PREDICTABLE CROSS-MODEL CONTROL WITHOUT SHARED COORDINATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Can information from one model guide a predictable intervention in another? We apply a source model's signed relation weights over paired examples to the receiver's own states, so no hidden vector crosses models. Across twelve directed paths among four decoder-only Transformer families and four controlled tasks, the resulting directions are closer to each receiver's relational reference than randomly permuted donors in every path-task cell. With identical donors, source weights produce larger held-out effects than a target-neutral control. A receiver score from disjoint audit items ranks held-out additive effects across the source construction and three controls, with Spearman correlations of 0.837 and 0.825 at two doses. In natural-context tests including an encoder-decoder model, source responses reduce receiver-response prediction error by 32.5% beyond label, action, and clean-output baselines. These results establish two concrete uses of cross-model information: constructing receiver-native directions whose additive effects can be ranked by a local measurement, and predicting output changes from observed source responses.

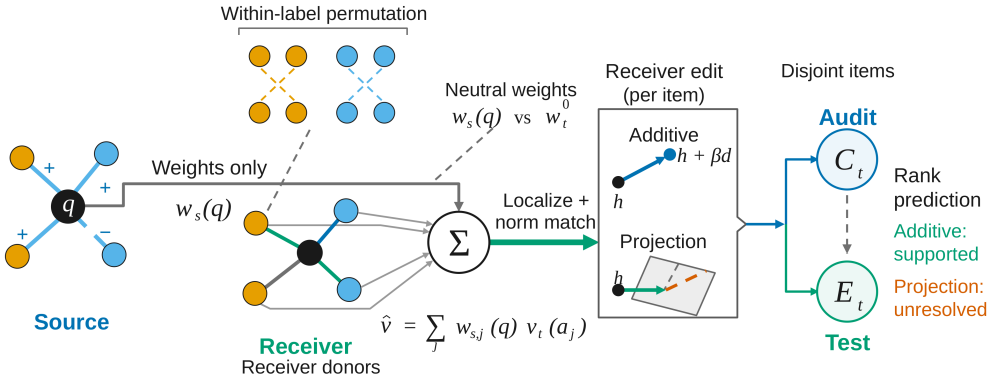


Figure 1: Signed source weights combine centered receiver innovations into a receiver-native proposal. Neutral weighting keeps the same donors; within-label permutation keeps source weights. After localization and norm matching, edits are evaluated on disjoint audit and test items. Audit scores C_t and test effects E_t are compared in a pooled analysis across source relational, task ridge, unrestricted and within-label permutations. Additive ranking is supported; input-dependent projection ranking remains unresolved. Source value is supported relative to neutral weighting; fine donor identity remains unresolved. Donor colors denote labels; geometry is schematic.

1 INTRODUCTION

Evidence for common structure across objectives, architectures, and modalities has made shared representations a plausible basis for model interoperability (Kornblith et al., 2019; Williams et al., 2021; Huh et al., 2024). Recent studies also find recurring features and relational structure across language models (Lan et al., 2024; Hu et al., 2026; Jiralerspong & Bricken, 2026). Relative representations support communication without a common coordinate system (Moschella et al., 2023),

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.

and translated activations can alter behavior across architectures (Armstrong et al., 2026). We ask what a source model adds to an edit inside another model, and whether the edit's effect can be predicted in advance.

Cross-model control requires evidence beyond representation similarity. Similarity metrics compare activation spaces after choosing an admissible alignment, while model stitching asks whether a learned connector preserves downstream behavior (Kornblith et al., 2019; Williams et al., 2021; Bansal et al., 2021). An activation intervention must also pass through the receiver's nonlinear computation and readout. Finite observations of natural states and clean outputs identify an unqueried response only if every distinction discarded by the representation summary is also null for receiver response. This gap matters for activation steering and model interoperability, where directions are increasingly used as causal interfaces (Todd et al., 2024; Li et al., 2023; Rimsky et al., 2024; Arditi et al., 2024).

Three tests distinguish recurring structure from its functional use. *Relational recurrence* compares how models weight the same paired support examples after a context change. *Source identification* measures the source's contribution relative to controls with the same receiver access. *Functional qualification* asks whether a receiver measurement on disjoint audit items predicts held-out intervention effects.

Our method uses source relation weights over paired support items to combine the corresponding receiver states, keeping every hidden vector inside the receiver (Figure 1). Matched controls hold receiver access and edit scale fixed. A receiver-local score then measures gain, finite-dose stability, and off-task response on audit items before additive steering and projection removal are evaluated on disjoint test items.

The first result identifies a recurring relation pattern with functional value relative to matched controls. Source-instantiated directions are closer to the receiver's relational reference than unrestricted donor assignment in all 48 directed path-task cells, with mean relation-weight cosine 0.834. Task-matched recurrence persists after removing label and neutral-prompt relation components. Source information also improves held-out effects over the target-neutral control. A separate natural-context experiment measures source information directly: observing the source's response improves receiver-response prediction in all twelve task and directed-model combinations, including an encoder-decoder model.

The second result makes intervention assessment prospective. A receiver-local score ranks held-out additive effects across the source construction and three controls at doses 0.50 and 1.00, with Spearman correlations of 0.837 and 0.825. Both remain positive after excluding any source, receiver, task, or incident model. Thus, the audit measurement helps compare candidate directions before observing their test effects.

Our contributions are:

- We characterize when a representation summary determines receiver response and bound how proposal error, receiver gain, and curvature govern finite-dose effects.
- We establish relational recurrence across twelve directed paths and test source specificity through label, neutral-relation, donor-assignment, and natural-response controls.
- We show that a disjoint receiver-local score predicts held-out additive steering across sources, receivers, tasks, and two intervention doses.

2 Related Work

2.1 Comparing representations

Representation comparison depends on which transformations are treated as irrelevant. Linear centered kernel alignment is invariant to isotropic scaling and orthogonal transformations (Kornblith et al., 2019). Generalized shape metrics make the admissible alignment group explicit (Williams et al., 2021), while model stitching tests whether a simple connector preserves downstream computation (Bansal et al., 2021). These approaches show that representation similarity depends on an equivalence class and a downstream test. We therefore report representation recurrence and intervention effects separately.

Foundation models provide growing evidence for recurring semantic organization. The Platonic Representation Hypothesis proposes convergence toward common structure as models and training signals improve (Huh et al., 2024). Sparse-autoencoder and crosscoder studies identify shared and model-specific features across language models (Lan et al., 2024; Lindsey et al., 2024; Jiralerspong & Bricken, 2026), while work on linear representations characterizes concept and relation geometry within individual models (Park et al., 2024; Hernandez et al., 2024). The closest relational study reports shared concept organization and context transformations across language-model families (Hu et al., 2026). We use its signed relational weighting rule as a testable proposal, then add matched controls that determine the resolution of the source information and held-out interventions that test functional use.

2.2 Cross-model alignment and relative representations

Global alignment methods learn a map between latent spaces, whereas relative representations describe examples through their relations to shared anchors. The latter can support latent communication without assuming a common basis (Moschella et al., 2023). Cross-architecture translation likewise reports behavioral correction together with only a weak relation between alignment quality and behavioral improvement (Armstrong et al., 2026). Our construction transfers source-internal coefficients and combines receiver-native donor states. It avoids direct coordinate comparison and uses paired receiver support to instantiate the target proposal.

Relative alignment leaves an identification problem. Support items may be organized mainly by answer label, allowing many donor assignments to realize similar predictions. An unrestricted permutation can reveal recurrence but cannot isolate exact donor identity. Our label-preserving derangement tests fine identity while holding target access, labels, and intervention magnitude fixed. Natural-response prediction provides a complementary test of information specific to the changed context.

2.3 Activation interventions and causal claims

Causal model editing includes parameter interventions such as Rank-One Model Editing (ROME), which changes factual associations in model weights (Meng et al., 2022). Function vectors, inference-time intervention, activation engineering, contrastive activation addition, and refusal-direction removal instead construct or manipulate task-relevant activation directions (Todd et al., 2024; Li et al., 2023; Turner et al., 2023; Rimsky et al., 2024; Arditi et al., 2024). These activation methods estimate and apply a direction within one model. Cross-model reuse adds receiver uncertainty because each receiver has its own local gain, curvature, and readout.

Recent work makes this receiver uncertainty concrete. Steering measurements can change with the readout, intervention scale, and operating point (Wu et al., 2026), while manifold-aware interventions study the path followed by a finite within-model edit (Wurgaft et al., 2026). Causal representation learning and causal abstraction likewise distinguish predictive latent variables from interventionally valid mechanisms (Schölkopf et al., 2021; Geiger et al., 2025). We operationalize this distinction across models by measuring local gain, finite-dose stability, and off-task response before held-out intervention outcomes are observed.

3 Relational Transport and Receiver Qualification

3.1 Context innovations

We first isolate the item-specific state change induced by task context. Let $h_{m,\ell_m}(x, c)$ be the final-token state of model m at cut ℓ_m for item x under task context c, and let c_0 retain the task vocabulary without its decisive relation. Using fit items F, we remove the mean context shift and average the two prompt surfaces:

$$\begin{aligned} v_m(x, c) &= \mathbb{E}_{f\in\{1,2\}}[h_{m,\ell_m}(x_f, c) - h_{m,\ell_m}(x_f, c_0) - \mu_{m,c}] , \\ \mu_{m,c} &= \mathbb{E}_{x\in F,f}[h_{m,\ell_m}(x_f, c) - h_{m,\ell_m}(x_f, c_0)] . \end{aligned} \tag{1}$$

We compare within-model relations between these centered context innovations and measure their functional effects separately.

3.2 RELATIONAL TRANSPORT WITHOUT COORDINATE MATCHING

Relational transport moves source coefficients while keeping hidden vectors in the receiver. For query q and paired support items $A = \{a_j\}_{j=1}^{K}$ with nonzero innovations, the source computes signed cosine similarities and normalized relation weights:

$$a_{s,j}(q) = \left\langle \frac{v_s(q)}{\|v_s(q)\|_2}, \frac{v_s(a_j)}{\|v_s(a_j)\|_2} \right\rangle, \qquad w_{s,j}(q) = \frac{a_{s,j}(q)}{\sum_k |a_{s,k}(q)|}. \tag{2}$$

For nonzero similarity vectors, the receiver-native proposal is

$$\widehat{v}_{t\leftarrow s}(q) = \sum_{j=1}^{K} w_{s,j}(q) v_t(a_j). \tag{3}$$

Only the coefficients cross models. Every vector in Eq. 3 belongs to receiver t. This relative construction follows the signed weighting rule studied in shared concept geometry (Hu et al., 2026) and is related to anchor-based latent communication (Moschella et al., 2023). Source relation weights are invariant to an orthogonal basis change or uniform positive rescaling of source innovations. The receiver proposal is equivariant to the corresponding transformation of receiver innovations. Neither property extends to an arbitrary affine transform.

The receiver's own relation weights define a target-native reference $\widetilde{v}_t(q)$. If the source and target similarity-vector ℓ_1 norms are at least $\gamma > 0$ and $R_t = \max_j \|v_t(a_j)\|_2$, normalization and the triangle inequality give

$$\|\widehat{v}_{t\leftarrow s}(q) - \widetilde{v}_t(q)\|_2 \le R_t \|w_s(q) - w_t(q)\|_1 \le \frac{2R_t}{\gamma} \|a_s(q) - a_t(q)\|_1. \tag{4}$$

The bound explains why recurrent relations can reduce receiver-native direction error. It contains no receiver readout, so it cannot determine a causal effect.

This missing information has an exact local form. Let Γ_t be a linear map from proposal coefficients to receiver edits, let J_t be the Jacobian from the intervention site to the output score, and define the response operator $\mathcal{R}_t = J_t \Gamma_t$. If S is a linear representation summary available to the transfer rule, then some linear map A_t recovers every local response from that summary precisely when

$$\mathcal{R}_t = A_t S \quad \Longleftrightarrow \quad \ker S \subseteq \ker \mathcal{R}_t. \tag{5}$$

Relational recurrence constrains S, but neither the Transformer computation nor the recurrence result guarantees this kernel inclusion. Appendix C.4 proves Eq. 5, gives a finite-access non-identifiability construction, and derives the response-query requirement on a proposal span.

Finite causal effects also depend on receiver-local response. Let $\Delta_t(\beta, d) = M_t(h + \beta d) - M_t(h)$. If the receiver margin has gradient norm at most G_t and an L_t-Lipschitz gradient along the two edit segments, Taylor's theorem gives

$$\left| \Delta_t(\beta, \widehat{d}) - \Delta_t(\beta, \widetilde{d}) \right| \le |\beta| G_t \|\widehat{d} - \widetilde{d}\|_2 + \frac{L_t \beta^2}{2} \left(\|\widehat{d}\|_2^2 + \|\widetilde{d}\|_2^2 \right). \tag{6}$$

The two bounds address different stages. Equation 4 controls the receiver-native proposal before tangent localization and norm matching. Equation 6 shows that finite effects for directions passed to the receiver also depend on local gain and curvature. Functional prediction therefore needs receiver-side information about local gain and finite-dose stability, which we estimate on a disjoint audit split. Appendix C gives the derivation.

3.3 RECEIVER-LOCAL FUNCTIONAL QUALIFICATION

Receiver-local qualification measures how one receiver responds to a proposed direction before held-out outcomes are observed. Each predicted vector is first projected into the tangent plane orthogonal to the receiver's neutral state and norm matched to a fixed relational reference. Let $M_t(h)$ be the signed full-string decision margin after continuing receiver t from activation h. On audit items, the class-balanced symmetric secant at dose ϵ is

$$g_{t,\epsilon}(d) = \mathbb{E}_{\text{bal}} \left[\frac{M_t(h + \epsilon d) - M_t(h - \epsilon d)}{2\epsilon} \right]. \tag{7}$$

The small-dose secant measures local gain. We denote its relative change between $\epsilon = 0.25$ and 0.50 by $I_t(d)$ and the ratio of mean absolute off-task response to $|g_{t,0.25}(d)|$ by $O_t(d)$. The receiver-local qualification score is

$$C_t(d) = \frac{\sigma\, g_{t,0.25}(d)}{1 + I_t(d) + O_t(d)}, \tag{8}$$

where $\sigma = +1$ for additive steering and $\sigma = -1$ for projection harm. The score is fixed before test effects are evaluated and ranks proposals for one receiver and intervention family.

4 Experimental Design

4.1 Models, tasks, and information splits

The controlled roster contains Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Gemma-2-9B-it, and Mistral-7B-Instruct-v0.3 (Qwen Team, 2024; Grattafiori et al., 2024; Gemma Team, 2024; Jiang et al., 2023). Every model serves as source and receiver, yielding 12 directed paths. We test entity binding, spatial relation, subject agreement, and two-hop precedence, for 48 path-task cells. Each task has 40 fit, 16 validation, 16 audit, and 40 test items, each rendered in two surface forms. Fit and validation provide $K = 56$ paired support items. Audit items estimate Eq. 8. Test items provide held-out outcomes.

The main cut is relative depth 0.75, corresponding to layers 21, 24, 32, and 24 in the four models. Support sizes $K \in \{8, 24, 56\}$ and fixed cuts at 0.625 and 0.875 provide protocol-fixed sensitivity checks.

