ARROWS RANK, PRODUCTS RELATE: HOW TO READ A TRANSFER MATRIX

Anonymous authors
Paper under double-blind review

ABSTRACT

We say that a son resembles his father, not that a father resembles his son: similarity has a direction, and psychology traced it to the items rather than to what they share. Interpretability reads such directions from transfer matrices: when an intervention learned on condition A moves condition B more than the reverse, an arrow is read as a mechanism flowing from A to B. We show that a transfer matrix is a ranking of the conditions times a symmetric relation, and that its arrows show almost only the ranking: across 152 published and re-measured matrices and Taskonomy, about 3% of arrow triples form a cycle, against 13% if pair-specific effects were as strong as the ranking. In interchange intervention the ranking is set by the recipient, whose own sentences fix how far the intervention pushes. In the syntax matrices we re-measured, the same push both ways removes more than half of the asymmetry on average, and where the push varies across coordinated verb phrases, it carries most of transfer's link to human acceptability. What two conditions share is read instead from the two-way product of transfers, which recovers known groupings and shows that rare syntactic constructions share no more with frequent ones than with the rest. An arrow can say that B receives a lot, almost never that B relies on A in particular.

1 INTRODUCTION

Interpretability often asks whether two behaviors of a model share a mechanism, and a common test is transfer: learn an intervention on one condition, apply it to another, and measure how much the model changes. Boguraev et al. (2025) ask this of seven English filler–gap constructions, in which a phrase is moved away from the position where it is interpreted. Each construction comes as pairs of sentences such as "The video showed that the employee replaced", which the model continues with "her", and "The video showed who the employee replaced", which it continues with a period because the object has moved to the front. For each construction they learn a direction in the model's hidden states that switches one continuation for the other, apply it to every other construction's sentences, and record the change in log-odds. The result is a 7×7 *transfer matrix* $T(A{\to}B)$, whose diagonal $T(A{\to}A)$ is a construction's effect on itself.

Such matrices are read through their asymmetries. In the release, the direction learned on non-finite embedded wh-questions moves pseudo-clefts by 5.72, while the direction learned on pseudo-clefts moves non-finite embedded wh-questions by only 0.76; we follow this pair throughout and call it the running pair. Drawn as arrows from A to B whenever $T(A{\to}B) > T(B{\to}A)$, the matrix becomes a network, which the authors read as frequency: rare constructions rely on more frequent, linguistically similar ones. Two constructions did not fit, and the authors explained them linguistically: the rare topicalization receives little, and the frequent relative clause gives little. Matrices of this kind are drawn and read the same way for languages, tasks, steering vectors and circuits (§3). Yet a transfer from A to B mixes something about A, something about B and something the two share, and an arrow does not say which of the three made it.

Psychology met the same problem with similarity. We say that a son resembles his father, not that a father resembles his son (Tversky, 1977). The frequency reading says the same of a rare construction and a frequent one. Shepard (1957) and Luce (1963) modeled such asymmetric data as a similarity shared by both items, times a bias toward one of them, so that every asymmetry sat in the biases, which psychologists could shift but only infer from the judgments they explained (Luce, 1963;

