0.7744	0.7828	0.6294	0.6931	0.5797	0.8741
Known / residual							
Gemma–BGE-M3	Known / residual	0.7219	0.6305	0.6336	0.6404	0.5571	0.8438
Gemma–Arctic	Known / residual	0.5972	0.3922	0.5488	0.6403	0.4530	0.8383
Gemma–Jina	Known / residual	0.6886	0.5953	0.5528	0.7103	0.5523	0.8267
Gemma–PPLX	Known / residual	0.5289	0.3681	0.5252	0.7024	0.4279	0.8041
Gemma–Qwen	Known / residual	0.7971	0.6902	0.6417	0.6327	0.5129	0.8191
Gemma–SFR	Known / residual	0.7116	0.6514	0.5932	0.5597	0.4907	0.8310
BGE-M3–Arctic	Known / residual	0.5843	0.5434	0.6187	0.7172	0.5733	0.8861
BGE-M3–Jina	Known / residual	0.6803	0.6745	0.5394	0.7170	0.4618	0.8533
BGE-M3–PPLX	Known / residual	0.5743	0.4371	0.4809	0.7323	0.4372	0.8240
BGE-M3–Qwen	Known / residual	0.6468	0.7310	0.6412	0.6497	0.4143	0.8480
BGE-M3–SFR	Known / residual	0.6185	0.7084	0.5021	0.5763	0.5276	0.8524
Arctic–Jina	Known / residual	0.5817	0.4407	0.4487	0.6499	0.4059	0.8506
Arctic–PPLX	Known / residual	0.5376	0.4088	0.4389	0.6753	0.4642	0.8056
Arctic–Qwen	Known / residual	0.5779	0.4550	0.5123	0.6110	0.3690	0.8342
Arctic–SFR	Known / residual	0.5189	0.4798	0.4254	0.5395	0.4853	0.8522
Jina–PPLX	Known / residual	0.5501	0.4518	0.4210	0.7323	0.3844	0.8072
Jina–Qwen	Known / residual	0.6943	0.7580	0.5485	0.7139	0.5813	0.8407
Jina–SFR	Known / residual	0.6536	0.6904	0.5115	0.6122	0.5641	0.8644
PPLX–Qwen	Known / residual	0.5683	0.5041	0.4978	0.6857	0.3972	0.7975
PPLX–SFR	Known / residual	0.5160	0.4632	0.5222	0.5933	0.4562	0.7988
Qwen–SFR	Known / residual	0.7651	0.7554	0.5799	0.6918	0.5775	0.8698
Held-out / full							
Gemma–BGE-M3	Held-out / full	0.6855	0.6639	0.6736	0.7872	0.6173	0.8344
Gemma–Arctic	Held-out / full	0.5603	0.4552	0.6043	0.7620	0.6595	0.8175
Gemma–Jina	Held-out / full	0.6359	0.6278	0.5985	0.7032	0.5105	0.8205
Gemma–PPLX	Held-out / full	0.4954	0.4726	0.5260	0.8107	0.5361	0.8088
Gemma–Qwen	Held-out / full	0.6496	0.5440	0.6381	0.7982	0.5161	0.8184
Gemma–SFR	Held-out / full	0.6403	0.6195	0.5591	0.6778	0.4787	0.8214
BGE-M3–Arctic	Held-out / full	0.5181	0.5989	0.5440	0.7909	0.6387	0.8303

Pair	Panel	Outcome	Polarity	Order	Role	Containment	Binding
BGE-M3–Jina	Held-out / full	0.5864	0.6221	0.5121	0.7365	0.4531	0.8215
BGE-M3–PPLX	Held-out / full	0.4197	0.5294	0.4628	0.7143	0.5471	0.7849
BGE-M3–Qwen	Held-out / full	0.5699	0.5369	0.5340	0.7525	0.4863	0.8319
BGE-M3–SFR	Held-out / full	0.5130	0.6520	0.4987	0.5830	0.4565	0.8216
Arctic–Jina	Held-out / full	0.5060	0.5168	0.5522	0.7416	0.5213	0.8246
Arctic–PPLX	Held-out / full	0.4859	0.4753	0.5689	0.6809	0.6170	0.7946
Arctic–Qwen	Held-out / full	0.4070	0.4515	0.5756	0.7383	0.4572	0.8256
Arctic–SFR	Held-out / full	0.4195	0.5241	0.5536	0.6336	0.4356	0.8209
Jina–PPLX	Held-out / full	0.4597	0.4990	0.4838	0.6945	0.5325	0.8093
Jina–Qwen	Held-out / full	0.7267	0.5419	0.5747	0.7391	0.5045	0.8363
Jina–SFR	Held-out / full	0.6586	0.6003	0.5196	0.6466	0.4665	0.8369
PPLX–Qwen	Held-out / full	0.4354	0.4198	0.5640	0.7328	0.4960	0.8126
PPLX–SFR	Held-out / full	0.4421	0.5011	0.5475	0.6139	0.4584	0.8150
Qwen–SFR	Held-out / full	0.7746	0.6177	0.6574	0.7305	0.6357	0.8537
Held-out / residual							
Gemma–BGE-M3	Held-out / residual	0.7217	0.6529	0.6313	0.7672	0.5823	0.8364
Gemma–Arctic	Held-out / residual	0.6371	0.4467	0.5365	0.7362	0.5946	0.8207
Gemma–Jina	Held-out / residual	0.6486	0.6252	0.5374	0.7268	0.5194	0.8223
Gemma–PPLX	Held-out / residual	0.6088	0.4676	0.4977	0.7984	0.4928	0.8097
Gemma–Qwen	Held-out / residual	0.7089	0.5726	0.6383	0.7913	0.5341	0.8209
Gemma–SFR	Held-out / residual	0.6735	0.5903	0.5590	0.6613	0.5188	0.8211
BGE-M3–Arctic	Held-out / residual	0.5841	0.5834	0.5245	0.7831	0.6323	0.8277
BGE-M3–Jina	Held-out / residual	0.5953	0.6177	0.5295	0.7585	0.4489	0.8201
BGE-M3–PPLX	Held-out / residual	0.5507	0.5148	0.4372	0.6974	0.4563	0.7883
BGE-M3–Qwen	Held-out / residual	0.6160	0.5635	0.5623	0.7136	0.4731	0.8274
BGE-M3–SFR	Held-out / residual	0.5939	0.6344	0.4957	0.5597	0.4324	0.8218
Arctic–Jina	Held-out / residual	0.5125	0.5138	0.5177	0.7536	0.4951	0.8230
Arctic–PPLX	Held-out / residual	0.5643	0.4765	0.5596	0.6580	0.5481	0.7983
Arctic–Qwen	Held-out / residual	0.5060	0.4688	0.5669	0.7045	0.4186	0.8240
Arctic–SFR	Held-out / residual	0.4721	0.4993	0.5409	0.5981	0.4291	0.8230
Jina–PPLX	Held-out / residual	0.5324	0.4898	0.4697	0.7140	0.4732	0.8075
Jina–Qwen	Held-out / residual	0.6813	0.5693	0.5357	0.7495	0.4531	0.8335
Jina–SFR	Held-out / residual	0.6707	0.5771	0.4980	0.6454	0.4248	0.8330
PPLX–Qwen	Held-out / residual	0.5384	0.4527	0.5470	0.7183	0.4654	0.8153
PPLX–SFR	Held-out / residual	0.5405	0.5129	0.5529	0.6126	0.4431	0.8146
Qwen–SFR	Held-out / residual	0.7943	0.6009	0.6225	0.7131	0.6364	0.8531