BoolQ (Clark et al., 2019) and the Recognizing Textual Entailment (RTE) task in the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) form a separate public stress stratum. It uses Qwen as source and the other three models as receivers.

A separate natural-response experiment evaluates Qwen, Mistral, and the instruction-finetuned encoder-decoder model FLAN-T5-XL (Chung et al., 2024). Two generated tasks track object exchanges and order relations. For each task, 32 scenes fit response predictors and 64 new scenes test them; two question positions are also withheld from fitting. All six directed model pairs are evaluated. Source responses to the changed contexts are observed, while receiver test responses are held out. Appendix G specifies the action and label baselines, four-candidate score coordinates, and scene-clustered inference.

4.2 Matched constructions and interventions

Table 1 holds receiver donors and edit scale fixed while changing the information used to construct a direction. This design distinguishes coarse recurrence from fine donor identity. A same-access task ridge learns an affine source-to-receiver map from the same paired support, providing a direct alignment baseline. Receiver-native relational predictions define the representation reference.

Table 1: Matched constructions vary source information while holding receiver access fixed. Donor-based arms use the same receiver donor set, and the ridge uses the same paired support.

Construction	Information varied	Question answered
Source relational	Source weights on receiver task donors	Candidate transfer
Target-neutral same-donor control	Receiver neutral-state weights on the same donors	Conditional source value
Task ridge	Paired support innovations	Same-access alignment baseline
Unrestricted derangement	Full permutation of receiver donors	Recurrence beyond arbitrary assignment
Label-preserving derangement	Within-label donor permutation	Fine identity beyond label-preserving assignment
Target relational	Receiver weights on receiver task donors	Target-native reference

Additive steering uses symmetric test doses $\beta \in \{0.50, 1.00\}$ and reports the class-balanced full-string margin secant. Projection removal uses a norm-matched displacement along each predicted

local axis at unit dose and reports the sign-oriented direct margin change. Appendix C specifies both implemented operators. Each endpoint uses a no-intervention reference from the same forward-scoring frame. Appendix E reports off-task response, paraphrase agreement, balanced top-one decisions, and numerical no-operation checks.

4.3 Inference

Dependence enters because each model participates in several directed paths and all methods share task items. Receiver qualification therefore correlates audit scores with held-out effects for four method roles in each of 48 directed path-task cells, yielding 192 values, and permutes roles only within cells. Holm correction covers the three endpoints. The crossed bootstrap resamples shared model nodes, tasks, and the disjoint audit and test items. We repeat the analysis after removing each source, receiver, task, or model together with all incident paths.

Each path-task cell contributes one paired difference to a method contrast. Sign randomization tests that contrast, Holm correction covers the nine endpoint-control comparisons, and a shared-node and task bootstrap estimates uncertainty. The public tasks remain descriptive because their six cells share one source. Appendix D gives the exact resampling procedure and decision criteria.

5 Results

5.1 Relational structure recurs across all directed paths

Relational structure recurs across every directed model pair without a shared hidden coordinate system. The source-instantiated prediction is closer to the target-relational reference than unrestricted donor derangement in all 48 path-task cells. Its pooled raw-distance advantage is 30.622 in receiver activation units before tangent localization and norm matching, with a shared-node and task 95% confidence interval (CI) of $[2.643, 69.262]$. Raw distances are scale dependent across receivers, so dimensionless summaries provide a second view: mean relation-weight cosine is 0.834, and independently chosen sources produce receiver-native proposals with mean pairwise cosine 0.846.

Task matching persists after explicit nuisance removal. A post hoc analysis projects relation matrices away from the fixed query-label and donor-label indicator spaces, then compares the identity task matching with $13{,}824$ model-wise task relabelings that remain consistent across model pairs. The same-minus-cross-task residual is $.255$ on audit queries and $.240$ on disjoint test queries, with Holm-adjusted $p = .0054$ on both splits. After also removing the measured neutral-prompt relation span, the gaps remain $.114$ and $.097$, with adjusted $p = .0069$ and $.0087$ (Appendix E.3). This aggregate task specificity is distinct from donor identity: relation-weight signs agree at 0.887, exact top-donor agreement is 0.159, and the representation interval against within-label donor derangement crosses zero. Figure 2 shows the resulting evidence boundary.

5.2 Source relations add held-out functional information

Source-derived relation weights add functional information beyond the target-neutral control. With identical receiver donors, they produce larger additive steering effects at both doses and greater decision-margin damage under removal. Mean advantages range from 1.040 to 2.343, all three shared-node and task 95% CIs exclude zero, and at least 40 of 48 cells favor the source-relational construction (Table 2). These effects also change decisions: descriptive surface-level balanced accuracy rises from 89.1% to 94.0% under steering at $\beta = 0.50$ (Appendix Table 20).

The matched controls locate the source advantage in how donors are organized. The relational construction exceeds unrestricted donor derangement at all three endpoints, with mean advantages from 1.282 to 4.510 and positive CI lower bounds. Against same-access task ridge and within-label donor permutation, every interval includes zero. Within-label permutation has slightly larger additive point estimates. Thus, source weights add value over neutral-state weighting and arbitrary donor assignment; the comparison with label-preserving organization remains unresolved (Figure 3).

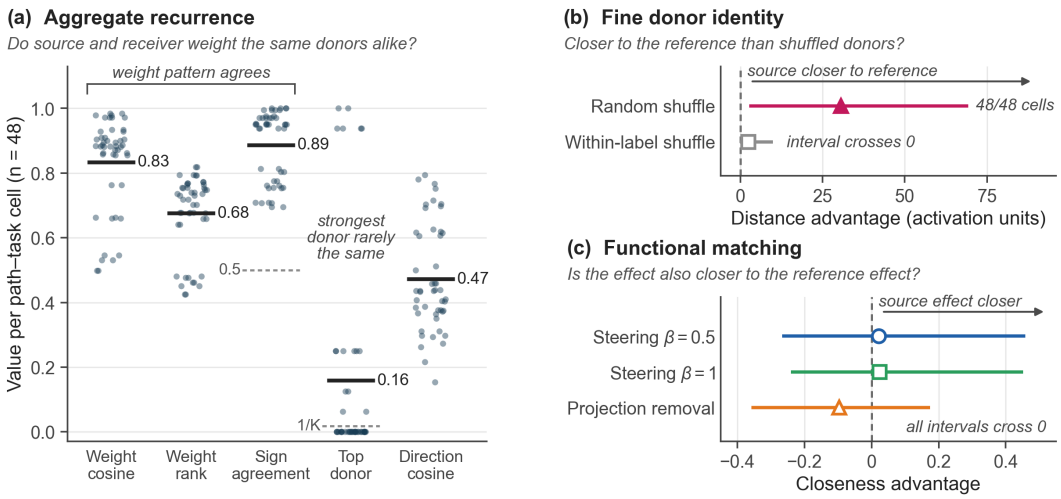


Figure 2: Matched tests separate relation recurrence, donor identity, and effect matching. (a) Points show cells and solid lines mark means; dashed 0.5 and $1/K$ lines are numerical references. (b) Distance advantages compare source transport with two donor permutations. (c) Effect matching takes the smaller source advantage over those permutations, measured by absolute difference from the target-relational effect. Bars are shared-node and task 95% CIs over 48 cells; filled markers indicate intervals excluding zero.

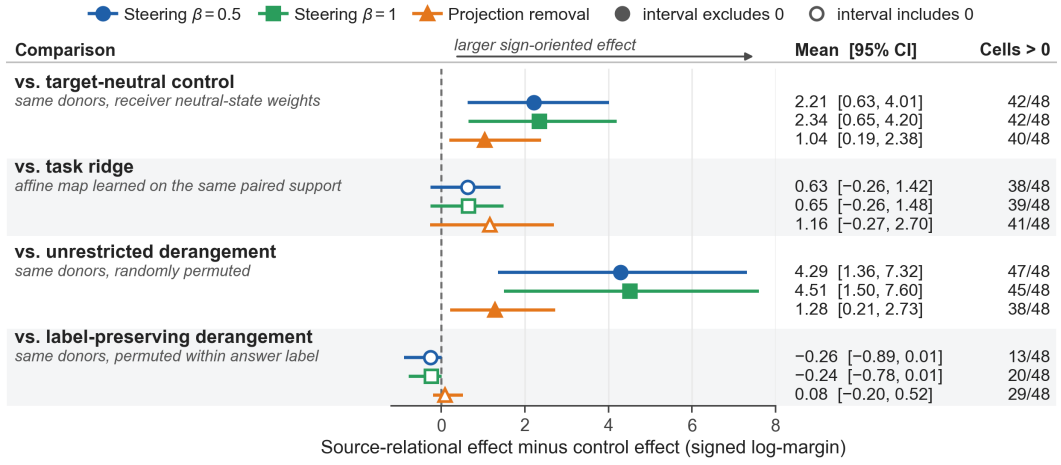


Figure 3: Source-relational effects relative to four matched controls. Points and bars show means and shared-node and task 95% CIs over 48 path-task cells. The right columns report the estimates and counts of positive cells. Positive values denote larger steering effects or greater removal damage. Filled markers indicate intervals excluding zero.

5.3 Receiver response predicts held-out additive steering

A receiver score computed before test outcomes ranks additive effects on disjoint items. Across the source construction and three controls, the correlations are $\rho = 0.837$ at $\beta = 0.50$ and $\rho = 0.825$ at $\beta = 1.00$, with joint audit-test 95% CIs $[0.475, 0.966]$ and $[0.450, 0.961]$ (Table 2). The ordering remains positive after excluding any source, receiver, task, or model together with its incident paths.

Local gain carries most of this ordering signal. Finite-dose stability and off-task response serve as receiver diagnostics, but do not increase aggregate rank correlation. Appendix Table 13 reports the component correlations and exact exclusion minima.

Table 2: Source-derived relations recur and add conditional functional value, while receiver-local response predicts additive steering at both doses. Representation rows report control-minus-source distance to the target-relational reference in pooled, scale-dependent receiver activation units. Source-value rows report source-relational minus target-neutral effect, with positive projection values denoting removal damage. Receiver rows report Spearman ρ with held-out effects.

Evidence	Comparison or endpoint	Estimate	95% CI
Relation recurrence	Unrestricted derangement	$+30.622$	$[2.643, 69.262]$
Fine donor identity	Label-preserving derangement	$+2.366$	$[-0.065, 9.968]$
Source value	Steering $\beta = 0.50$	$+2.213$	$[0.627, 4.009]$
Source value	Steering $\beta = 1.00$	$+2.343$	$[0.648, 4.198]$
Source value	Projection removal	$+1.040$	$[0.187, 2.384]$
Receiver qualification	Steering $\beta = 0.50$	$\rho = .837$	$[0.475, 0.966]$
Receiver qualification	Steering $\beta = 1.00$	$\rho = .825$	$[0.450, 0.961]$
Receiver qualification	Projection removal	$\rho = .645$	$[-0.025, 0.890]$

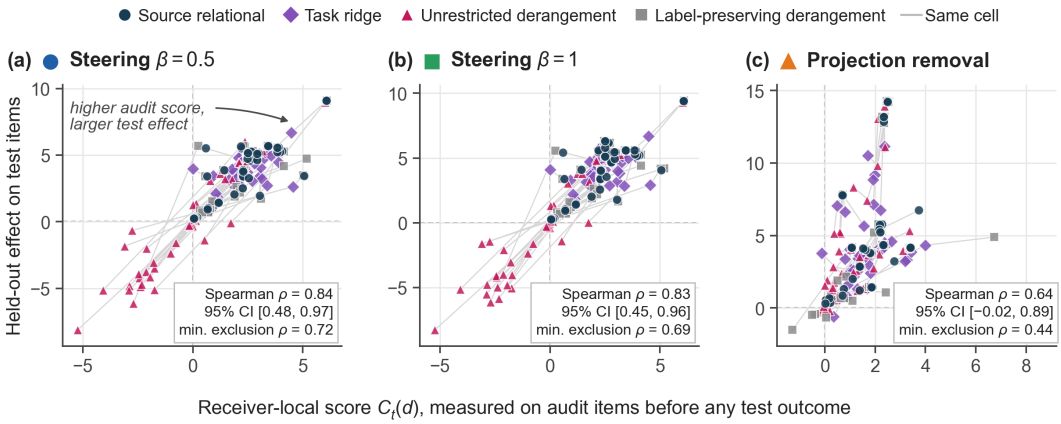


Figure 4: Audit scores predict held-out additive effects across four constructions. Each panel plots the raw scores and effects of 192 method units in 48 path-task cells; thin lines join methods within a cell. Colors and shapes identify methods. Insets report pooled Spearman correlations, joint audit-test 95% CIs, and minimum correlations after source, receiver, task, or incident-model exclusions.

BoolQ and RTE reproduce the direction of the controlled findings in a small public stress test. Across 36 method units in six Qwen-to-receiver task cells, receiver correlations are 0.903, 0.903, and 0.609. Source-relational effects exceed same-access ridge in all six cells at every endpoint. Because all six cells share one source, we treat these public results as descriptive and report every estimate in Appendix F.

5.4 Receiver qualification is operator-specific

Additive steering and projection removal impose different demands on receiver prediction. Steering applies a fixed direction, whereas removal changes the displacement with the input. For removal, source information yields greater margin damage than the target-neutral control, but the score-effect correlation of $\rho = 0.645$ has a joint audit-test CI $[-0.025, 0.890]$. The prospective ranking result is therefore specific to additive steering.

A direct mechanism test measures whether task-matched structure also recurs in local hidden-response kernels. Finite-difference responses correlate with autograd directional derivatives above .996 at both doses. Neither conditioned cross-model response comparison separates from coherent task relabelings: their gaps are .0004 and .0104, both with Holm-adjusted $p = .509$. Appendix E.6 reports the definitions, chain-rule check, and complete results.

5.5 Natural responses carry context-specific source information

Source responses to natural context changes improve prediction beyond the tested label, action, and clean-output baselines. On new scenes and withheld questions, a four-scalar predictor reduces normalized squared error from $.705$ to $.476$, a 32.5% reduction. The difference has a scene-clustered 95% CI of $[-.268, -.193]$ and Holm-adjusted $p < .001$. All twelve task and directed-model combinations improve. Full linear prediction reproduces the gain; ordinary same-information regression performs similarly, at $.477$ (Appendix Table 29).

The predictive information depends on the particular context. Label- and action-matched source outputs from another scene or task raise error to $.929$ or $.940$. A post hoc check retains an aggregate gain when both models answer correctly (Appendix G.6). These results extend context-specific output prediction to the tested encoder-decoder model; hidden-intervention evidence remains receiver specific.

6 Discussion

The results support two uses of cross-model information without a global coordinate map. Source relations construct receiver-native intervention directions, and measured source responses predict receiver output changes. Each use has its own test of functional value.

Receiver-local measurement supports prospective comparison of intervention directions. Existing steering methods estimate directions from demonstrations or contrasts (Todd et al., 2024; Li et al., 2023; Rimsky et al., 2024; Arditi et al., 2024). Cross-model reuse must additionally account for the receiver's response. Across the source construction and three controls, a disjoint audit strongly orders held-out additive effects. This procedure can prioritize candidates when moving an intervention between model families. Its operator-specific performance motivates measuring the response to the intended edit.

Natural-response prediction identifies source information specific to the changed context. Shared-geometry studies motivate searching for recurring structure (Huh et al., 2024; Hu et al., 2026), but high similarity can also reflect common labels or task-independent organization. Here the source response improves prediction beyond known label changes and clean-output action baselines, and matched wrong-scene and cross-task responses lose that advantage. Ordinary regression performs similarly to the structured predictor, locating the gain in the observed source response rather than the fitting algorithm.

The distinction between observable response sharing and hidden intervention response explains the combined findings. Translated activations can change behavior across architectures even when alignment quality weakly predicts behavioral correction (Armstrong et al., 2026). Within-model steering also depends on local geometry and operating point (Wurgaft et al., 2026; Wu et al., 2026). Our factorization condition identifies the additional assumption needed for a representation summary to determine a hidden response. The natural-response result observes the source's changed-context output directly; it does not supply that assumption for an unqueried receiver edit.

The architectural scope depends on the measured function. Hidden intervention qualification covers four decoder-only families; natural-response prediction includes an encoder-decoder model. The latter predicts score changes, including shared errors, rather than improving accuracy. Its correct-answer subset also retains a gain. Together, these findings make source information and its intended functional use separately testable.

7 Conclusion

Cross-model control can reuse relations without aligning hidden coordinates. Source relations recur across twelve directed paths and add held-out effects beyond a receiver-neutral control. Natural-context responses provide additional context-specific predictive information, including for an encoder-decoder model. For hidden interventions, a disjoint receiver audit strongly orders additive steering on held-out items at two test doses. Cross-model reuse is therefore supported by both source information and explicit measurement of its intended functional use.

AI USE STATEMENT

Generative AI tools supported mathematical formalization, experimental code checking, and manuscript polishing. The authors developed the theoretical framework and experimental design, interpreted the results, and verified the citations, equations, code, figures, insights, claims, and prose. The authors take responsibility for all content.

Figure 1 was reconstructed with native PowerPoint shapes and editable text from an AI-assisted layout draft.

ETHICS STATEMENT

The study uses synthetic controlled prompts, public BoolQ and RTE examples, and publicly available model checkpoints. It includes no human-subject experiment and releases no personal data. Activation interventions can alter model behavior, so the evaluation reports off-task response and does not treat measured control as a safety guarantee.

REPRODUCIBILITY STATEMENT

The complete experimental code and key results are included in the Supplementary Material.

Sections 3 and 4 define the estimands, constructions, interventions, splits, and statistical decision rules. Appendices A–H provide task construction, derivations, complete results, sensitivity analyses, and an artifact map. The supplementary code runs in one Colab cell, downloads pinned public models and datasets, generates the controlled prompts, and computes the reported measurements and analyses without author-provided intermediate results. The supplement contains analysis tables and figure inputs. Raw measurements and execution caches are generated by the code and omitted from the upload to meet the size limit.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Marcus Armstrong, Navid Ayoobi, and Arjun Mukherjee. Thinking in different spaces: Domain-specific latent geometry survives cross-architecture translation. *arXiv preprint arXiv:2603.20406*, 2026. URL https://arxiv.org/abs/2603.20406.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Zhimin Hu, Lanhao Niu, and Sashank Varma. Language models represent and transform concepts with shared geometry. *arXiv preprint arXiv:2607.04525*, 2026. URL https://arxiv.org/abs/2607.04525.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 20617–20642, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Thomas Jiralerspong and Trenton Bricken. Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between LLMs. *arXiv preprint arXiv:2602.11729*, 2026. URL https://arxiv.org/abs/2602.11729.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders. *arXiv preprint arXiv:2410.06981*, 2024. URL https://arxiv.org/abs/2410.06981.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 39643–39666, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://doi.org/10.1109/JPROC.2021.3058954.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.15213.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

Yuqi Wu, Shengming Zhao, and Jie Chen. When is a steerable concept representation real? measurement confounds in a cross-family audit of neuroscience parallels in LLMs. *arXiv preprint arXiv:2608.08159*, 2026. URL https://arxiv.org/abs/2608.08159.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint arXiv:2605.05115*, 2026. URL https://arxiv.org/abs/2605.05115.

The appendix follows the evidence from construction to functional use. Appendices A–D specify models, prompts, mathematical definitions, and inference. Appendix E reports relational recurrence, cross-task specificity, matched source value, receiver qualification, and the response-kernel test, followed by robustness and all controlled cells. Appendix F gives the public stress test. Appendix G develops the separate natural-response experiment, including all predictors, controls, and correctness analyses. Appendix H maps these claims to reproducible artifacts.

A Protocol Details

A.1 Models and directed paths

We evaluate four instruction-tuned decoder-only Transformer models. Table 3 records their public repositories. The released access manifest records the resolved repository revisions. Every core model serves as a receiver and every other core model serves as a source, giving $4 \times 3 = 12$ directed source-to-receiver paths. The controlled functional analysis contains all twelve paths and four tasks. The public stress test fixes Qwen2.5 as the source and uses the other three core models as receivers. An ancillary representation-only analysis adds Qwen3.6-27B to test scale sensitivity; it does not enter any intervention or receiver-qualification estimate.

Table 3: Public model repositories and decoder-block sites. The released access manifest records the resolved repository revisions.

Short name	Repository	.625	.750	.875
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	18	21	24
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	20	24	28
Gemma-2-9B	`google/gemma-2-9b-it`	26	32	37
Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.3`	20	24	28
Qwen3.6-27B	`Qwen/Qwen3.6-27B`	40	48	56

All model states are taken at a normalized depth. If a model has L decoder blocks and the requested fraction is q, the site is

$$\ell(q, L) = \mathrm{clip}(\mathrm{round}(qL), 1, L - 1). \tag{9}$$

The central representation and intervention site is $q = 0.75$. Representation diagnostics also use $q \in \{0.625, 0.75, 0.875\}$, and causal-cut sensitivity repeats a fixed six-method subset at 0.625 and 0.875. These alternate sites are sensitivity analyses, not a layer search.

A.2 Splits and information access

Table 4 gives item counts per task. Every item has two surface forms, so activation and scoring rows are twice the item counts. Controlled labels are balanced within every split. Fit and validation items form the donor support, audit items construct receiver-local qualification scores, and test items are held out for the reported causal effects. The full controlled support therefore has $40 + 16 = 56$ donors. Nested representation-only diagnostics use balanced support sizes 8, 24, and 56. The corresponding public supports are 8, 24, and 64.

Table 4: Item counts per task and split. Each item is rendered in two surface forms.

Stratum	Fit	Val.	Audit	Test	Support
Controlled task	40	16	16	40	56
Public task	48	16	32	64	64

The access level differs by comparator and is part of the estimand. Relational transport observes paired source and receiver support innovations. The task ridge observes the same paired support and is fit only on fit and validation items. The target-neutral same-donor control observes receiver baseline states and the same receiver task donors, but not source task innovations. The target-relational construction uses receiver queries and receiver donors and serves as a target-native reference.

A.3 Protocol development and evidentiary roles

The receiver-local qualification formula, audit doses, and decision rule were developed in a separate numeric-task study using pretrained models before outcomes from the present language-task roster were examined. Development-study outcomes do not enter any estimate reported here. In the present experiment, audit items compute the fixed score and disjoint test items provide the intervention effects. We use *prospective* to mean audit-to-test prediction under this fixed protocol.

B Task Generation and Prompts

B.1 Controlled task roster

The controlled tasks cover four kinds of linguistic relation.

1. **Entity binding**: a nonce person is assigned one of two occupations, and the query asks whether a proposed person-occupation pair matches the stated assignment.
2. **Spatial relation**: a pair is placed left or right and the query asks for the stated or opposite relation.
3. **Subject agreement**: a head noun and an intervening distractor require choosing *Are* or *Is*.
4. **Two-hop precedence**: two ordered links must be composed to answer a forward or reversed query.

Nonce names, occupations, nouns, predicates, relation orientation, and query orientation are generated deterministically from the item index before model access. Within each split, lexical or fact factors occur under both labels. Spatial and precedence tasks contain the full fact-orientation by query-orientation factorial. Fit and validation use surface templates 1 and 2, audit uses templates 3 and 4, and test uses templates 5 and 6. This separation prevents an audit or test paraphrase from appearing in model fitting.

Table 5 shows the exact baseline and contextual surface template 1 (S1) pair for each controlled task. Braced symbols denote deterministic item fields. The baseline contains the critical vocabulary but withholds the relation, while the contextual prompt states the relation needed for the answer. Five additional template pairs implement the split-specific paraphrases. The accompanying prompt roster contains every realized string.

Table 5: Exact first surface templates. Baseline and contextual forms retain the critical vocabulary while differing in relational information.

Entity binding
Baseline: The story contains the name {n} and the occupation word {r}, without assigning a job. Is {n} a {q}? Answer Yes or No:
Contextual: Story fact: {n} works as a {r}. Is {n} a {q}? Answer Yes or No:
Spatial relation
Baseline: The scene lists {a}, {b}, and the relation word {r}, but gives no arrangement. Is {a} to the {q} of {b}? Answer Yes or No:
Contextual: Scene fact: {a} is to the {r} of {b}. Is {a} to the {q} of {b}? Answer Yes or No:
Subject agreement
Baseline: Vocabulary only: {h}, {d}, {p}; no sentence structure is specified. Choose a token for a blank. Answer Are or Is:
Contextual: Complete the sentence: The {h} near the {d} __ {p}. Answer Are or Is:
Two-hop precedence
Baseline: Vocabulary only: {l}, {m}, {m}, {r}, precedes, precedes; no ordering is asserted. Does {ql} precede {qr}? Answer Yes or No:
Contextual: Two facts: {l} precedes {m}, and {m} precedes {r}. Does {ql} precede {qr}? Answer Yes or No:

B.1.1 SURFACE-FAMILY SEPARATION

Each controlled task uses six paired baseline and contextual surfaces. Table 5 gives S1 exactly. Fit and validation use S1–S2, audit uses S3–S4, and test uses S5–S6, so no surface template crosses the donor, audit, and test roles. The released prompt roster contains the complete S1–S6 catalog and every realized string. Appendix E separately checks agreement between the two held-out test surfaces and recomputes effects on clean-correct items.

All prompts are wrapped with each model's official chat template. The common system message is "You are a math assistant. Answer with exactly one word." Candidate order is $(\textit{No}, \textit{Yes})$ for entity binding, spatial relation, and precedence, and $(\textit{Are}, \textit{Is})$ for subject agreement. The scorer sums the conditional log probability of every token in the candidate content rather than using only the first tokenizer identifier (ID).

B.2 PUBLIC STRESS TASKS

BoolQ (Clark et al., 2019) and RTE from GLUE (Wang et al., 2018) form a separate stress stratum. Dataset revisions are pinned in the accompanying manifest. Fit and validation items are drawn from each training split. Audit and test items are drawn from each validation split. Selection is deterministic and label balanced before model access, with no origin row shared across child splits. Each item again has two prompt surfaces. BoolQ baselines omit the passage, and RTE baselines omit the premise. Public results remain separate from the controlled confirmatory analysis because they include one source and only six source-receiver-task cells.

The exact public prompt families are as follows, where {p}, {q}, and {h} denote passage, question, and hypothesis.

BoolQ S1 *Baseline:* Question: {q}? No supporting passage is provided. Answer Yes or No:
Contextual: Passage: {p} Question: {q}? Answer Yes or No:

BoolQ S2 *Baseline:* Without a passage, consider the question: {q}? Reply Yes or No:
Contextual: Read this passage: {p} Based on it, is the answer to '{q}' yes? Reply Yes or No:

RTE S1 *Baseline:* Hypothesis: {h} No premise is provided. Does the hypothesis follow? Answer Yes or No:
Contextual: Premise: {p} Hypothesis: {h} Does the hypothesis follow from the premise? Answer Yes or No:

RTE S2 *Baseline:* Consider only this hypothesis: {h} Without a premise, is it entailed? Reply Yes or No:
Contextual: Given '{p}', is '{h}' entailed? Reply Yes or No:

C METHOD DERIVATION

C.1 CONTEXT INNOVATIONS

Let $b^m_{\tau if} \in \mathbb{R}^{d_m}$ be the baseline state of model m for task τ, item i, and surface form f, and let $h^m_{\tau if}$ be the state under the contextual prompt. We first estimate a task-wide displacement using fit items,

$$\mu^m_\tau = \frac{1}{|\mathcal{F}_\tau|} \sum_{(i,f)\in\mathcal{F}_\tau} \left(h^m_{\tau if} - b^m_{\tau if}\right). \tag{10}$$

The item innovation averages the two surfaces after subtracting this displacement,

$$v^m_{\tau i} = \frac{1}{2} \sum_{f=1}^{2} \left(h^m_{\tau if} - b^m_{\tau if} - \mu^m_\tau\right). \tag{11}$$

This decomposition removes a prompt-wide mean shift before cross-model comparison and prevents the two paraphrases from acting as independent observations.

C.2 SIGNED RELATIONAL TRANSPORT

For source s, receiver t, query i, and paired support $\mathcal{D}$, the source computes signed-cosine similarities

$$a_{ij}^s = \frac{\langle v_{\tau i}^s, v_{\tau j}^s \rangle}{\|v_{\tau i}^s\|_2 \|v_{\tau j}^s\|_2}, \qquad w_{ij}^s = \frac{a_{ij}^s}{\sum_{k \in \mathcal{D}} |a_{ik}^s|}. \tag{12}$$

The receiver-native prediction is

$$\widehat{v}_{\tau i}^{s \to t} = \sum_{j \in \mathcal{D}} w_{ij}^s v_{\tau j}^t. \tag{13}$$

The support item IDs and order match exactly across models. Equation 13 transfers weights, not hidden coordinates. It applies the signed relational rule from the main text to the centered innovations in Eq. 11.

We compare four source-paired roles for every directed path. The *relational* arm uses Eq. 13. The *unrestricted donor derangement* keeps the source weights but applies a deterministic fixed-point-free permutation to receiver donors. The *label-preserving derangement* applies a fixed-point-free permutation separately within the two label strata. The *task ridge* learns an affine source-to-receiver map. Its regularization factor is selected from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ on validation items, scaled by the fit Gram trace, and the selected map is refit on fit plus validation items.

Two receiver-native arms clarify source value. The *target-neutral same-donor* arm centers target baseline queries and target baseline support, computes their signed-cosine weights, and combines the same target task donors used by the source arm. The *target-relational same-donor* arm computes both query weights and donors from target innovations. A target-native oracle and an isotropic random direction are retained as diagnostics, but they are not members of the primary four-role comparison.

C.3 FROM DIRECTION ERROR TO FUNCTIONAL ERROR

Let M_t be the receiver's signed decision margin and define $\Delta_t(\beta, d) = M_t(h + \beta d) - M_t(h)$. Taylor's theorem around h gives

$$\Delta_t(\beta, d) = \beta \langle \nabla M_t(h), d \rangle + R_t(d), \qquad |R_t(d)| \le \frac{L_t \beta^2}{2} \|d\|_2^2, \tag{14}$$

when the gradient is L_t-Lipschitz along the edit segment. Subtracting Eq. 14 for $\widehat{d}$ and $\widetilde{d}$, applying the triangle inequality, and using $\|\nabla M_t(h)\|_2 \le G_t$ yields Eq. 6. The first term propagates implemented-direction error through the receiver's local gain. The two remainder terms depend on receiver curvature and intervention scale. The relational bound applies before tangent localization and norm matching, so it does not by itself bound this implemented error.