Table 30: Lexical response correlations for defined fixed-family aggregates, followed by all unavailable panels. An unavailable aggregate has at least one undefined required member; it is not zero. Panels cross known/held-out vocabulary with full/residual responses.

Distinction	Panel	View	Gemma	BGE-M3	Arctic	Jina	PPLX	Qwen	SFR
Outcome	Known / full	Char	0.281	0.298	0.270	0.340	0.260	0.283	0.322
Outcome	Known / residual	Char	0.085	0.104	0.119	0.124	0.117	0.147	0.193
Outcome	Held-out / full	Char	0.205	0.303	0.310	0.285	0.214	0.203	0.210
Outcome	Held-out / residual	Char	0.093	0.093	0.165	0.155	0.187	0.140	0.177
Polarity	Known / full	Surface	0.088	0.085	0.066	0.133	0.023	0.081	0.078
Polarity	Known / residual	Surface	0.092	0.085	0.065	0.130	0.030	0.086	0.081
Polarity	Held-out / full	Surface	0.044	0.160	0.167	0.106	0.144	0.088	0.186
Polarity	Held-out / residual	Surface	0.057	0.135	0.168	0.107	0.146	0.095	0.170

Distinction	Panel	View	Gemma	BGE-M3	Arctic	Jina	PPLX	Qwen	SFR
Unavailable aggregates: all seven encoder comparisons									
Outcome	Word, Surface: all four panels								
Polarity	Word, Char: all four panels								
Order	Word, Char, Surface: all four panels								
Role	Word, Char, Surface: all four panels								
Containment	Word, Char, Surface: all four panels								
Binding	Word, Char, Surface: all four panels								

D.4 Geometry resampling and complete intervals

The bootstrap samples complete scenes as in Appendix B.3, then recomputes the cosine-matrix rank correlation on the expanded sample. Repeated scenes produce repeated responses. Comparisons between distinct copies of the same response add diagonal contributions, and weighted midranks reproduce this expanded tied sample. An interval requires a defined point estimate and at least 90% defined draws.

This estimator exhibits finite-sample bootstrap bias. For example, known-vocabulary full binding has point estimate 0.83673 and percentile interval [0.83807,0.84863]. A percentile interval need not contain the original estimate. Independent direct expansion verifies the computation, not its repeated-sampling coverage; we report the declared procedure without recentering its intervals after seeing results. Intervals condition on fixed encoders, families and fits, and individual intervals are not multiplicity-adjusted discoveries.

All 48 geometry intervals follow: 24 observed means, 12 residual-minus-full differences and 12 held-out-minus-known vocabulary differences. They are grouped by question, with full-precision values retained in the source manifest.

Table 31: All 48 response-geometry measurements and contrasts, with conditional 95% percentile scene intervals.

Distinction	Measurement or contrast	Estimate	Lower	Upper
Observed correspondence				
Outcome	Observed: known, full	+0.62062	+0.61105	+0.65574
Outcome	Observed: known, residual	+0.62443	+0.61253	+0.66653
Outcome	Observed: held-out, full	+0.55188	+0.53368	+0.60352
Outcome	Observed: held-out, residual	+0.60718	+0.58738	+0.65757
Polarity	Observed: known, full	+0.56800	+0.54686	+0.61426
Polarity	Observed: known, residual	+0.56329	+0.54251	+0.61013
Polarity	Observed: held-out, full	+0.54618	+0.53532	+0.58789
Polarity	Observed: held-out, residual	+0.54430	+0.53400	+0.58490
Order	Observed: known, full	+0.55132	+0.54503	+0.58506
Order	Observed: known, residual	+0.53255	+0.52612	+0.56735
Order	Observed: held-out, full	+0.55946	+0.54369	+0.60035
Order	Observed: held-out, residual	+0.54097	+0.52721	+0.58312
Role	Observed: known, full	+0.67689	+0.67658	+0.71424
Role	Observed: known, residual	+0.65636	+0.65717	+0.69446
Role	Observed: held-out, full	+0.71754	+0.70785	+0.75516
Role	Observed: held-out, residual	+0.70764	+0.70024	+0.74461
Containment	Observed: known, full	+0.48716	+0.46956	+0.54157
Containment	Observed: known, residual	+0.48063	+0.46527	+0.53290
Containment	Observed: held-out, full	+0.52498	+0.52285	+0.55653
Containment	Observed: held-out, residual	+0.49867	+0.49650	+0.53140
Binding	Observed: known, full	+0.83673	+0.83807	+0.84863
Binding	Observed: known, residual	+0.83560	+0.83700	+0.84749
Binding	Observed: held-out, full	+0.82099	+0.82284	+0.83335

Distinction	Measurement or contrast	Estimate	Lower	Upper
Binding	Observed: held-out, residual	+0.82104	+0.82296	+0.83333
Training-mean subtraction				
Outcome	Residual minus full: known	+0.00381	-0.01018	+0.01980
Outcome	Residual minus full: held-out	+0.05531	+0.04708	+0.05990
Polarity	Residual minus full: known	-0.00471	-0.00982	+0.00075
Polarity	Residual minus full: held-out	-0.00189	-0.00827	+0.00438
Order	Residual minus full: known	-0.01877	-0.02286	-0.01383
Order	Residual minus full: held-out	-0.01849	-0.02208	-0.01307
Role	Residual minus full: known	-0.02053	-0.02445	-0.01262
Role	Residual minus full: held-out	-0.00989	-0.01290	-0.00456
Containment	Residual minus full: known	-0.00653	-0.01330	+0.00150
Containment	Residual minus full: held-out	-0.02632	-0.03167	-0.01971
Binding	Residual minus full: known	-0.00112	-0.00169	-0.00047
Binding	Residual minus full: held-out	+0.00005	-0.00041	+0.00046
Vocabulary shifts				
Outcome	Held-out minus known: full	-0.06874	-0.09970	-0.03420
Outcome	Held-out minus known: residual	-0.01724	-0.05921	+0.02365
Polarity	Held-out minus known: full	-0.02181	-0.05533	+0.01876
Polarity	Held-out minus known: residual	-0.01899	-0.05099	+0.01929
Order	Held-out minus known: full	+0.00814	-0.01933	+0.03405
Order	Held-out minus known: residual	+0.00842	-0.01694	+0.03402
Role	Held-out minus known: full	+0.04065	+0.02004	+0.05229
Role	Held-out minus known: residual	+0.05129	+0.02767	+0.06340
Containment	Held-out minus known: full	+0.03782	+0.00525	+0.06746
Containment	Held-out minus known: residual	+0.01804	-0.01207	+0.04645
Binding	Held-out minus known: full	-0.01574	-0.02114	-0.00959
Binding	Held-out minus known: residual	-0.01457	-0.01982	-0.00870

E TRANSFER OBJECTIVES, SUPERVISION AND CONTROLS

E.1 TRAINING POOLS AND MAP FITTING

Transfer uses the centered $z_i(t)$ view and separate affine readouts from Section 2.1. Vectors are columns, and $W_{B\to A}$ has shape $d_A \times d_B$; implementations with rows store transposed maps. All maps retain the native dimensions. The same mean μ_i, fitted on the original 5,000 ordinary training texts, is used in every condition.

Table 32: Supervision per map direction and focus-distinction/family fold. Counts are not independent annotations summed across overlapping folds.

Condition	Ordinary texts	Semantic texts	Semantic information
Ordinary (O)	5,000	0	Paired text identities across encoders
Endpoints (E)	4,616	384	Selection of controlled sentences
Full/mean/variation (F/M/V)	4,616	384	Same pool, plus 192 labeled contrasts

The semantic pool contains 16 four-cell training scenes per distinction: 384 sentences and 192 oriented contrasts across six distinctions. For the focus distinction, sampling uses that fold's 72 training scenes and excludes the held family. Other distinctions use fixed pools from their globally training-only groups. Family quotas differ by at most one; category/template usage is balanced by deterministic seed-42 ordering. Validation, final and stress groups are excluded. This focus-family exclusion does not exclude every related grammatical device from other distinctions or ordinary text.

Semantic-map examples overlap the focus readout's training set on 64 sentences. Thus 288 readout examples plus 384 map examples give a union of 608, with 96 separate readout-validation examples.

Selection of semantic endpoints and labels for other distinctions are part of the map's supervision; mean matching uses the same labeled contrast inventory as full matching.

Ordinary anchors are paired unique-document Dolma excerpts, capped at 256 whitespace words. For each direction, we fit O at $\alpha \in \{.01, .1, 1\}$ and select the lowest endpoint MSE after mapped-vector normalization on 1,000 separate ordinary validation texts; smaller α breaks ties. All semantic modes share the selected ridge coefficient

$$\rho_B = \alpha \operatorname{tr}(X_{O,B}^\top X_{O,B}/5000)/d_B,$$

where $X_{O,B}$ contains the centered-unit ordinary training rows. There is no semantic-validation selector or search over the contrast-penalty multiplier.

For a given pool, the base objective is

$$L_0(W) = N^{-1} \sum_{n=1}^{N} \|W z_{B,n} - z_{A,n}\|^2 + \rho_B \|W\|_F^2, \qquad N = 5000. \tag{3}$$

O minimizes this loss on ordinary texts and E on the mixed pool. F/M/V minimize the mixed-pool base loss plus the corresponding penalty below, with multiplier one. For each distinction, its 32 training contrasts give $e_{c,k} = \Delta z_{A,c,k} - W \Delta z_{B,c,k}$ and the shared training-only scale $\kappa_c^{-1} = (7 \cdot 32)^{-1} \sum_{i,k} \|\Delta z_{i,c,k}\|^2$:

$$L_F = \tfrac{1}{6} \sum_c \kappa_c \, 32^{-1} \sum_k \|e_{c,k}\|^2, \tag{4}$$

$$L_M = \tfrac{1}{6} \sum_c \kappa_c \, \|\bar{e}_c\|^2, \qquad L_V = \tfrac{1}{6} \sum_c \kappa_c \, 32^{-1} \sum_k \|e_{c,k} - \bar{e}_c\|^2. \tag{5}$$

The identity $L_F = L_M + L_V$ concerns the added penalties, not two copies of the base objective, normalized inference or final AUC. At inference, $W z_B$ is normalized before applying the frozen source readout.