C.4 WHEN A REPRESENTATION SUMMARY DETERMINES RECEIVER RESPONSE

The representation-to-function gap is an identification problem with an exact linear condition. Fix receiver t, intervention state $h_0 \in \mathcal{H}_t$, and differentiable score map $F_t : \mathcal{H}_t \to \mathbb{R}^c$. Let $J_t = DF_t(h_0)$ be the downstream Jacobian. A linear synthesis map $\Gamma_t : \mathcal{C} \to \mathcal{H}_t$ converts proposal coefficients $z \in \mathcal{C}$ into receiver-native edits, so the local response operator is

$$\mathcal{R}_t = J_t \Gamma_t, \qquad \mathcal{R}_t z = DF_t(h_0)[\Gamma_t z]. \tag{15}$$

Under any invertible hidden-coordinate change $h' = Gh$, the factors become $\Gamma_t' = G\Gamma_t$ and $J_t' = J_t G^{-1}$, while $\mathcal{R}_t$ remains unchanged. The functional object is therefore the composed response, not either hidden-coordinate factor.

Let $S : \mathcal{C} \to \mathcal{Z}$ denote the linear representation summary available without querying receiver response. The summary determines every response in $\mathcal{C}$ through some linear map $A_t : \mathcal{Z} \to \mathbb{R}^c$ if and only if

$$\mathcal{R}_t = A_t S \quad \Longleftrightarrow \quad \ker S \subseteq \ker \mathcal{R}_t. \tag{16}$$

Necessity follows because $Sz = 0$ implies $\mathcal{R}_t z = A_t S z = 0$. For sufficiency, define $\widetilde{A}_t(Sz) = \mathcal{R}_t z$ on $\operatorname{im} S$. If $Sz = Sz'$, then $z - z' \in \ker S \subseteq \ker \mathcal{R}_t$, so this definition is independent of the

representative. Any linear extension of $\widetilde{A}_t$ to $\mathcal{Z}$ gives the required A_t. Equation 16 states the missing assumption precisely: every proposal distinction discarded by the representation summary must also be irrelevant to receiver response.

Finite access to natural states does not supply this condition. Suppose a protocol observes distinct natural states $h_1, \dots, h_n$, their clean scores, and representation statistics, but does not query the response at $h_\star = h_1$ along $d_\star \neq 0$. Choose a with $a^\top d_\star \neq 0$ and define

$$\phi(h) = a^\top (h - h_\star) \prod_{i=2}^{n} \frac{\|h - h_i\|_2^2}{\|h_\star - h_i\|_2^2}. \tag{17}$$

The product equals one when $n = 1$. Then $\phi(h_i) = 0$ for every observed state, while $D\phi(h_\star)[d_\star] = a^\top d_\star \neq 0$. For any differentiable base score map F_0, nonzero $u \in \mathbb{R}^c$, and $\eta > 0$, the two differentiable score maps

$$F_+(h) = F_0(h) + \eta\phi(h)u, \qquad F_-(h) = F_0(h) - \eta\phi(h)u \tag{18}$$

agree on every observed clean score but differ in their local response along $d_\star$ by $2\eta(a^\top d_\star)u$. Thus finite representation and clean-output observations cannot identify an unqueried response over a receiver class that contains both maps. This construction concerns the stated access regime; direct Jacobian-vector products or response queries add the missing information.

C.5 Response queries on a proposal span

Receiver measurement becomes sufficient once it covers the proposal span. Let $\mathcal{D}_Q$ be all directions whose responses must be predicted, let $\mathcal{P}_Q = \operatorname{span} \mathcal{D}_Q$, and let $r = \dim \mathcal{P}_Q > 0$. Assume one noiseless query returns the full vector $J_t d$ for a chosen $d \in \mathcal{P}_Q$. Fewer than r linearly independent queries cannot identify an unrestricted linear $J_t|_{\mathcal{P}_Q}$ in the worst case. Indeed, their span is a proper subspace $\mathcal{U}$. A nonzero linear functional λ that vanishes on $\mathcal{U}$ and a nonzero $u \in \mathbb{R}^c$ define an indistinguishable rank-one response difference $\Delta J(d) = \lambda(d)u$, yet ΔJ is nonzero on at least one direction in $\mathcal{D}_Q$. Conversely, querying any basis $b_1, \dots, b_r$ recovers every response because

$$d = \sum_{j=1}^{r} \alpha_j b_j \quad \Longrightarrow \quad J_t d = \sum_{j=1}^{r} \alpha_j J_t b_j. \tag{19}$$

The exact worst-case query count is therefore r for uniform recovery of an unrestricted linear response operator on the span. The case $r = 0$ is trivial. A single known direction can be queried once, and lower counts can succeed under additional structure on J_t.

Noisy measurements add a conditioning requirement. Write the basis matrix as $B = [b_1, \dots, b_r]$ and $T_t = J_t B$. If row $x_k^\top$ of $X \in \mathbb{R}^{m \times r}$ specifies queried direction Bx_k, then the observations satisfy $Y = XT_t^\top + E$. For full-column-rank X, define the least-squares estimate by $\widehat{T}_t^\top = X^\dagger Y$. It obeys

$$\|\widehat{T}_t - T_t\|_F \leq \|X^\dagger\|_2 \|E\|_F = \frac{\|E\|_F}{\sigma_{\min}(X)}. \tag{20}$$

For a general test edit d, let $d_\parallel = BB^\dagger d = B\alpha$ be its orthogonal projection onto $\mathcal{P}_Q$ and let $d_\perp = d - d_\parallel$. A finite-dose prediction $\beta\widehat{T}_t\alpha$ satisfies

$$\begin{aligned} &\|F_t(h_0 + \beta d) - F_t(h_0) - \beta\widehat{T}_t\alpha\|_2 \\ &\quad \leq |\beta| \|J_t d_\perp\|_2 + |\beta| \frac{\|E\|_F}{\sigma_{\min}(X)} \|\alpha\|_2 + \frac{L_t \beta^2}{2} \|d\|_2^2, \end{aligned} \tag{21}$$

when DF_t is L_t-Lipschitz in the induced operator norm along the edit segment. The three terms measure proposal-span approximation, receiver calibration, and finite-dose remainder. For proposals inside $\mathcal{P}_Q$, $d_\perp = 0$. Reusing one calibrated operator across base states additionally requires control of $|\beta| \|[J_t(h_q) - J_t(h_{\mathrm{cal}})]B\alpha\|_2$. These terms explain why a low-dimensional source proposal can reduce receiver measurements only when it spans the test directions, is well probed, and remains locally stable.

C.6 TANGENT LOCALIZATION AND INTERVENTION OPERATORS

Let $\bar{b}_i = b_i/\|b_i\|_2$ and let $\widehat{v}_i$ be a predicted receiver vector. Additive steering first localizes the vector to the plane orthogonal to the baseline state,

$$u_i = \widehat{v}_i - \langle \widehat{v}_i, \bar{b}_i \rangle \bar{b}_i. \tag{22}$$

Within each receiver-task split, every method is norm matched row by row to a relational reference source selected cyclically from the frozen model order. The reference source therefore does not depend on observed outcomes. For nonzero u_i, the implemented direction is $d_i = (\|u_i^{\mathrm{ref}}\|_2/\|u_i\|_2)u_i$, and the additive intervention is $\delta_i = \pm\beta d_i$ at the final prompt token of the selected decoder block. Audit doses are $\beta \in \{0.25, 0.5\}$ and test doses are $\beta \in \{0.5, 1.0\}$.

Projection removal is input dependent. Define the baseline-tangent contextual innovation

$$r_i = (h_i - b_i - \mu) - \langle h_i - b_i - \mu, \bar{b}_i \rangle \bar{b}_i, \tag{23}$$

and localize the unit prediction $q_i = \widehat{v}_i/\|\widehat{v}_i\|_2$ as

$$\widetilde{q}_i = q_i - \langle q_i, \bar{b}_i \rangle \bar{b}_i. \tag{24}$$

The unscaled removal delta at unit dose is

$$p_i = -\frac{\langle r_i, \widetilde{q}_i \rangle}{\|\widetilde{q}_i\|_2^2} \widetilde{q}_i. \tag{25}$$

For nonzero p_i, the implemented displacement is $\delta_i^{\mathrm{proj}}(\gamma) = \gamma(\|p_i^{\mathrm{ref}}\|_2/\|p_i\|_2)p_i$. A zero reference gives a zero displacement; a degenerate direction cannot be matched to a nonzero reference. Projection audit doses are $\gamma \in \{0.25, 0.5\}$, and the held-out endpoint uses $\gamma = 1$. This is a unit dose of the norm-matched intervention; exact component removal occurs when the norm ratio is one. The reported projection outcome changes sign so that a positive value denotes margin damage under this displacement.

C.7 FULL-STRING CAUSAL OUTCOME

For each surface, let S_c be the teacher-forced log probability sum of candidate $c \in \{0, 1\}$. With binary label y, the signed clean margin is

$$M = (S_1 - S_0)(2y - 1). \tag{26}$$

Each intervention and its zero-delta reference are duplicated within the same forward batch. The causal surface effect is the patched margin minus its paired reference margin. Two surface effects are averaged to the item level, and task summaries average the two label-class means. Additive test outcomes use the symmetric secant

$$E_\beta = \frac{\Delta M(+\beta) - \Delta M(-\beta)}{2\beta}. \tag{27}$$

Projection uses the direct, sign-reversed effect of the norm-matched $\delta_i^{\mathrm{proj}}(1)$. A zero-delta hook is also compared with an unhooked forward on identical tensors, and a compact subset is replayed with fused and single-arm scoring.

D RECEIVER-LOCAL QUALIFICATION AND STATISTICAL CRITERIA

D.1 RECEIVER-LOCAL QUALIFICATION

The receiver-local qualification score is computed only on audit items. For a method, receiver, and task, let $g_{.25}$ and $g_{.5}$ be class-balanced symmetric derivatives at the two audit doses. Define relative instability and off-task response as

$$I = \frac{|g_{.5} - g_{.25}|}{\max(|g_{.25}|, 10^{-12})}, \qquad O = \frac{g_{.25}^{\mathrm{off}}}{\max(|g_{.25}|, 10^{-12})}, \tag{28}$$

where $g_{.25}^{\text{off}}$ is the class-balanced mean absolute response when the same intervention is applied to the other controlled task prompts. The additive score is

$$C_{\text{add}} = \frac{g_{.25}}{1 + I + O}. \tag{29}$$

For projection, $g_{.25}$ is sign reversed so that larger C_{proj} predicts more harm under removal. The output also reports an empirical dose-sensitivity diagnostic, $\min(1, 2\eta|g_{.25}|/\widehat{L})$, where $\widehat{L} = |g_{.5} - g_{.25}|/(0.5 - 0.25)$ and $\eta = 0.25$; it is set to one when $\widehat{L}$ is zero to numerical precision. This quantity summarizes the change between the two measured secants. It is not a certified curvature bound and does not enter Eq. 29.

D.2 Primary uncertainty criteria

The receiver test correlates audit qualification scores with held-out test effects over 192 source-receiver-task-role units, grouped into 48 directed path-task cells. Significance uses $20{,}000$ random permutations of score values within each four-role cell. The three endpoint p-values are corrected by Holm's procedure. A receiver endpoint passes only if Spearman $\rho \geq 0.5$, Holm-adjusted $p < 0.05$, correlations remain positive after removing each source, receiver, and task and all paths incident to each model, and the 2.5th percentile of the disjoint audit/test crossed bootstrap is positive.

Method comparisons operate on 48 paired path-task differences for each endpoint and control. They use $20{,}000$ random sign flips and one Holm family over the three endpoints by three controls. Passing requires a positive mean, adjusted $p < 0.05$, positive leave-axis means, and a positive crossed-bootstrap lower bound. The crossed bootstrap performs $10{,}000$ draws, resampling shared model nodes and tasks. A model draw weights a directed edge by the product of its source and receiver multiplicities, which preserves the dependence created when one model appears in many directed paths. Audit and test items are resampled independently for qualification intervals.

The representation criterion compares Euclidean distance to the target-relational prediction. Its unrestricted advantage is

$$\|\widehat{v}^{\text{der}} - \widehat{v}^{t\to t}\|_2 - \|\widehat{v}^{s\to t} - \widehat{v}^{t\to t}\|_2, \tag{30}$$

and the label-preserving version replaces the unrestricted derangement with the within-label derangement. Identifying fine donor identity requires both comparisons to pass. Functional matching is tested separately by asking whether the source-relational effect is closer to the target-relational effect than both derangement effects. These criteria prevent representation recurrence from substituting for functional matching.

The post hoc cross-task analysis uses the coherent model-wise null in Eq. 32. Its exact upper-tail p-value is the fraction of $13{,}824$ mappings whose mean cosine is at least the identity score. Holm correction covers five profiles, three fixed depths, and two query splits. Label-stratified query resampling measures uncertainty conditional on the fixed model-task roster and does not replace the exact mapping test.

The response-kernel mechanism test uses the three tasks with common No/Yes candidate strings and the same coherent construction, giving $6^3 = 216$ mappings. Its two test-split comparisons form one Holm family. A mechanism claim additionally requires a positive audit gap, positive leave-one-model gaps, and a positive cross-surface gap. Finite differences validate local response measurement but do not change this decision rule.

E Complete Controlled Diagnostics

The complete controlled results follow the same scientific order as the main text: relational recurrence, conditional source value, and receiver-side prediction of held-out effects. Robustness and behavioral analyses then test the stability and interpretation of these findings.

E.1 Relational recurrence and identified resolution

At full support and the central site, the source-relational prediction is closer to the target-relational prediction than unrestricted donor derangement in all 48 directed path-task cells. The pooled raw-

distance advantage is 30.622 in receiver activation units before tangent localization and norm matching, with shared-node and task 95% CI $[2.643, 69.262]$. Its magnitude is scale dependent across receivers. The $48/48$ cellwise sign and dimensionless relation-weight statistics provide the cross-model evidence. Against the label-preserving derangement, the pooled mean advantage is 2.366 with CI $[-0.065, 9.968]$. The experiment therefore supports aggregate recurrence without resolving fine donor identity.

Table 6 gives all nine functional comparisons. Relational transport meets the criterion against unrestricted donor derangement for all three endpoints. Its intervals against same-access ridge and label-preserving derangement include zero.

Table 6: Relational effect minus matched control effect over 48 directed path-task cells. LP denotes the label-preserving control. Positive projection values denote removal damage. Passing uses the conjunction in Appendix D, not the adjusted p-value alone.