E.2 Complete transfer profiles

These tables retain all six distinctions and four evaluation conditions. Accuracy uses source-validation thresholds, so a high ranking AUC need not imply compatible thresholded decisions. Target-encoder and family profiles expose heterogeneity behind the 42-direction means.

Table 33: Transfer AUC, fixed 42-direction means. Target is the matching centered-affine readout fitted on the same label budget.

Distinction	Condition	Target	O	E	F	M	V
Outcome	Context	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Outcome	Construction	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Outcome	Lexical	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Outcome	Joint	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Polarity	Context	1.0000	0.9992	1.0000	1.0000	1.0000	0.9982
Polarity	Construction	1.0000	0.9972	0.9997	0.9998	0.9997	0.9980
Polarity	Lexical	1.0000	0.9981	0.9998	1.0000	1.0000	0.9975
Polarity	Joint	1.0000	0.9941	0.9986	0.9988	0.9993	0.9951
Order	Context	0.9995	0.6478	0.7040	0.9211	0.8466	0.7085
Order	Construction	0.6979	0.5819	0.5844	0.6172	0.5541	0.5543
Order	Lexical	0.8713	0.6155	0.6361	0.8532	0.7969	0.6502
Order	Joint	0.6358	0.5607	0.5517	0.6045	0.5445	0.5366
Role	Context	0.9980	0.6317	0.6878	0.8684	0.8573	0.6782
Role	Construction	0.8607	0.5943	0.6220	0.6924	0.6883	0.6096
Role	Lexical	0.8397	0.6331	0.6475	0.8135	0.8339	0.6570
Role	Joint	0.7461	0.6030	0.6052	0.7063	0.7295	0.6097
Containment	Context	0.9908	0.5937	0.6241	0.8524	0.8360	0.6491
Containment	Construction	0.8033	0.5672	0.5797	0.6976	0.7154	0.6041
Containment	Lexical	0.7845	0.5514	0.5621	0.7389	0.7353	0.5823
Containment	Joint	0.6340	0.5375	0.5402	0.6220	0.6336	0.5599
Binding	Context	0.8484	0.5033	0.5049	0.5533	0.5498	0.5048
Binding	Construction	0.7111	0.5038	0.5052	0.5341	0.5326	0.5024
Binding	Lexical	0.6014	0.5034	0.5034	0.5369	0.5342	0.5048
Binding	Joint	0.5651	0.5035	0.5034	0.5225	0.5215	0.5024

Table 34: Accuracy for the same interfaces using source-validation thresholds.

Distinction	Condition	Target	O	E	F	M	V
Outcome	Context	0.9924	0.6486	0.8330	0.9830	0.9868	0.7590
Outcome	Construction	0.9261	0.6462	0.8240	0.9598	0.9638	0.7558
Outcome	Lexical	0.9543	0.6422	0.8372	0.9704	0.9805	0.7603
Outcome	Joint	0.8891	0.6380	0.8319	0.9456	0.9565	0.7589
Polarity	Context	0.9899	0.6161	0.7425	0.9730	0.9696	0.6615
Polarity	Construction	0.8291	0.6358	0.7549	0.9094	0.9212	0.6955
Polarity	Lexical	0.9260	0.6101	0.7420	0.9532	0.9606	0.6555
Polarity	Joint	0.8161	0.6274	0.7540	0.9050	0.9179	0.6919
Order	Context	0.9830	0.5323	0.5774	0.7638	0.7494	0.5874
Order	Construction	0.5665	0.5193	0.5290	0.5376	0.5101	0.5175
Order	Lexical	0.7611	0.5271	0.5619	0.7013	0.6881	0.5669
Order	Joint	0.5645	0.5165	0.5232	0.5413	0.5165	0.5150
Role	Context	0.9833	0.5407	0.6057	0.7431	0.7486	0.6005
Role	Construction	0.7159	0.5305	0.5751	0.6208	0.6211	0.5590
Role	Lexical	0.7091	0.5482	0.5762	0.6689	0.7018	0.5826
Role	Joint	0.6417	0.5369	0.5529	0.5990	0.6126	0.5531
Containment	Context	0.9709	0.5290	0.5701	0.7029	0.7066	0.5892
Containment	Construction	0.6598	0.5248	0.5343	0.5891	0.5970	0.5539
Containment	Lexical	0.6812	0.5185	0.5311	0.6108	0.6126	0.5453
Containment	Joint	0.5783	0.5151	0.5186	0.5483	0.5515	0.5309
Binding	Context	0.7821	0.5014	0.5004	0.5259	0.5241	0.5022
Binding	Construction	0.5736	0.5016	0.5018	0.5162	0.5137	0.5004
Binding	Lexical	0.5569	0.5016	0.5008	0.5160	0.5118	0.5021
Binding	Joint	0.5311	0.5009	0.5027	0.5104	0.5080	0.5005

Table 35: Transfer by target encoder, averaging its six source rules. C/J denote construction/joint shifts; F/M denote full/mean penalties.