Endpoint	Control	Mean	Positive	95% CI	Holm p	Criterion
Steer .5	Task ridge	.630	38/48	[-.256, 1.416]	.000450	No
Steer .5	Unrestricted	4.290	47/48	[1.357, 7.318]	.000450	Yes
Steer .5	LP der.	-.257	13/48	[-.890, .010]	1.000	No
Steer 1	Task ridge	.654	39/48	[-.263, 1.484]	.000450	No
Steer 1	Unrestricted	4.510	45/48	[1.500, 7.597]	.000450	Yes
Steer 1	LP der.	-.239	20/48	[-.782, .008]	1.000	No
Projection	Task ridge	1.157	41/48	[-.272, 2.700]	.000450	No
Projection	Unrestricted	1.282	38/48	[.206, 2.730]	.000450	Yes
Projection	LP der.	.084	29/48	[-.197, .520]	.442	No

E.2 Support, depth, and cross-source stability

Relational recurrence is stable across the three protocol-fixed support sizes. Table 7 shows that weight cosine remains between .821 and .834, flattened weight-rank correlation remains between .670 and .685, and sign agreement remains between .883 and .896. Exact top-donor agreement decreases as the support grows, while the aggregate relation statistics remain stable. This pattern is consistent with recurrence at the level of signed relation structure rather than one exact neighbor. Figure 5 summarizes the support and depth checks together with the analytic error bounds.

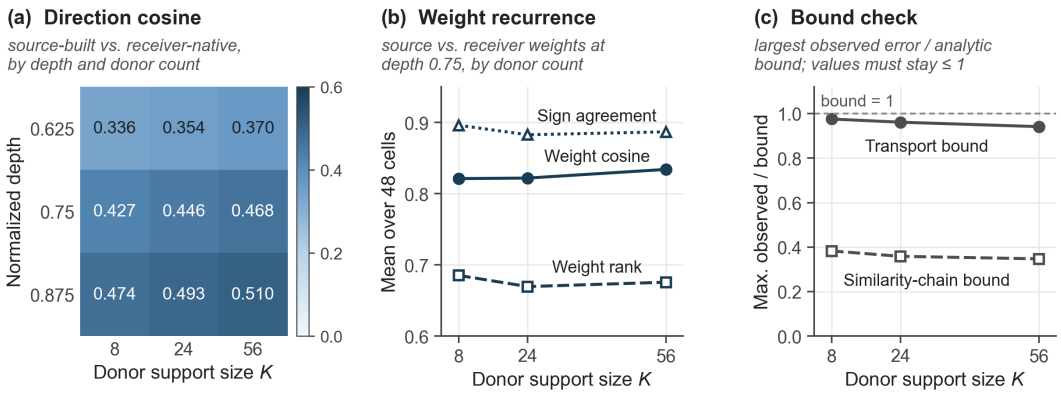


Figure 5: Relational recurrence is stable across protocol-fixed depths and support sizes, and all observed error ratios satisfy the analytic bound. (a) Receiver prediction cosine rises with normalized depth and support size. (b) Relation-weight cosine, rank correlation, and sign agreement remain stable across $K \in \{8, 24, 56\}$ at the central depth. (c) The maximum observed transport and similarity-chain ratios remain below their analytic upper bound.

Table 7: Controlled relation recurrence at normalized depth .75 across donor support sizes. U and LP are pooled receiver-space Euclidean-distance advantages against unrestricted and label-preserving donor derangements. Their raw magnitudes are scale dependent across receivers.

K	Weight cosine	Weight rank	Sign agree.	Top donor	U / LP
8	.821	.685	.896	.263	32.147 / 3.066
24	.822	.670	.883	.237	31.768 / 2.146
56	.834	.676	.887	.159	30.622 / 2.366

The task-level decomposition in Table 8 explains why aggregate recurrence and fine identity must remain separate. Entity binding and spatial relation have the highest weight cosine and sign agreement, yet their exact top-donor agreement is near zero. Subject agreement has lower aggregate recurrence but much higher top-donor agreement and a larger LP distance advantage. The pooled LP interval still crosses zero, so these descriptive task differences do not establish fine donor identity.

Table 8: Task-level recurrence at normalized depth .75 and $K = 56$, averaged over twelve directed model paths. U and LP use pooled, scale-dependent receiver-space Euclidean distance.

Task	Cosine	Rank	Sign	Top donor	U	LP
Entity binding	.939	.780	.974	.021	37.975	.169
Spatial relation	.905	.730	.961	.000	42.601	.029
Subject agreement	.786	.602	.749	.604	11.641	9.362
Two-hop precedence	.707	.590	.864	.010	30.272	-.096

Three external sources independently produce a receiver-native prediction for each receiver-task cell. Their pairwise agreement provides a source-consensus diagnostic that is distinct from fitting a common coordinate map. At the central depth and full support, mean pairwise cosine is .846 and the mean minimum pairwise cosine is .774. Table 9 shows that consensus is strongest for entity binding and precedence and weakest for spatial relation and agreement.

Table 9: Cross-source prediction consensus at normalized depth .75 and $K = 56$, averaged over four receivers. Each receiver-task cell contains predictions from the other three models.

Task	Mean pairwise cosine	Minimum pairwise cosine
Entity binding	.949	.924
Spatial relation	.716	.574
Subject agreement	.730	.618
Two-hop precedence	.987	.981
All tasks	.846	.774

E.3 Cross-task specificity of relation weights

A cross-task analysis tests whether recurrence follows task identity rather than a task-independent relation pattern. It reconstructs Eq. 2 from the saved context innovations at each fixed depth, using $K = 56$ support items and two disjoint query splits. The four tasks share index and label schedules, but those schedules do not assign semantic correspondence across tasks. Similarities are averaged over the six unordered model pairs because relation-weight cosine is symmetric in the two models.

Raw similarities confirm that high recurrence alone is insufficient to establish task specificity. On audit queries at depth .75, entity binding and spatial relation have a cross-task cosine of .9059, close to their respective same-task values of .9389 and .9049. The cross-task value averages both task assignments within each model pair. These descriptive values motivate removing the measured label and neutral-prompt components before testing task matching.

The primary residual removes the registered binary-label subspaces from both axes of each relation matrix. Let $W_{m,\tau} \in \mathbb{R}^{n_Q \times n_D}$ contain the weights for model m and task τ. Let Z_τ^Q and Z_τ^D contain the two query-label and donor-label indicators, with corresponding orthogonal projectors P_τ^Q and P_τ^D. We define

$$R_{m,\tau} = (I - P_\tau^Q) W_{m,\tau} (I - P_\tau^D). \tag{31}$$

This projection removes every component in the stated query-label and donor-label indicator spaces. It does not claim to remove arbitrary nonlinear label information. A second profile projects $\mathrm{vec}\, R_{m,\tau}$ away from the fixed span of relation residuals computed from the corresponding neutral prompts. This step treats neutral-prompt relations as a measured nuisance rather than a zero baseline.

The exact null preserves one task labeling within each model across all of its model pairs. For M models and T tasks, choose permutations $\pi_m \in S_T$, fix π_1 to the identity, and score

$$\mathcal{A}(\boldsymbol{\pi}) = \frac{1}{\binom{M}{2} T} \sum_{m<n} \sum_{\tau=1}^{T} C^{mn}_{\pi_m(\tau),\pi_n(\tau)}, \tag{32}$$

where C^{mn}_{ab} is the mean query cosine between tasks a and b in models m and n. Simultaneously relabeling every model leaves the relative assignment unchanged, so the null contains $(T!)^{M-1}$ distinct mappings. The four-model, four-task analysis therefore enumerates $24^3 = 13{,}824$ mappings. This model-wise construction preserves reverse-edge and cycle consistency that an independently permuted edge does not.

Table 10: Task-matched recurrence after removing registered label and neutral-relation components. Same and cross are mean query cosines over six unordered model pairs at depth .75 and $K = 56$. Confidence intervals use label-stratified query resampling. Exact p values compare the identity score with all 13,824 coherent model-wise relabelings. Holm correction covers the fixed cross-task analysis family.

Profile	Split	Same	Cross	Δ	95% CI	Exact p	Holm p
Label-indicator component	Audit	.7585	.6999	.0585	[.0437, .0766]	.043837	.089193
Label-indicator component	Test	.6910	.6562	.0348	[.0291, .0405]	.029731	.089193
Two-sided label residual	Audit	.2909	.0355	.2554	[.2365, .2743]	.000217	.005425
Two-sided label residual	Test	.3112	.0708	.2404	[.2255, .2556]	.000217	.005425
Neutral-conditioned residual	Audit	.1167	.0028	.1139	[.0763, .1557]	.000362	.006872
Neutral-conditioned residual	Test	.1137	.0166	.0970	[.0731, .1210]	.000940	.008681

Task matching remains after both nuisance removals and replicates across disjoint query splits. The label-indicator component has high same-task and cross-task similarity, and its small gap does not survive the fixed Holm correction. The two-sided residual gap is .2554 on the audit split and .2404 on the test split. Neutral conditioning reduces the magnitude, as expected when measured relation structure is removed, but leaves gaps of .1139 and .0970. All four residual identity scores remain significant after correction. This analysis was specified after the initial controlled outcomes were inspected. We therefore report it as a post hoc analysis replicated on disjoint audit and test query splits, rather than as a preregistered task-population test.

Table 11: Same-task minus cross-task gaps across the three fixed depths. Residual removes the registered query-label and donor-label components. Conditioned additionally removes the neutral-relation span. Every entry has Holm-adjusted $p < .009$.

Depth	Residual audit	Residual test	Conditioned audit	Conditioned test
.625	.2792	.2461	.1627	.1138
.750	.2554	.2404	.1139	.0970
.875	.2576	.2359	.0946	.0873

The aggregate result is broad but not uniform. On the neutral-conditioned test profile, five of six model pairs and three of four tasks have positive same-minus-cross gaps. Two-hop precedence is the only negative task at $-.0066$, while the smallest leave-one-model and leave-one-task aggregates remain positive at .0549 and .0329. Exchanging the two prompt surfaces retains a gap of .0639 with exact $p = .0137$. Restricting the analysis to the three tasks with the same No/Yes answer strings yields a gap of .0348, exact $p = .0231$, and interval $[.0161, .0532]$. At depths .625 and .875, the neutral-conditioned test gaps are .1138 and .0873, with Holm-adjusted $p = .0074$ and .0076. The aggregate gap does not depend on retaining any single model or task and remains positive under the cross-surface, common-answer, and fixed-depth checks.

An ancillary representation-only scale check adds a 27B decoder-only model and expands the comparison from six to ten unordered model pairs. At the central depth, its test two-sided residual gap is .2150 with exact $p = 3.0 \times 10^{-5}$ across 24^4 coherent mappings. The neutral-conditioned gap is .0568 with exact $p = .0057$. This extension supports the representation result only; the four-model roster remains the basis of the intervention and receiver-qualification results.

E.4 Conditional source value and functional matching

Source information has positive conditional value relative to the target-neutral control for all three endpoints. For projection, a positive difference means that removing the source-relational component causes more margin damage. The mean advantages are 2.213, 2.343, and 1.040 for steering at 0.5, steering at 1.0, and projection, respectively, with shared-node and task 95% CIs $[0.627, 4.009]$, $[0.648, 4.198]$, and $[0.187, 2.384]$. The differences are positive in 42, 42, and 40 of the 48 cells, respectively, and every mean remains positive after removing all paths incident to any one model.

The path-level effects show where the aggregate result is strongest and where it is weakest. Gemma and Llama receivers have positive source value in all 12 directed source-task cells for all three endpoints. Mistral is positive in $10/12$, $10/12$, and $11/12$ cells, while Qwen is positive in $8/12$, $8/12$, and $5/12$. The negative cells are concentrated in a small set of Qwen and Mistral conditions. They remain in every primary aggregate and leave-axis calculation.

The complete 48-cell table in Appendix E.8 jointly reports the two representation controls and all three conditional source-value endpoints for every central path-task cell.

The target-relational matching test asks whether the source-relational effect is closer to the target-relational effect than both unrestricted and label-preserving derangements. Its mean minimum closeness advantages are $.0209$, $.0236$, and $-.0960$ for steering 0.5, steering 1.0, and projection. The shared-node and task 95% CIs are $[-.267, .458]$, $[-.240, .452]$, and $[-.358, .174]$. All three include zero, leaving functional matching unresolved.

E.5 Prospective receiver qualification and operator boundary

Table 12 reports the complete receiver-side criterion. Both additive endpoints meet the audit-to-test criterion, including every leave-axis check. Projection has a positive descriptive correlation, but its joint audit-test CI includes zero.

Table 12: Receiver-local qualification over 192 method units. Intervals use the disjoint audit/test crossed bootstrap. The lower block gives the minimum correlations after leaving out one source (LSO), receiver (LRO), task (LTO), or incident model (LIO).

Endpoint	ρ	95% CI	Holm p	Criterion
Steering 0.5	.8365	[.4752, .9662]	.000150	Yes
Steering 1.0	.8254	[.4505, .9607]	.000150	Yes
Projection	.6447	[-.0247, .8899]	.000150	No
Endpoint	**LSO**	**LRO**	**LTO**	**LIO**
Steering 0.5	.8053	.7934	.7187	.7875
Steering 1.0	.7974	.7915	.6933	.7743
Projection	.6189	.5612	.4379	.5897

Table 13 clarifies the score's interpretation. Small-dose gain is the strongest single correlate of both additive endpoints. The composite retains strong held-out ordering, while instability and off-task response act as diagnostics rather than improving aggregate rank correlation.

Two diagnostics separate representation quality from effect size. Source-relational dose-sensitivity radii average $.961$ for steering and $.948$ for projection, reaching the cap in 93.8% and 83.3% of units. Isotropic directions have radii $.372$ and $.347$, near-zero cosine, and steering effects $.010$ and $.012$. The target-native oracle has cosine 1.000 but steering effects 1.933 and 2.068.

Table 13: Spearman correlations between qualification components and held-out effects. Instability and off-task response enter as negative penalties, so positive entries in those columns indicate that smaller penalties align with larger effects.

Endpoint	Gain	$-I$	$-O$	Composite
Steering 0.5	.8600	.2966	.1995	.8365
Steering 1.0	.8752	.2457	.1442	.8254
Projection	.7396	-.2048	-.0279	.6447

E.6 Response-kernel mechanism boundary

A direct first-order test asks whether the task specificity of relation weights is also present in receiver response. The test uses entity binding, spatial relation, and two-hop precedence because they share the No/Yes candidate strings. For receiver t, task κ, and query i, let $\mu_i^{t,\kappa}(h)$ be the full-candidate correct-answer margin at the intervention state and define

$$g_i^{t,\kappa} = \nabla_h \mu_i^{t,\kappa}(h), \qquad K_{ij}^{t,\kappa} = (g_i^{t,\kappa})^\top v_j^{t,\kappa}. \tag{33}$$

The scalar matrix K records the local margin response to each receiver donor innovation. Under an invertible hidden-coordinate change $h' = Gh$, the gradient and donor innovation transform as $g' = G^{-\top} g$ and $v' = Gv$, so $K'_{ij} = (g'_i)^\top v'_j = K_{ij}$. Let $A^{s,\tau}$ denote the source relation matrix after the two-sided label projection and before neutral conditioning. It constructs $d_i = \sum_j A_{ij}^{s,\tau} v_j^{t,\kappa}$. Linearity gives the identity

$$\sum_i D\mu_i^{t,\kappa}[d_i] = \sum_{ij} A_{ij}^{s,\tau} K_{ij}^{t,\kappa} = \langle A^{s,\tau}, K^{t,\kappa} \rangle_F. \tag{34}$$

Thus the receiver kernel supplies precisely the first-order information that representation weights alone omit.

Two complementary comparisons test for a shared task-matched response pattern. The relation-response comparison measures row-wise cosine between the neutral-conditioned version of $A^{s,\tau}$ and a receiver response kernel after the same conditioning. Because $A^{s,\tau}$ and its neutral projector are built from cosine relation matrices, this complete statistic is invariant to orthogonal basis changes and uniform positive scaling, not to arbitrary invertible changes. The response-response comparison measures row-wise cosine between two models' response kernels after removing the span of their neutral-donor response kernels. It acts only on the invariant scalar K matrices and therefore retains invariance to arbitrary invertible hidden-coordinate changes. Both comparisons contrast matched tasks with cross-task pairs under 216 coherent model-wise relabelings.

Table 14: The prespecified response-kernel mechanism tests do not qualify a task-matched advantage over coherent cross-task relabeling. Confidence intervals resample matched test items. The two exact test p values form one Holm family. Audit gaps and leave-axis checks are claim qualifiers.

Comparison	Same	Cross	Δ	95% CI	Exact p	Holm p	Audit Δ
Relation response	.00428	.00384	.00045	[-.01275, .01334]	.4861	.5093	-.00835
Response response	.01009	-.00032	.01042	[-.00801, .02999]	.2546	.5093	-.00657

Neither comparison qualifies a shared task-matched response pattern. Both test intervals in Table 14 cross zero, both audit gaps reverse sign, and the leave-one-model and leave-one-task checks include negative gaps. Removing neutral components is not the sole cause: the unconditioned relation-response and response-response gaps are .00543 and .00317, with Holm-adjusted $p = .6204$ for both. The result leaves positive values inside both confidence intervals, so it does not establish that response recurrence is absent. It shows that representation recurrence alone does not qualify a shared predictive-response kernel on this roster.

The leave-axis and surface checks reach the same boundary. Table 15 reports the minimum test gap after excluding one model or task, the cross-surface gap, and the corresponding unconditioned

Table 15: Sensitivity of the two response-kernel comparisons. LOM and LOT are the minimum test gaps after leaving out one model or one task. Cross-surface evaluates exchanged prompt surfaces, and unconditioned omits neutral-span removal. Negative exclusion gaps prevent either comparison from qualifying.

Comparison	Test Δ	LOM min.	LOT min.	Surface Δ	Raw Δ
Relation response	.00045	-.00615	-.00404	.00538	.00543
Response response	.01042	-.00308	-.00490	-.01134	.00317

comparison. Each prespecified family has a negative exclusion gap, and the response-response comparison also reverses across prompt surfaces.

The implementation diagnostics isolate this scientific boundary from response measurement error. Equation 34 closes across all 72 edge, split, and task entries with maximum absolute error 1.22×10^{-12}. Symmetric finite differences closely track the autograd directional derivatives, with Pearson correlations .99847 and .99635 at doses .25 and .50 and sign agreement .9444 and .9618. These checks validate the gradient, intervention, scorer, and finite-dose calculation for a given direction. They do not supply cross-model response recurrence.

Table 16: Finite differences reproduce the autograd directional derivatives for the tested directions. Each row pools 576 source-receiver-task-item observations. Completed-turn Pearson uses a second full-candidate scoring path.

Dose	n	Pearson	Spearman	Sign agreement	Turn Pearson
.25	576	.99847	.99403	.9444	.99639
.50	576	.99635	.99772	.9618	.99457

The response decomposition shows that the remaining task-specific component is small. Across the 24 model, split, and primary-task cells, the registered binary label-indicator component accounts for a median .999338 of response-kernel energy, with range $[.995716, .999984]$. The remaining two-sided label-residual response component is heterogeneous, even though the corresponding representation residual is task matched. This finding supports the central separation: shared relations can propose a receiver-native edit, while the receiver's downstream Jacobian determines its functional realization.

E.7 Robustness and behavioral interpretation

Table 17 reports clean surface balanced accuracy before intervention. The main statistical analysis retains every protocol-fixed cell, including Mistral subject agreement at .5000 and Llama spatial relation at .5875. No clean-performance threshold removes a model-task cell.

Table 17: Clean surface balanced accuracy on controlled test items.

Model	Entity	Spatial	Agreement	Precedence
Qwen2.5-7B	.9875	.8875	1.0000	.9375
Llama-3.1-8B	.7625	.5875	.8750	.9250
Gemma-2-9B	1.0000	.8750	.9625	1.0000
Mistral-7B	1.0000	1.0000	.5000	.9625

Across all 256 method units, audit representation cosine correlates with steering at 0.5, steering at 1.0, and projection at .412, .438, and .200, respectively. Test representation cosine gives .456, .490, and .241 in the same endpoint order. The maximum relative residual outside the donor span correlates at $-.444$, $-.451$, and $-.477$. These construction-family comparisons include the target-native oracle and isotropic random diagnostic, so they are descriptive rather than a within-method mechanism test. Representation metrics contain functional information but remain weaker than the receiver-local qualification score for additive steering.

E.7.1 DEPTH AND CLEAN-ANSWER SENSITIVITY

Table 18 compares the central 0.75 cut with the two fixed auxiliary cuts over 96 cell-method units per endpoint. All rank correlations are positive, and sign agreement ranges from .917 to .979. The stronger agreement at 0.875 does not select that depth as a replacement. The main result remains defined at 0.75.

Table 18: Sensitivity of central effects to two protocol-fixed causal cuts.

Cut	Endpoint	Spearman ρ	Sign agreement
.625	Steering 0.5	.694	.927
.625	Steering 1.0	.693	.917
.625	Projection	.581	.917
.875	Steering 0.5	.901	.958
.875	Steering 1.0	.927	.958
.875	Projection	.948	.979

E.7.2 PROMPT-SURFACE AND DISCRETE BEHAVIOR CHECKS

Two prompt-surface checks leave the primary interpretation unchanged. Table 19 recomputes each role on the 42 cells where both prompt forms are answered correctly and both label classes remain. Source-relational effects stay close to their same-cell full estimates, and per-item sign agreement is .942, .953, and .888 across the three endpoints.

Margin effects also produce visible decision changes. Table 20 reports surface-level balanced top-one accuracy and clean-to-intervened transitions. Source-relational steering raises accuracy from .891 to .940 and .939 at the two doses. Projection removal lowers it to .720, which is the expected direction because the projection endpoint measures damage after removing a useful component. Surface counts repeat receiver items across source paths and are therefore descriptive rather than independent sample sizes.

Table 19: Correct-only and paraphrase diagnostics for the primary roles. "Full" is recomputed on the same cells for which both label classes retain at least one clean-correct item. "Abs. gap" is the mean absolute cellwise difference, not the difference between displayed aggregate means. These 42-cell subsets do not replace the 48-cell primary analysis.

Endpoint	Role	Cells	Full	Correct-only	Abs. gap	Sign agree.
Steer .5	Relational	42/48	4.122	4.132	.093	.942
Steer .5	Task ridge	42/48	3.467	3.479	.070	.955
Steer .5	Unrestricted	42/48	-.273	-.265	.056	.949
Steer .5	Label-preserving	42/48	4.364	4.353	.087	.961
Steer 1	Relational	42/48	4.328	4.351	.080	.953
Steer 1	Task ridge	42/48	3.645	3.660	.063	.961
Steer 1	Unrestricted	42/48	-.324	-.316	.049	.953
Steer 1	Label-preserving	42/48	4.550	4.548	.074	.966
Projection	Relational	42/48	5.246	5.413	.225	.888
Projection	Task ridge	42/48	3.976	4.114	.171	.869
Projection	Unrestricted	42/48	3.825	3.921	.178	.785
Projection	Label-preserving	42/48	5.149	5.320	.248	.843

The capability-conditioned subset provides a final descriptive check. Thirty-one path-task cells contain at least one item on which the source answers both surfaces correctly and the receiver does not. Relational transport exceeds unrestricted donor assignment by 4.069, 4.046, and $.695$ for the two steering doses and projection, with positive-cell fractions .871, .871, and .774. The corresponding label-preserving differences are $-.679$, $-.608$, and $.461$, so this subset retains the primary boundary on fine donor identity.

Table 20: Surface-level discrete behavior for the primary roles. Help denotes clean-incorrect to correct, and harm denotes clean-correct to incorrect. Projection applies removal, so harm is the expected direction of its sign-oriented endpoint.

Endpoint	Role	Clean acc.	Post acc.	Help	Harm	Net
Steer .5	Relational	.891	.940	211	24	187
Steer .5	Task ridge	.891	.936	174	4	170
Steer .5	Unrestricted	.891	.877	80	134	-54
Steer .5	Label-preserving	.891	.946	228	20	208
Steer 1	Relational	.891	.939	246	62	184
Steer 1	Task ridge	.891	.948	232	15	217
Steer 1	Unrestricted	.891	.784	106	519	-413
Steer 1	Label-preserving	.891	.945	268	62	206
Projection	Relational	.891	.720	99	759	-660
Projection	Task ridge	.891	.791	73	458	-385
Projection	Unrestricted	.891	.786	79	485	-406
Projection	Label-preserving	.891	.722	109	760	-651

E.8 Complete controlled cell results

The following table reports every central path-task cell behind the recurrence and conditional source-value summaries. Rows are descriptive. Inference remains at the shared-node and task level defined in Appendix D.

Table 21: Complete central controlled cells. U and LP are the unrestricted and label-preserving receiver-space Euclidean-distance advantages before tangent localization and norm matching; their absolute magnitudes are not comparable across receivers. $\Delta_{.5}$, Δ_1, and Δ_P are source-relational minus target-neutral signed log-margin effects. Positive steering values favor source-relational transport; positive projection values denote greater removal damage. Per-cell rows are descriptive and receive no cellwise significance test.

Source	Receiver	Task	U	LP	$\Delta_{.5}$	Δ_1	Δ_P
Qwen2.5	Llama3.1	Entity	3.486	0.065	1.178	1.172	0.172
Qwen2.5	Llama3.1	Spatial	5.444	0.066	4.447	4.364	1.305
Qwen2.5	Llama3.1	Agreement	2.077	2.448	0.320	0.345	0.755
Qwen2.5	Llama3.1	Precedence	3.868	-0.049	0.919	0.896	0.263
Qwen2.5	Gemma2	Entity	129.436	0.526	1.174	1.238	0.016
Qwen2.5	Gemma2	Spatial	154.154	-0.240	4.725	4.818	0.174
Qwen2.5	Gemma2	Agreement	33.750	24.449	3.131	3.357	6.245
Qwen2.5	Gemma2	Precedence	125.317	-0.384	0.434	0.487	0.248
Qwen2.5	Mistral	Entity	2.823	-0.014	3.803	4.488	2.384
Qwen2.5	Mistral	Spatial	2.635	-0.006	4.203	4.658	6.959
Qwen2.5	Mistral	Agreement	0.596	0.580	0.888	0.895	0.598
Qwen2.5	Mistral	Precedence	0.520	-0.023	6.445	6.227	0.680
Llama3.1	Qwen2.5	Entity	16.932	0.014	3.406	4.002	0.581
Llama3.1	Qwen2.5	Spatial	30.403	0.303	2.773	2.823	-0.125
Llama3.1	Qwen2.5	Agreement	2.685	1.586	-0.197	-0.179	-0.209
Llama3.1	Qwen2.5	Precedence	10.458	0.038	-0.380	-0.247	-0.222
Llama3.1	Gemma2	Entity	126.944	0.668	0.695	0.735	0.019
Llama3.1	Gemma2	Spatial	115.836	-0.028	2.196	2.270	0.113
Llama3.1	Gemma2	Agreement	40.774	36.439	0.164	0.143	0.180
Llama3.1	Gemma2	Precedence	130.956	-0.303	0.351	0.436	0.220
Llama3.1	Mistral	Entity	3.799	0.028	3.363	3.964	2.349
Llama3.1	Mistral	Spatial	3.285	0.018	-1.424	-0.685	6.607
Llama3.1	Mistral	Agreement	0.902	0.727	-0.290	-0.258	-0.315
Llama3.1	Mistral	Precedence	0.705	-0.022	6.088	5.935	0.608
Gemma2	Qwen2.5	Entity	17.336	0.143	4.116	4.749	0.614
Gemma2	Qwen2.5	Spatial	32.672	-0.019	6.125	6.259	-0.056
Gemma2	Qwen2.5	Agreement	5.258	2.546	0.988	0.991	1.248
Gemma2	Qwen2.5	Precedence	16.121	0.052	-0.262	-0.134	-0.058
Gemma2	Llama3.1	Entity	4.785	0.028	1.166	1.172	0.181
Gemma2	Llama3.1	Spatial	4.904	0.007	4.422	4.352	1.306
Gemma2	Llama3.1	Agreement	2.755	3.609	0.194	0.241	0.455
Gemma2	Llama3.1	Precedence	3.986	-0.033	0.939	0.914	0.258
Gemma2	Mistral	Entity	3.293	0.023	3.791	4.468	2.288
Gemma2	Mistral	Spatial	3.589	-0.018	4.205	4.647	6.949
Gemma2	Mistral	Agreement	0.947	0.900	0.398	0.424	0.244
Gemma2	Mistral	Precedence	0.713	-0.024	6.439	6.220	0.677
Mistral	Qwen2.5	Entity	20.508	0.101	4.080	4.731	0.580
Mistral	Qwen2.5	Spatial	26.319	0.337	6.072	6.239	-0.102

Continued on next page

Table 21 continued

Source	Receiver	Task	U	LP	$\Delta_{.5}$	Δ_1	Δ_P
Mistral	Qwen2.5	Agreement	1.744	0.070	0.541	0.537	0.561
Mistral	Qwen2.5	Precedence	14.528	0.025	-0.409	-0.282	-0.252
Mistral	Llama3.1	Entity	4.000	0.062	1.175	1.174	0.173
Mistral	Llama3.1	Spatial	4.404	0.006	4.423	4.360	1.317
Mistral	Llama3.1	Agreement	3.854	3.264	0.286	0.273	0.481
Mistral	Llama3.1	Precedence	1.518	-0.060	0.925	0.897	0.255
Mistral	Gemma2	Entity	122.354	0.382	1.148	1.227	0.023
Mistral	Gemma2	Spatial	127.570	-0.074	4.720	4.811	0.168
Mistral	Gemma2	Agreement	44.352	35.727	1.915	1.832	2.730
Mistral	Gemma2	Precedence	54.578	-0.369	0.437	0.500	0.258

F PUBLIC STRESS TEST

The public stratum tests whether the controlled pattern remains visible on BoolQ and RTE. Across 36 method units in six Qwen-to-receiver task cells, receiver-local qualification correlates with the two additive endpoints at .903 and with projection at .609. Relational recurrence is weaker than in the controlled tasks but increases with support: weight cosine rises from .362 at $K = 8$ to .450 at $K = 64$. Because all six cells share one source, these results provide a directional stress test rather than a second confirmatory estimate.