Distinction	Target	F C	M C	F J	M J
Outcome	Gemma	1.0000	1.0000	1.0000	1.0000
Outcome	BGE-M3	1.0000	1.0000	1.0000	1.0000
Outcome	Arctic	1.0000	1.0000	1.0000	1.0000
Outcome	Jina	1.0000	1.0000	1.0000	1.0000
Outcome	PPLX	1.0000	1.0000	1.0000	1.0000
Outcome	Qwen	1.0000	1.0000	1.0000	1.0000
Outcome	SFR	1.0000	1.0000	1.0000	1.0000
Polarity	Gemma	0.9999	1.0000	1.0000	0.9996
Polarity	BGE-M3	1.0000	1.0000	0.9998	1.0000
Polarity	Arctic	0.9992	0.9978	0.9925	0.9957
Polarity	Jina	1.0000	1.0000	0.9996	0.9999
Polarity	PPLX	0.9998	1.0000	1.0000	1.0000
Polarity	Qwen	1.0000	1.0000	1.0000	1.0000
Polarity	SFR	1.0000	1.0000	1.0000	1.0000
Order	Gemma	0.7056	0.6749	0.6614	0.6164
Order	BGE-M3	0.5051	0.3936	0.5375	0.4340
Order	Arctic	0.6072	0.5745	0.6074	0.5872
Order	Jina	0.4385	0.4510	0.4705	0.4701
Order	PPLX	0.5566	0.4719	0.5485	0.4831
Order	Qwen	0.7379	0.7036	0.6902	0.6656
Order	SFR	0.7695	0.6096	0.7158	0.5550
Role	Gemma	0.7631	0.7624	0.7386	0.7783
Role	BGE-M3	0.7142	0.6586	0.7139	0.6894
Role	Arctic	0.6577	0.6871	0.5885	0.6274
Role	Jina	0.6104	0.5595	0.6735	0.6565
Role	PPLX	0.7539	0.7276	0.7458	0.7545
Role	Qwen	0.6907	0.6937	0.7676	0.7956
Role	SFR	0.6569	0.7288	0.7162	0.8047
Containment	Gemma	0.7279	0.7892	0.6334	0.6951
Containment	BGE-M3	0.6370	0.6968	0.5784	0.5933
Containment	Arctic	0.7474	0.8128	0.6762	0.7190
Containment	Jina	0.6215	0.6933	0.5799	0.6503
Containment	PPLX	0.7192	0.6853	0.6135	0.5905
Containment	Qwen	0.7898	0.6730	0.6489	0.5730
Containment	SFR	0.6406	0.6576	0.6237	0.6137
Binding	Gemma	0.5107	0.5117	0.5064	0.5112
Binding	BGE-M3	0.5068	0.5202	0.5075	0.5157
Binding	Arctic	0.5067	0.5073	0.5056	0.5047
Binding	Jina	0.5558	0.5407	0.5376	0.5294
Binding	PPLX	0.5813	0.5827	0.5514	0.5503
Binding	Qwen	0.5475	0.5336	0.5282	0.5214
Binding	SFR	0.5301	0.5323	0.5209	0.5180

Table 36: Full/mean transfer by held family, fixed 42-direction means. C/J denote construction/joint shifts. Evaluations share final scenes.

Distinction	Held family	F C	M C	F J	M J
Outcome	finite-result	1.0000	1.0000	1.0000	1.0000
Outcome	adjectival-result	1.0000	1.0000	1.0000	1.0000
Outcome	nominal-result	1.0000	1.0000	1.0000	1.0000
Polarity	finite-negation	1.0000	1.0000	0.9997	0.9999
Polarity	truth-predication	0.9997	0.9994	0.9979	0.9988
Order	temporal-comparison	0.6774	0.5934	0.6699	0.5901
Order	ordinal-sequence	0.6763	0.6244	0.6374	0.5840
Order	perfect-aspect	0.4979	0.4446	0.5062	0.4594
Role	transitive-voice	0.7724	0.7305	0.7530	0.7430
Role	event-nominal	0.8283	0.8536	0.7912	0.8594
Role	event-receipt	0.6182	0.6141	0.7061	0.7221
Role	participant-nouns	0.5506	0.5549	0.5749	0.5935
Containment	locative-nesting	0.8535	0.8883	0.7254	0.7662
Containment	transitive-containment	0.7377	0.8077	0.6698	0.7114
Containment	container-roles	0.5531	0.5681	0.5173	0.5180
Containment	contents-membership	0.6463	0.5976	0.5754	0.5387
Binding	badge-possession	0.5418	0.5375	0.5288	0.5271
Binding	badge-wearing	0.5378	0.5307	0.5231	0.5193
Binding	badge-assignment	0.5257	0.5287	0.5162	0.5176
Binding	badge-attribute	0.5312	0.5336	0.5221	0.5221

E.3 Endpoint weighting, alternative pairings and energy scales

The controls use both directions of Gemma–SFR, BGE–Qwen, Jina–PPLX and Arctic–Gemma. Their reference is F on these same eight directions, not the 42-direction mean. All retain cross-model text identities.

For endpoint errors a, b on a meaningful contrast, endpoint boost (EB) adds the factor-scaled penalty $2(\|a\|^2 + \|b\|^2)$. The identity

$$2(\|a\|^2 + \|b\|^2) = \|a - b\|^2 + \|a + b\|^2$$

shows that EB adds pressure on common endpoint errors as well as differences. It is neither a label-free nor a loss-equivalent replacement for full matching.

Table 37: Alternative contrast graphs. The same text-identity graph is used across encoders, so these controls do not shuffle cross-model correspondence.

Control	Operation	Preserved structure
J0	Pair endpoints randomly within each distinction, with degree one	Endpoint inventory and distinction membership
J1	Fix negative endpoints; derange positive partners within distinction, family and background-assignment groups	Endpoint degrees and unscaled signed mean response

Fixed weighting reuses the meaningful-contrast κ_c. Energy matching multiplies it by the ratio of meaningful to alternative-graph contrast energy, averaged over the seven training representations. J1 preserves the unscaled mean, not the weighted mean penalty after this adjustment. Metrics for graph seeds 701, 702 and 703 are averaged; neither logits nor a favorably selected graph seed are used.

Table 38: Eight-direction controls, AUC. F (8) is the matched reference; graph entries average the three fixed seeds.

Distinction	Condition	F (8)	EB	J0 fix	J0 energy	J1 fix	J1 energy
Outcome	Construction	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Outcome	Joint	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Polarity	Construction	0.9995	1.0000	0.9998	0.9998	0.9998	0.9997
Polarity	Joint	0.9976	0.9985	0.9993	0.9991	0.9983	0.9987
Order	Construction	0.6382	0.6625	0.6158	0.6037	0.6552	0.6213
Order	Joint	0.6172	0.6086	0.5803	0.5662	0.6063	0.5782
Role	Construction	0.7061	0.7521	0.7040	0.6614	0.7330	0.7087
Role	Joint	0.7009	0.6754	0.6613	0.6376	0.7035	0.6933
Containment	Construction	0.6944	0.7464	0.6817	0.6456	0.7193	0.7114
Containment	Joint	0.6156	0.6142	0.6014	0.5752	0.6066	0.6017
Binding	Construction	0.5395	0.5522	0.5164	0.5063	0.5481	0.5134
Binding	Joint	0.5263	0.5259	0.5101	0.5042	0.5237	0.5071

E.4 Complete transfer contrasts

The paired bootstrap follows Appendix B.3, sharing complete-scene multiplicities across directions and methods. Maps and readouts remain fixed. All 168 contrasts are grouped below: 48 alignment-condition comparisons, 60 matched-target gaps and 60 endpoint/pairing-control comparisons. The latter use the same eight-direction reference as Table 38.

Table 39: All 168 map AUC contrasts, grouped by question, with conditional 95% scene intervals. Differences are first minus second; (8) denotes the matched diagnostic directions.

Distinction	Measurement or contrast	Estimate	Lower	Upper
Alignment constraints				
Outcome	E minus O: construction	+0.00000	+0.00000	+0.00000
Outcome	F minus E: construction	+0.00000	+0.00000	+0.00000
Outcome	F minus M: construction	+0.00000	+0.00000	+0.00000
Outcome	F minus V: construction	+0.00000	+0.00000	+0.00000
Outcome	E minus O: joint	+0.00000	+0.00000	+0.00000
Outcome	F minus E: joint	+0.00000	+0.00000	+0.00000
Outcome	F minus M: joint	+0.00000	+0.00000	+0.00000
Outcome	F minus V: joint	+0.00000	+0.00000	+0.00000
Polarity	E minus O: construction	+0.00246	+0.00203	+0.00286
Polarity	F minus E: construction	+0.00016	+0.00006	+0.00028
Polarity	F minus M: construction	+0.00015	+0.00002	+0.00032
Polarity	F minus V: construction	+0.00188	+0.00156	+0.00215
Polarity	E minus O: joint	+0.00450	+0.00365	+0.00527
Polarity	F minus E: joint	+0.00019	-0.00003	+0.00037
Polarity	F minus M: joint	-0.00048	-0.00075	-0.00024
Polarity	F minus V: joint	+0.00375	+0.00312	+0.00430
Order	E minus O: construction	+0.00254	+0.00079	+0.00376
Order	F minus E: construction	+0.03280	+0.02882	+0.03711
Order	F minus M: construction	+0.06306	+0.06105	+0.06593
Order	F minus V: construction	+0.06287	+0.05836	+0.06827
Order	E minus O: joint	-0.00897	-0.00999	-0.00835
Order	F minus E: joint	+0.05273	+0.04928	+0.05629
Order	F minus M: joint	+0.05999	+0.05754	+0.06345
Order	F minus V: joint	+0.06785	+0.06462	+0.07167
Role	E minus O: construction	+0.02769	+0.02402	+0.03041
Role	F minus E: construction	+0.07042	+0.06524	+0.07606

Distinction	Measurement or contrast	Estimate	Lower	Upper
Role	F minus M: construction	+0.00413	-0.00103	+0.01077
Role	F minus V: construction	+0.08275	+0.07551	+0.09006
Role	E minus O: joint	+0.00222	+0.00113	+0.00329
Role	F minus E: joint	+0.10107	+0.09754	+0.10435
Role	F minus M: joint	-0.02320	-0.02618	-0.01947
Role	F minus V: joint	+0.09661	+0.09327	+0.09936
Containment	E minus O: construction	+0.01249	+0.01112	+0.01367
Containment	F minus E: construction	+0.11794	+0.11652	+0.12026
Containment	F minus M: construction	-0.01779	-0.02018	-0.01462
Containment	F minus V: construction	+0.09357	+0.09135	+0.09618
Containment	E minus O: joint	+0.00279	+0.00210	+0.00364
Containment	F minus E: joint	+0.08175	+0.07780	+0.08726
Containment	F minus M: joint	-0.01157	-0.01521	-0.00729
Containment	F minus V: joint	+0.06208	+0.05860	+0.06662
Binding	E minus O: construction	+0.00133	+0.00090	+0.00175
Binding	F minus E: construction	+0.02897	+0.02871	+0.03159
Binding	F minus M: construction	+0.00149	+0.00082	+0.00342
Binding	F minus V: construction	+0.03174	+0.03150	+0.03428
Binding	E minus O: joint	-0.00016	-0.00042	+0.00022
Binding	F minus E: joint	+0.01914	+0.01912	+0.02130
Binding	F minus M: joint	+0.00100	+0.00087	+0.00237
Binding	F minus V: joint	+0.02011	+0.02005	+0.02229
Matched-target gaps				
Outcome	O minus target: construction	+0.00000	+0.00000	+0.00000
Outcome	E minus target: construction	+0.00000	+0.00000	+0.00000
Outcome	F minus target: construction	+0.00000	+0.00000	+0.00000
Outcome	M minus target: construction	+0.00000	+0.00000	+0.00000
Outcome	V minus target: construction	+0.00000	+0.00000	+0.00000
Outcome	O minus target: joint	+0.00000	+0.00000	+0.00000
Outcome	E minus target: joint	+0.00000	+0.00000	+0.00000
Outcome	F minus target: joint	+0.00000	+0.00000	+0.00000
Outcome	M minus target: joint	+0.00000	+0.00000	+0.00000
Outcome	V minus target: joint	+0.00000	+0.00000	+0.00000
Polarity	O minus target: construction	-0.00278	-0.00322	-0.00230
Polarity	E minus target: construction	-0.00032	-0.00049	-0.00017
Polarity	F minus target: construction	-0.00016	-0.00027	-0.00007
Polarity	M minus target: construction	-0.00031	-0.00054	-0.00012
Polarity	V minus target: construction	-0.00203	-0.00236	-0.00167
Polarity	O minus target: joint	-0.00584	-0.00695	-0.00463
Polarity	E minus target: joint	-0.00134	-0.00177	-0.00091
Polarity	F minus target: joint	-0.00115	-0.00158	-0.00073
Polarity	M minus target: joint	-0.00067	-0.00100	-0.00036
Polarity	V minus target: joint	-0.00490	-0.00567	-0.00403
Order	O minus target: construction	-0.11603	-0.12131	-0.11012
Order	E minus target: construction	-0.11349	-0.11884	-0.10802
Order	F minus target: construction	-0.08069	-0.08532	-0.07591
Order	M minus target: construction	-0.14375	-0.14831	-0.13959
Order	V minus target: construction	-0.14356	-0.15015	-0.13757
Order	O minus target: joint	-0.07511	-0.07914	-0.07074
Order	E minus target: joint	-0.08407	-0.08857	-0.07958
Order	F minus target: joint	-0.03134	-0.03489	-0.02780
Order	M minus target: joint	-0.09133	-0.09600	-0.08763
Order	V minus target: joint	-0.09919	-0.10395	-0.09500
Role	O minus target: construction	-0.26645	-0.27220	-0.25853
Role	E minus target: construction	-0.23876	-0.24655	-0.23006