The public results retain the main directional contrasts. Source-relational minus target-neutral effect is positive in all six cells at all three endpoints. At $K = 64$, the mean representation advantage over unrestricted derangement is 9.622, and source-relational effects exceed task ridge and unrestricted derangement in all six cells at every endpoint. Label-preserving differences are less consistent, with positive fractions of $5/6$, $4/6$, and $4/6$. Target-relational functional matching is also descriptive, with mean minimum closeness advantages .168, .167, and .458. Every corrected six-cell comparison has Holm-adjusted $p = .140625$, so the stress test does not change the controlled identification boundary. Representation advantages are measured in each receiver's activation units before localization and norm matching, so their absolute magnitudes are not compared across receivers. Tables 23 and 24 report the complete cellwise estimates.

The public tasks provide a separate descriptive cross-task comparison. Table 22 applies the two-sided label and neutral-span projections to all six unordered pairs among the four core models. Same-task recurrence exceeds the BoolQ-to-RTE comparison on both query splits. With two tasks, the coherent model-wise null has only $2^3 = 8$ mappings, so its smallest attainable exact p is .125.

Table 22: Descriptive public cross-task specificity at $K = 64$. Values average six unordered model pairs. The coherent exact p is .125 for every row because only eight model-wise relabelings exist.

Profile	Split	Same	Cross	Δ
Two-sided label residual	Audit	.3606	.0081	.3524
Two-sided label residual	Test	.3837	-.0037	.3874
Neutral-conditioned residual	Audit	.3175	.0036	.3139
Neutral-conditioned residual	Test	.3333	-.0030	.3363

Table 23: Complete public six-cell recurrence and source-value diagnostics. U and LP are representation-distance advantages. Δ columns are source-relational minus target-neutral signed log-margin effects. Positive projection values denote greater removal damage.

Receiver	Task	Clean	Weight cos.	U	LP	$\Delta_{.5}$	Δ_1	Δ_P
Llama	BoolQ	.883	.368	.719	.660	1.415	1.389	1.900
Llama	RTE	.773	.623	2.542	.514	1.818	1.813	1.353
Gemma	BoolQ	.898	.350	20.627	12.738	1.915	1.928	2.078
Gemma	RTE	.828	.542	31.808	17.563	2.191	2.227	2.377
Mistral	BoolQ	.844	.364	.759	.669	2.070	2.092	3.302
Mistral	RTE	.797	.452	1.273	.484	3.466	3.542	5.450

Table 24: Source-relational receiver-local qualification scores and held-out effects in the public six-cell stratum. These values do not enter the controlled criterion.

Receiver	Task	C_{add}	C_{proj}	$E_{.5}$	E_1	E_{proj}
Llama	BoolQ	.574	1.193	1.146	1.137	2.415
Llama	RTE	.780	1.287	1.538	1.529	1.970
Gemma	BoolQ	.759	.620	1.280	1.295	3.376
Gemma	RTE	.811	1.510	1.734	1.782	3.433
Mistral	BoolQ	.889	1.412	1.768	1.789	6.736
Mistral	RTE	1.274	2.035	3.044	3.150	7.417

G Predicting Natural Context Responses

Corresponding source responses contain predictive information beyond known actions, correct labels, and baseline outputs. This experiment tests that information directly: given how a source model's four-candidate scores change after a natural context edit, can a fitted predictor estimate the receiver's score change on a new scene? The source-response increment is positive in both prespecified predictor families and all twelve task-direction cells. We first specify the observable response and its information requirements, then report the complete comparisons and correctness strata.

G.1 Generated scenes and natural actions

The task construction provides exact edits with known answer changes. In object tracking, four people initially hold four distinct objects and then perform three exchanges. A question asks for the final holder of one object. An edit swaps two initial holdings while retaining the subsequent exchanges and the question. In ordering, three pairwise relations determine the order of four distinct markers. A question asks which marker occupies a given rank. An edit exchanges two marker roles throughout the relations. These are generated task instances with exact symbolic answers, separate from the public stress tests in the main experiment.

Every scene supports the same eight nonidentity actions. Six actions exchange one of the six unordered pairs of four roles. The other two apply $A = (0, 1)$ followed by $B = (1, 2)$, or B followed by A. These compositions differ because the transpositions share a role. Including the base context gives nine contexts for each scene, question, and form. In object tracking, we index roles before the fixed sequence of subsequent exchanges and pull scores back through those exchanges. This makes the action permutations identical across scenes. Ordering already uses the same role indexing.

The evaluation separates new scenes, new lexical identities, and questions absent from calibration. Each task has 32 fit scenes with question indices 0 and 1, and 64 test scenes with all four question indices. The fit and test sets draw names, object names, and marker names from disjoint interleaved vocabulary pools. Each scene samples four distinct entities without replacement. The two surface forms alternate with scene index in both splits. The first 16 test scenes per task also receive the other form. Thus, the second-form comparison changes the wording of the same scene using another known renderer. The exact fit and test prompt sets have no intersection.

The sample contains 768 semantic groups per model, where a group fixes a scene, question, and form. The count is $2(32 \cdot 2 + 64 \cdot 4 + 16 \cdot 4)$, with nine contexts per group. We evaluate Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, and the encoder-decoder FLAN-T5-XL (Qwen Team, 2024; Jiang et al., 2023; Chung et al., 2024), giving 20,736 model-prompt rows. All six directed model pairs are included. The primary test has 128 distinct task-scenes, with dependent questions, actions, and model directions within each scene.

The prompts make the edited relation explicit while keeping the response format fixed. Table 25 gives exact base strings. In the object example, the first single action exchanges the initial possessions of Luma and Fara while retaining the other possessions and all three later exchanges. The correct option remains 2. In the ordering example, exchanging the blue and lilac marker roles changes the correct option from 4 to 2. These examples show why the analysis includes both label-preserving and label-changing edits.

Table 25: Exact base prompts from the generated natural-response test set. Both examples ask question 2 under the primary form. Every natural action retains the options and final answer instruction.

Objects. Four people hold one object each. Luma holds the opal pin. Fara holds the denim pouch. Yara holds the marble cube. Sela holds the jade coin. Then Luma and Fara swap objects. Then Luma and Yara swap objects. Then Yara and Sela swap objects. Who holds the jade coin at the end? Options: 1. Luma 2. Yara 3. Sela 4. Fara Solve the problem carefully. Return only one option number: 1, 2, 3, or 4.
Ordering. Four markers occupy distinct positions from earliest to latest. hazel marker is earlier than jade marker. jade marker is earlier than lilac marker. lilac marker is earlier than blue marker. Which marker is position 3? Options: 1. jade marker 2. blue marker 3. hazel marker 4. lilac marker Solve the problem carefully. Return only one option number: 1, 2, 3, or 4.

G.2 Four-event scores and three-dimensional responses

All models are scored on the same four native answer events. We use each decoder model's official chat template and the full assistant-content events for option numbers 1 through 4, excluding the common template trailer. Qwen receives the system instruction "You are a careful reasoning assistant." Mistral receives the user message alone. FLAN-T5 receives the prompt directly and is scored from its native decoder start. When a candidate has a shared content prefix, its log probability includes that prefix. Prompts are not truncated, and scores retain the native model probabilities before normalization over the four candidates.

The response coordinates remove a common score offset without fitting a cross-model basis. Let $\ell_m(x) \in \mathbb{R}^4$ contain the four log probabilities after mapping option numbers to the scene's canonical roles. Define $C_4 = I_4 - \mathbf{1}\mathbf{1}^\top/4$ and $z_m(x) = C_4\ell_m(x)$. We use the fixed orthonormal Helmert contrast matrix

$$H = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} \\ -1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} \\ 0 & -2/\sqrt{6} & 1/\sqrt{12} \\ 0 & 0 & -3/\sqrt{12} \end{pmatrix}, \qquad q_m(x) = H^\top z_m(x) \in \mathbb{R}^3. \tag{35}$$

Here $H^\top H = I_3$ and $HH^\top = C_4$, so the three coordinates preserve every centered four-event score difference. They are output contrasts, independent of the models' hidden dimensions.

Known action and label changes define baselines for the natural response. Let $T_a x$ be the naturally edited context and let P_a be its canonical role permutation, with $P_a e_j = e_{a(j)}$. For model m, define

$$\begin{aligned} N_m(x,a) &= q_m(T_a x) - q_m(x), \\ B_m(x,a) &= H^\top(P_a - I_4)z_m(x), \\ L(x,a) &= H^\top\big(e_{y(T_a x)} - e_{y(x)}\big). \end{aligned} \tag{36}$$

N_m is the measured natural score response. B_m permutes the base scores according to the known action, and L records the change in the generator's correct answer. All three vectors lie in the same fixed three-dimensional space. This construction lets the comparison ask what the observed source response adds after the task's action and label information are already available.

The prediction contract allows the same labels and action maps to every fitted method. At test time, target-only predictors receive $q_t(x)$, P_a, and the correct roles in both contexts. Source-base predictors additionally receive $q_s(x)$. Source-response predictors additionally receive $q_s(T_a x)$. The changed receiver output $q_t(T_a x)$ is reserved for evaluation. Correct roles are supplied by the task

Table 26: Fixed natural-response predictors. Symbols shorten the complete result tables. A scalar coefficient multiplies an entire three-dimensional vector, while a full linear coefficient is a 3×3 matrix. All fitted methods have zero intercept.

Symbol	Predictor	Inputs	Parameters
E_t	Scalar target	B_t, L	2
E_b	Scalar source base	B_t, L, B_s	3
E_n	Scalar source response	B_t, L, B_s, N_s	4
F_t	Full linear target	B_t, L	18
F_b	Full linear source base	B_t, L, B_s	27
F_n	Full linear source response	B_t, L, B_s, N_s	36
O	Ordinary output regression	$q_t(x), q_s(x), q_s(T_a x), B_t, B_s, L$	54
C	Direct source response	N_s	0

generator, so the label features constitute known-answer access. The comparison measures conditional prediction with this access, including the additional source evaluation on the changed context.

G.3 Eight fixed predictors and their relation

Two predictor families test the same source-response increment with different coefficient sharing. The scalar family forms a linear combination of its three-dimensional input vectors, using one scalar coefficient per vector. For example,

$$\widehat{N}_t = \alpha_t B_t + \alpha_L L + \alpha_s B_s + \alpha_N N_s. \tag{37}$$

The four coefficients are shared across the three output coordinates. Consequently, the prediction transforms with a common orthogonal change of output basis. The full linear family instead assigns an unrestricted 3×3 matrix to each input vector. Table 26 specifies all eight methods and their parameter counts.

Calibration uses the same observations for every fitted predictor. We solve unregularized least squares separately for each task and directed model pair, using 32 fit scenes, two questions, and six single actions. The scalar family fits $1{,}152$ scalar coordinate observations, corresponding to the full linear family's 384 three-dimensional response observations. Fitting excludes compound actions, second-form pairs, and test receiver responses. We use no additional feature centering, rank selection, or regularization search. Numerical rank follows the least-squares solver's machine-precision criterion.

Ordinary output regression contains the full linear source-response predictor. To see this, write the latter as

$$\widehat{N}_t = W_t B_t + W_L L + W_s B_s + W_N N_s. \tag{38}$$

Since $N_s = q_s(T_a x) - q_s(x)$, ordinary regression produces the same prediction with coefficient blocks

$$[0,\, -W_N,\, W_N,\, W_t,\, W_s,\, W_L]\,. \tag{39}$$

The fitted ordinary model has no larger calibration squared error in any of the twelve cells. Its primary test error is 0.4771, close to the four-scalar model's 0.4760. The full linear source-response model has error 0.4924. The useful comparison is therefore the source-response increment within each family. The four-scalar construction attains nearly the same aggregate error as the 54-parameter ordinary regression on this sample.

The response features also specify the identity action exactly. Setting $P_a = I_4$ and retaining the base context gives $B_t = B_s = L = N_s = 0$. All response-feature predictors then return zero. Ordinary output regression can return a nonzero value because its base-score inputs remain nonzero. This identity-input calculation uses the saved base outputs and adds no natural-context measurement.

G.4 Primary estimand, uncertainty, and complete results

The primary endpoint measures prediction on new scenes and questions absent from calibration. It uses the primary form, question indices 2 and 3, and all six single actions. For an endpoint set $\mathcal{I}$,

Table 27: Both prespecified source-response contrasts pass the scene-clustered primary test. Differences are candidate minus source-base normalized error. Intervals resample scenes within each task; Holm correction covers the two contrasts.

Contrast	Difference	95% interval	Holm p
E_n vs. E_b	-0.2291	[-0.2680, -0.1934]	0.00020
F_n vs. F_b	-0.2387	[-0.2793, -0.2021]	0.00020

Table 28: Both source-response contrasts improve in every task and model direction on held questions with single actions. Values are normalized response squared error (lower is better). Q denotes Qwen2.5-7B-Instruct, M denotes Mistral-7B-Instruct-v0.3, and F denotes FLAN-T5-XL.

Task	Direction	E_t	E_b	E_n	F_t	F_b	F_n	O	C
Ordering	$F \to M$	1.067	1.096	0.887	1.085	1.111	0.899	0.873	0.818
Ordering	$F \to Q$	0.440	0.450	0.425	0.443	0.455	0.431	0.425	0.953
Ordering	$M \to F$	1.183	1.196	1.000	1.207	1.232	1.033	1.015	6.960
Ordering	$M \to Q$	0.440	0.418	0.384	0.443	0.430	0.396	0.400	1.013
Ordering	$Q \to F$	1.183	1.197	1.137	1.207	1.228	1.164	1.158	36.614
Ordering	$Q \to M$	1.067	1.112	1.043	1.085	1.133	1.079	1.119	4.574
Objects	$F \to M$	0.960	0.945	0.392	0.995	0.989	0.391	0.372	0.388
Objects	$F \to Q$	0.893	0.886	0.380	0.928	0.925	0.411	0.379	0.403
Objects	$M \to F$	0.891	0.891	0.361	0.944	0.947	0.372	0.331	0.507
Objects	$M \to Q$	0.893	0.893	0.453	0.928	0.931	0.474	0.429	0.539
Objects	$Q \to F$	0.891	0.890	0.395	0.944	0.949	0.427	0.358	0.646
Objects	$Q \to M$	0.960	0.953	0.482	0.995	1.017	0.522	0.486	0.662

normalized response squared error is

$$R(\widehat{N}) = \frac{\sum_{(x,a,s,t)\in\mathcal{I}} \|N_t(x,a) - \widehat{N}_{t\leftarrow s}(x,a)\|_2^2}{\sum_{(x,a,s,t)\in\mathcal{I}} \|N_t(x,a)\|_2^2}. \tag{40}$$

This is a ratio of pooled sums, with lower values indicating more accurate score-response prediction. The zero-response predictor has $R = 1$. Each primary scene contributes six model directions, two questions, and six actions, for 72 dependent records and $9{,}216$ records in total.

Inference keeps all observations from the same scene together. We bootstrap the 64 test scenes within each task $10{,}000$ times and recompute the ratio of sums. For each prespecified contrast, we also sum the paired squared-error difference within each task-scene and apply $10{,}000$ random sign flips to these 128 cluster differences. The one-sided sign-flip test assumes that cluster differences are sign-symmetric under its null. We correct the two primary comparisons, E_n versus E_b and F_n versus F_b, by Holm's method. A source-response increment requires negative error differences, upper 95% bootstrap bounds below zero, and both adjusted p values below .05.