Distinction	Measurement or contrast	Estimate	Lower	Upper
Role	F minus target: construction	-0.16834	-0.18024	-0.15519
Role	M minus target: construction	-0.17247	-0.18601	-0.15912
Role	V minus target: construction	-0.25109	-0.25679	-0.24391
Role	O minus target: joint	-0.14307	-0.14920	-0.13760
Role	E minus target: joint	-0.14085	-0.14711	-0.13526
Role	F minus target: joint	-0.03978	-0.04543	-0.03478
Role	M minus target: joint	-0.01658	-0.02498	-0.00977
Role	V minus target: joint	-0.13639	-0.14227	-0.13067
Containment	O minus target: construction	-0.23606	-0.24022	-0.23104
Containment	E minus target: construction	-0.22357	-0.22771	-0.21877
Containment	F minus target: construction	-0.10563	-0.10957	-0.10018
Containment	M minus target: construction	-0.08784	-0.09143	-0.08327
Containment	V minus target: construction	-0.19920	-0.20356	-0.19348
Containment	O minus target: joint	-0.09654	-0.10359	-0.09117
Containment	E minus target: joint	-0.09375	-0.10094	-0.08814
Containment	F minus target: joint	-0.01200	-0.01605	-0.00814
Containment	M minus target: joint	-0.00043	-0.00521	+0.00395
Containment	V minus target: joint	-0.07408	-0.08065	-0.06866
Binding	O minus target: construction	-0.20723	-0.21340	-0.20393
Binding	E minus target: construction	-0.20590	-0.21198	-0.20264
Binding	F minus target: construction	-0.17693	-0.18154	-0.17295
Binding	M minus target: construction	-0.17842	-0.18382	-0.17486
Binding	V minus target: construction	-0.20867	-0.21479	-0.20538
Binding	O minus target: joint	-0.06160	-0.06879	-0.06072
Binding	E minus target: joint	-0.06175	-0.06889	-0.06085
Binding	F minus target: joint	-0.04262	-0.04842	-0.04106
Binding	M minus target: joint	-0.04362	-0.05002	-0.04262
Binding	V minus target: joint	-0.06273	-0.06995	-0.06171
Endpoint and pairing controls (eight directions)				
Outcome	EB minus F (8): construction	+0.00000	+0.00000	+0.00000
Outcome	J0-fixed minus F (8): construction	+0.00000	+0.00000	+0.00000
Outcome	J0-energy minus F (8): construction	+0.00000	+0.00000	+0.00000
Outcome	J1-fixed minus F (8): construction	+0.00000	+0.00000	+0.00000
Outcome	J1-energy minus F (8): construction	+0.00000	+0.00000	+0.00000
Outcome	EB minus F (8): joint	+0.00000	+0.00000	+0.00000
Outcome	J0-fixed minus F (8): joint	+0.00000	+0.00000	+0.00000
Outcome	J0-energy minus F (8): joint	+0.00000	+0.00000	+0.00000
Outcome	J1-fixed minus F (8): joint	+0.00000	+0.00000	+0.00000
Outcome	J1-energy minus F (8): joint	+0.00000	+0.00000	+0.00000
Polarity	EB minus F (8): construction	+0.00047	+0.00016	+0.00086
Polarity	J0-fixed minus F (8): construction	+0.00034	+0.00010	+0.00065
Polarity	J0-energy minus F (8): construction	+0.00025	+0.00004	+0.00052
Polarity	J1-fixed minus F (8): construction	+0.00030	+0.00008	+0.00058
Polarity	J1-energy minus F (8): construction	+0.00020	+0.00004	+0.00044
Polarity	EB minus F (8): joint	+0.00092	+0.00025	+0.00162
Polarity	J0-fixed minus F (8): joint	+0.00172	+0.00103	+0.00243
Polarity	J0-energy minus F (8): joint	+0.00157	+0.00097	+0.00220
Polarity	J1-fixed minus F (8): joint	+0.00075	+0.00021	+0.00132
Polarity	J1-energy minus F (8): joint	+0.00117	+0.00071	+0.00166
Order	EB minus F (8): construction	+0.02425	+0.01900	+0.02881
Order	J0-fixed minus F (8): construction	-0.02237	-0.02734	-0.01745
Order	J0-energy minus F (8): construction	-0.03447	-0.04011	-0.02894
Order	J1-fixed minus F (8): construction	+0.01702	+0.01051	+0.02290
Order	J1-energy minus F (8): construction	-0.01687	-0.02277	-0.01172
Order	EB minus F (8): joint	-0.00866	-0.01160	-0.00565

Distinction	Measurement or contrast	Estimate	Lower	Upper
Order	J0-fixed minus F (8): joint	-0.03689	-0.03957	-0.03427
Order	J0-energy minus F (8): joint	-0.05098	-0.05407	-0.04795
Order	J1-fixed minus F (8): joint	-0.01090	-0.01428	-0.00782
Order	J1-energy minus F (8): joint	-0.03899	-0.04243	-0.03610
Role	EB minus F (8): construction	+0.04602	+0.04002	+0.05028
Role	J0-fixed minus F (8): construction	-0.00212	-0.00737	+0.00174
Role	J0-energy minus F (8): construction	-0.04463	-0.05074	-0.03944
Role	J1-fixed minus F (8): construction	+0.02688	+0.02238	+0.02980
Role	J1-energy minus F (8): construction	+0.00266	-0.00378	+0.00743
Role	EB minus F (8): joint	-0.02552	-0.02901	-0.02214
Role	J0-fixed minus F (8): joint	-0.03960	-0.04313	-0.03625
Role	J0-energy minus F (8): joint	-0.06332	-0.06797	-0.05865
Role	J1-fixed minus F (8): joint	+0.00263	-0.00022	+0.00525
Role	J1-energy minus F (8): joint	-0.00761	-0.01199	-0.00321
Containment	EB minus F (8): construction	+0.05198	+0.04741	+0.05559
Containment	J0-fixed minus F (8): construction	-0.01275	-0.01681	-0.00918
Containment	J0-energy minus F (8): construction	-0.04883	-0.05283	-0.04512
Containment	J1-fixed minus F (8): construction	+0.02487	+0.01960	+0.02894
Containment	J1-energy minus F (8): construction	+0.01696	+0.01147	+0.02155
Containment	EB minus F (8): joint	-0.00138	-0.00483	+0.00200
Containment	J0-fixed minus F (8): joint	-0.01418	-0.01772	-0.01061
Containment	J0-energy minus F (8): joint	-0.04046	-0.04537	-0.03561
Containment	J1-fixed minus F (8): joint	-0.00903	-0.01270	-0.00587
Containment	J1-energy minus F (8): joint	-0.01387	-0.01794	-0.00987
Binding	EB minus F (8): construction	+0.01277	+0.00971	+0.01529
Binding	J0-fixed minus F (8): construction	-0.02306	-0.02674	-0.02129
Binding	J0-energy minus F (8): construction	-0.03319	-0.03733	-0.03133
Binding	J1-fixed minus F (8): construction	+0.00858	+0.00530	+0.01111
Binding	J1-energy minus F (8): construction	-0.02603	-0.02990	-0.02429
Binding	EB minus F (8): joint	-0.00039	-0.00207	+0.00106
Binding	J0-fixed minus F (8): joint	-0.01617	-0.01900	-0.01515
Binding	J0-energy minus F (8): joint	-0.02213	-0.02555	-0.02096
Binding	J1-fixed minus F (8): joint	-0.00260	-0.00442	-0.00143
Binding	J1-energy minus F (8): joint	-0.01919	-0.02231	-0.01812

F Analysis chronology and reproducibility

F.1 What was known when analyses were specified

The stages below distinguish fitting and selection from final evaluation. Native results and earlier exploratory outcomes were already visible when the map and geometry analyses were designed. Their later freeze does not restore hypothesis blindness or account for adaptive design.

Table 40: Chronology of the controlled analyses. Detailed timestamps and source identities are retained in the reproducibility records.

Stage	Information available and subsequent action
Native fitting	Readouts and lexical controls frozen before native final scoring.
Map/geometry design	Native and earlier exploratory results inspected before specifying these analyses.
Map/geometry evaluation	Selected maps, preprocessing and analysis choices frozen before new final scoring.
Numerical verification	Saved inputs, predictions, matrices and statistics independently reconstructed after evaluation.

Final outcomes do not select regularization, penalty multiplier, graph seed or which conditions are reported. MultiNLI has its own training/validation selection and subsequent public-dev evaluation, with separate labels and source records.

F.2 Computational checks and resources

Separately implemented checks reconstruct split membership, input overlap, map selection, objective identities, predictions, geometry matrices and resampling statistics. They verify computation on the recorded data, not linguistic validity or stronger semantic interpretations. The maximum native score discrepancy is 3.84×10^{-13}; map/geometry score and matrix discrepancies are below 5×10^{-13}, with zero threshold-decision disagreements. Geometry null and bootstrap reconstructions agree within 5.56×10^{-16}. The independently reconstructed MultiNLI logits differ by at most 1.82×10^{-5} with no argmax disagreements. Ties, repeated scenes, undefined panels and restart consistency are included in the checks.

All 1,800 native affine candidates converge. Of 504 native neural fits, 244 reach the 100-epoch cap and 260 stop by patience; neither certifies a global optimum. All 35 selected MultiNLI affine fits converge. Computation records distinguish allocated resources from measured utilization and exclude earlier cached artifacts:

Table 41: Scoped controlled-analysis resource accounting. CPU hours are requested allocation hours, not measured processor utilization or total project cost.

Stage	GPU-container hours	Requested CPU-container hours
Native readout stage	0.5864	1.5033
Map/geometry stage	1.8633	12.3711

The retained source records include exact model/component versions, fitting configurations, saved scores, source hashes and deterministic table/figure generators. Numeric importers preserve every reported contrast and its original source path; presentation changes do not refit models or recompute intervals. A repository link will be provided after acceptance.