Both primary comparisons support an incremental contribution from the corresponding source response. Scalar prediction error falls from 0.7051 to 0.4760, a 32.5% reduction. Full linear error falls from 0.7311 to 0.4924, a 32.6% reduction. Table 27 gives the risk-difference intervals and adjusted tests. Both sign-flip results reach the resolution of the $10{,}000$ resamples, with Holm-adjusted $p \simeq 0.00020$. Table 28 shows that both contrasts improve in every task and directed model pair.

The source-response increment also persists in the other prespecified test settings. Table 29 reports all eight combinations of form, question family, and action family for every method. On held questions with compound actions, scalar error falls from 0.6779 to 0.4354. On the second form with held questions and single actions, it falls from 0.7355 to 0.5012. Compound-input prediction receives the source model's measured response to that compound input. It tests response prediction on actions absent from fitting, while the second-form comparison tests the paired alternative wording of the same scenes.

Table 29: Source responses improve prediction across all eight test settings. Each entry is normalized response squared error (lower is better). Fit questions are question indices used in calibration, evaluated on new test scenes; held questions were not used in calibration. Method symbols are defined in Table 26.

Form	Questions	Actions	E_t	E_b	E_n	F_t	F_b	F_n	O	C
Primary	Fit	Single	0.578	0.572	0.385	0.598	0.596	0.401	0.379	0.963
Primary	Held	Single	0.706	0.705	0.476	0.726	0.731	0.492	0.477	1.239
Primary	Fit	Compound	0.549	0.540	0.357	0.564	0.564	0.375	0.347	0.923
Primary	Held	Compound	0.682	0.678	0.435	0.688	0.692	0.450	0.427	1.204
Second	Fit	Single	0.574	0.569	0.379	0.598	0.598	0.392	0.371	0.933
Second	Held	Single	0.743	0.735	0.501	0.779	0.782	0.525	0.505	1.231
Second	Fit	Compound	0.528	0.520	0.344	0.551	0.548	0.357	0.331	0.892
Second	Held	Compound	0.653	0.645	0.417	0.676	0.676	0.433	0.406	1.250

Table 30: Replacing the corresponding source context raises response-prediction error. Both controls match the question, form, canonical correct role, and action maps. Each comparison covers all 9,216 directed primary action records from 128 scenes.

Method	Same context	Other scene	Other task
E_t	0.706	0.706	0.706
E_b	0.705	0.714	0.720
E_n	0.476	0.929	0.940
F_t	0.726	0.726	0.726
F_b	0.731	0.747	0.755
F_n	0.492	0.957	0.972
O	0.477	0.895	0.941
C	1.239	1.892	1.953

G.5 Matched source-context controls

Corresponding context information accounts for the source-response advantage under matched action and answer structure. The other-scene control replaces all source base and changed-context fields with those from another scene in the same task. Donors are sorted by scene index and selected by a fixed cyclic shift within the same question, form, and canonical base-answer stratum. The other-task control uses the corresponding question, form, and canonical base-answer stratum in the other task, cycling through its sorted donor list. Both controls preserve the complete action maps and all nine canonical correct answers. Target fields and fitted coefficients stay fixed.

Replacing the corresponding source context increases error for both source-response families. Table 30 gives all eight methods on the primary endpoint. The scalar source-response error rises from 0.4760 to 0.9294 with another scene and 0.9401 with another task. The target-only methods remain unchanged because they do not use any replaced field. These comparisons show that the observed gain depends on the corresponding source context beyond the matched labels and action maps.

The controls cover every test group, with reuse visible at the group level. Table 31 reports all form and question combinations. Single and compound actions share those mappings. In the primary endpoint, other-scene donors are one-to-one over the selected groups. Other-task donors can serve up to three target groups. Because model directions, actions, and reused donors are dependent, these controls are descriptive comparisons without additional item-level significance tests. Cross-task pairing matches the specified factors without assuming a semantic correspondence between unrelated scene indices.

G.6 Correctness, label strata, and task capability

The predictive increment is present in correct-response subsets as well as in shared errors. Table 32 reports all primary-endpoint label and correctness strata. When the target answers both base and changed contexts correctly, scalar error falls from 0.2629 to 0.2311. When source and target

Table 31: Context controls have complete coverage, with donor reuse reported explicitly. The last two columns give distinct donor groups / maximum target groups per donor. Single and compound actions use the same group mappings. A group is one scene, question, and form, independent of model direction and action.

Form	Questions	Task	Groups	Other scene	Other task
Primary	Fit	Ordering	128	128/1	117/2
Primary	Fit	Objects	128	128/1	116/2
Primary	Held	Ordering	128	128/1	105/3
Primary	Held	Objects	128	128/1	110/3
Second	Fit	Ordering	32	32/1	32/1
Second	Fit	Objects	32	32/1	32/1
Second	Held	Ordering	32	32/1	31/2
Second	Held	Objects	32	32/1	32/1

Table 32: Pooled tasks: complete primary-endpoint label and correctness strata. Values are normalized response squared error (lower is better); n counts directed action records, not independent scenes. Correctness strata are descriptive post-result analyses with unchanged fitted predictors.

Condition	Stratum	n	E_b	E_n	F_b	F_n	O
Correct label	Changes	4608	0.671	0.409	0.697	0.425	0.397
	Preserved	4608	0.857	0.771	0.881	0.789	0.829
Target after action	Correct	3076	0.417	0.303	0.430	0.313	0.306
	Wrong	6140	1.016	0.663	1.056	0.686	0.661
Target before / after	Correct / correct	1896	0.263	0.231	0.271	0.238	0.249
	Correct / wrong	1116	1.091	0.730	1.113	0.746	0.702
	Wrong / correct	1180	0.798	0.481	0.824	0.500	0.448
	Wrong / wrong	5024	0.990	0.640	1.036	0.665	0.647
Source / target after	Correct / correct	1722	0.512	0.287	0.527	0.301	0.279
	Correct / wrong	1354	0.970	0.919	1.000	0.938	0.950
	Wrong / correct	1354	0.311	0.320	0.323	0.327	0.337
	Wrong / wrong	4786	1.029	0.595	1.071	0.618	0.584

both answer the changed context correctly, it falls from 0.5124 to 0.2871. Label-preserving and label-changing edits both retain the increment. These results establish that the pooled improvement extends beyond examples with incorrect target answers.

Correctness stratification also reveals which forms of agreement contribute to the pooled result. When both models answer the changed context incorrectly, scalar error falls from 1.0286 to 0.5946, contributing 71.9% of the total absolute squared-error reduction. When the source is wrong and the target is correct, error increases slightly from 0.3108 to 0.3205. Thus, the predicted quantity includes systematic mistakes as well as correct score changes. The correctness strata were computed after the primary analysis with all fitted predictors unchanged. They are descriptive analyses without additional hypothesis tests.

Correctness-conditioned gains differ across tasks (Tables 33 and 34). In ordering, error rises from 0.2565 to 0.2768 when the receiver answers the changed context correctly, whereas object tracking improves. Restricting both models to correct changed-context answers retains a gain in both tasks. These differences locate the source of the pooled prediction gain without changing any fitted predictor or primary observation.

Native task accuracy supplies a separate interpretation of the response metric. Table 35 reports all model, task, question, and action-family accuracies, computed by the four-candidate argmax. Pooled over the two held questions, object-tracking accuracy with single actions ranges across models from 25.0% to 29.8%, near the 25% random-choice reference. Ordering accuracy on the same question family ranges from 30.6% to 44.5%. Predicting these models' score changes measures their observed behavior, including errors. Increasing task accuracy is a distinct objective.

The natural-response evidence identifies a concrete source of conditional predictability across model architectures. Corresponding source outputs improve target response prediction after action, la-

Table 33: Ordering: complete primary-endpoint label and correctness strata. Values are normalized response squared error (lower is better); n counts directed action records, not independent scenes. Correctness strata are descriptive post-result analyses with unchanged fitted predictors.

Condition	Stratum	n	E_b	E_n	F_b	F_n	O
Correct label	Changes	2304	0.491	0.442	0.502	0.453	0.453
	Preserved	2304	0.761	0.705	0.771	0.720	0.721
Target after action	Correct	1780	0.256	0.277	0.264	0.284	0.286
	Wrong	2828	1.130	0.951	1.148	0.971	0.969
Target before / after	Correct / correct	1134	0.200	0.213	0.204	0.218	0.225
	Correct / wrong	618	1.512	1.252	1.518	1.265	1.236
	Wrong / correct	646	0.548	0.606	0.573	0.624	0.601
	Wrong / wrong	2210	1.018	0.862	1.039	0.885	0.891
Source / target after	Correct / correct	888	0.277	0.268	0.286	0.278	0.270
	Correct / wrong	892	0.993	1.030	1.005	1.042	1.070
	Wrong / correct	892	0.241	0.284	0.247	0.288	0.298
	Wrong / wrong	1936	1.202	0.909	1.223	0.934	0.917

Table 34: Objects: complete primary-endpoint label and correctness strata. Values are normalized response squared error (lower is better); n counts directed action records, not independent scenes. Correctness strata are descriptive post-result analyses with unchanged fitted predictors.

Condition	Stratum	n	E_b	E_n	F_b	F_n	O
Correct label	Changes	2304	0.868	0.373	0.912	0.395	0.336
	Preserved	2304	2.250	1.723	2.474	1.790	2.397
Target after action	Correct	1296	0.880	0.378	0.912	0.398	0.366
	Wrong	3312	0.925	0.431	0.982	0.456	0.413
Target before / after	Correct / correct	762	0.697	0.355	0.740	0.374	0.417
	Correct / wrong	498	0.817	0.391	0.849	0.409	0.354
	Wrong / correct	534	0.978	0.391	1.004	0.410	0.339
	Wrong / wrong	2814	0.967	0.447	1.034	0.474	0.436
Source / target after	Correct / correct	834	0.879	0.317	0.903	0.336	0.294
	Correct / wrong	462	0.911	0.624	0.987	0.665	0.632
	Wrong / correct	462	0.883	0.620	0.946	0.642	0.654
	Wrong / wrong	2850	0.926	0.409	0.981	0.431	0.388

bel, and base-output information is supplied, including on questions and compound inputs absent from calibration. This output-level result complements the separate hidden-response test in Appendix E.6. It supports the value of measured source responses while leaving hidden-state execution and receiver-query reduction as distinct questions.

Table 35: Natural four-candidate accuracy (%) on all primary test questions. Each row uses 64 scenes, giving 64 base, 384 single-action, and 128 compound-action events. Questions 2 and 3 are withheld from response-predictor fitting. The random-choice reference is 25%.

Task	Model	Question	Base	Single	Compound
Ordering	F	0	79.7	74.0	73.4
		1	65.6	65.9	55.5
		2	46.9	40.6	41.4
		3	40.6	40.9	42.2
	M	0	40.6	40.9	46.9
		1	40.6	45.1	46.9
		2	35.9	37.5	32.0
		3	20.3	23.7	25.8
	Q	0	90.6	90.1	89.8
		1	50.0	44.8	49.2
		2	18.8	16.4	25.0
		3	65.6	72.7	71.1
Objects	F	0	29.7	28.9	25.8
		1	29.7	29.7	31.2
		2	31.2	30.7	27.3
		3	26.6	28.4	35.2
	M	0	31.2	29.9	23.4
		1	25.0	24.7	28.1
		2	18.8	25.0	24.2
		3	26.6	25.0	28.9
	Q	0	31.2	33.3	29.7
		1	39.1	34.4	39.1
		2	32.8	31.8	32.8
		3	28.1	27.9	34.4

H REPRODUCIBILITY

The controlled experiments use seed 20270227, pinned model and dataset revisions, bfloat16 (BF16) inference without quantization, and official model chat templates. They produce a fixed prompt roster, per-model context innovations, every audit and test prediction, same-forward paired candidate scores, item-level causal effects, aggregate tables, and bootstrap intervals. The controlled intervention analysis contains 17,792 item-level effect records, 192 primary source-receiver-task-role units, 48 directed path-task cells, and 48 crossed-bootstrap interval records. The cross-task analysis reuses archived support, audit, and test innovations. The response-kernel analysis adds receiver gradients and finite-difference evaluations for the same fixed tasks and splits. The natural-response experiment adds separate context and candidate-score arrays, fit coefficients, held-out predictions, and scene-level errors as specified in Appendix G.

The released single-cell program starts from pinned public checkpoints and datasets, generates the controlled prompts, and measures all activation fields and candidate scores. It then computes receiver-local qualification scores, interventions, cross-task comparisons, response-kernel diagnostics, and natural-response predictions. It requires no author-provided activation fields or fitted predictors. A fresh run saves prompt rosters, raw paired scores, predictions, item-level effects, and aggregate analyses. The submitted supplement contains the code, analysis results, and full-precision figure and table inputs; large raw arrays and execution caches are regenerated by the code.

The Supplementary Material includes `figure_source/Figure_1.pptx`, the editable PowerPoint source for Figure 1. Its source CSV records schematic drawing coordinates, not measured activations.

H.1 IMPLEMENTATION CONSISTENCY

Implementation checks confirm that the recorded intervention effects do not depend on the scoring path. Zero-delta hooks match unhooked forwards, reference scores remain fixed across intervention

frames, and a compact 126-row replay gives identical paired-candidate and decision-margin effects under fused and single-arm scoring. These checks do not add scientific samples.

Table 36: Implementation consistency checks. Values are maximum absolute differences over the indicated roster.

Check	Roster	Maximum difference
No-op hook vs. unhooked	All audit and test frames	.000000
Reference across frames	All audit and test frames	.000000
Fused vs. single candidate effect	126 replay rows	.000000
Fused vs. single margin effect	126 replay rows	.000000

H.2 Artifact-to-claim map

Table 37 links each evidence question to its principal analysis output. Within the supplement's `results/` directory, the controlled, cross-task, response-kernel, and natural-response analyses use the prefixes `core/`, `cross_task/`, `response/`, and `natural_response/`. The analysis tables and item-level errors provide the reported estimates, uncertainty intervals, and control comparisons.

Table 37: Evidence-to-artifact map. Paths are relative to the supplementary `results/` directory.

Evidence question	Principal outputs
Prompts and splits	`core/results/natural_prompt_rows.csv`
Relational recurrence	`core/results/controlled_relation_weight_recurrence.csv`
Cross-task specificity	`cross_task/neutral_baseline_split_depth_metrics.csv`, `cross_task/controlled_split_depth_metrics.csv`
Matched source value	`core/results/source_conditional_value_gates.csv`
Functional matching	`core/results/target_relational_functional_gates.csv`
Receiver qualification	`core/results/receiver_external_gates.csv`, `core/results/crossed_bootstrap_intervals.csv`
Response-kernel mechanism	`response/results/co_primary_tests.csv`, `response/results/finite_difference_pooled_diagnostics.csv`
Behavior and public stress	`core/results/discrete_behavior_diagnostics.csv`, `core/results/public_receiver_statistics.csv`
Natural-response prediction	`natural_response/analysis/primary_inference.csv`, `natural_response/analysis/matched_controls.csv`
Correctness strata	`natural_response/analysis/descriptive_correctness_strata.csv`
Implementation consistency	`core/results/fused_single_parity.csv`