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

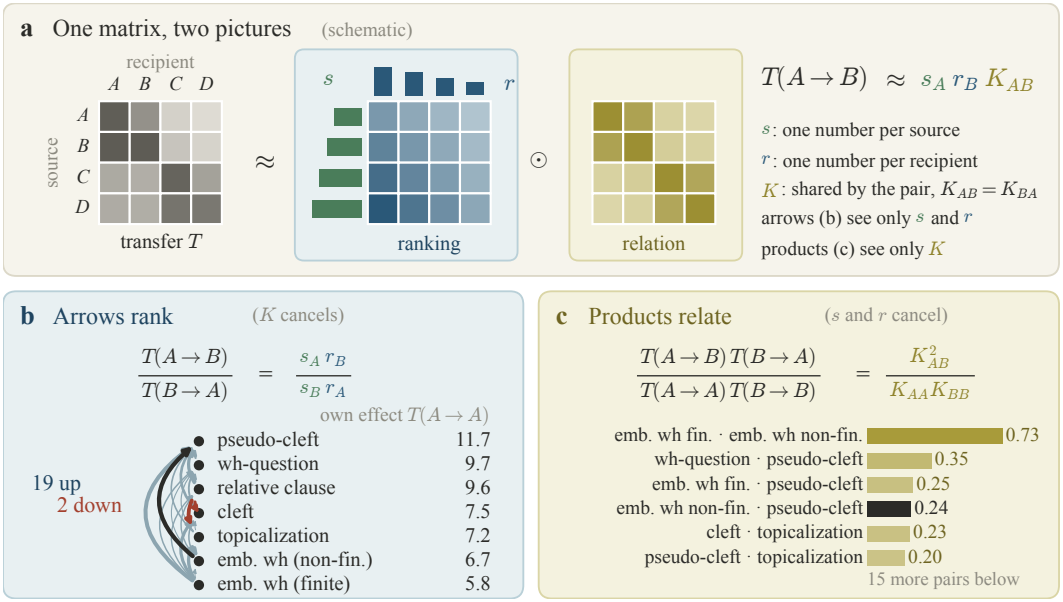


Figure 1: **A transfer matrix is a ranking times a symmetric relation.** (a) Schematic: a transfer $T(A{\to}B)$, the effect on B of an intervention learned on A, is roughly one number per source (s) times one per recipient (r) times a symmetric relation K (Eq. 1). (b) The ratio of the two transfers cancels K, so arrows follow one number per condition. Below, a released matrix of Boguraev et al. (2025) (Pythia-1.4B), constructions ordered by their own effect $T(A{\to}A)$, a stand-in for the recipient's number (§3): 19 of its 21 arrows point up, toward the larger own effect, and 2 down (red); black: the running pair, non-finite embedded wh-question to pseudo-cleft, 5.72 vs 0.76. (c) The product of the two transfers over the two diagonals cancels s and r. Below, the same matrix read by the relation, the square root of this ratio (Eq. 3): its six most related pairs, the running pair (black) fourth.

Nosofsky, 1991). We argue that a transfer has the same form, with a number for the source as well as for the recipient,

$$T(A{\to}B) \;\approx\; s_A \cdot r_B \cdot K_{AB}, \qquad K_{AB} = K_{BA}, \tag{1}$$

known in statistics as quasi-symmetry (Caussinus, 1965). In the ratio of the two transfers K cancels and each condition keeps one number, s/r, so the arrows can only form a ranking. In the product of the two transfers over the two diagonals, s and r cancel and the shared part remains, a symmetric relation (Figure 1). Arrows can say that B receives a lot, almost never that B relies on A in particular.

For interchange intervention, a standard test of shared mechanisms, we can say what sets the ranking. The intervention sets the recipient's coordinate along the source's direction to that of the recipient's other sentence, which is steering along that direction, except that the experimenter does not choose how far to push. That push, the *step*, is the distance between the recipient's own two sentences along the source's direction. It is roughly the recipient's *volume*, the same distance along its own direction, times the overlap of the two directions. The overlap is the same both ways and belongs to K, so the recipient's number r follows its volume. Rare constructions tend to have larger volumes, so the arrows tend to point at them, and a large step reads as a strong link from that particular source: an illusion of amount, akin to the dormant-pathway illusion of subspace patching (Makelov et al., 2024). Fixing the step leaves a smaller ranking, set at the source by how central its direction is, its mean $|\cos|$ with the other directions.

We make three contributions. (1) Arrows rank: across 152 published and re-measured matrices and Taskonomy's vision tasks (Zamir et al., 2018), about 3% of arrow triples cycle (§3). (2) The recipient sets the step: in interchange intervention its volume, measured in advance, predicts 77–97% of the syntax arrows, the same step both ways removes 55% of the asymmetry, and the step changes a published reading of human acceptability (§4). (3) Products relate: two-way products read what conditions share and show that rare constructions do not rely on frequent ones in particular (§5).

Table 1: **The quantities.** g, b: the two sentences of a pair; u_A: the unit direction learned on A.

Quantity	Definition	Where
Ranking share	share of the log two-way ratios fit by one number per condition	§3
Relation of A, B	$\sqrt{T(A{\to}B)\,T(B{\to}A)/T(A{\to}A)\,T(B{\to}B)}$, what A and B share	Eq. 3, §5
Step $A{\to}B$	median $\lvert(g-b)\cdot u_A\rvert$ over B's evaluation pairs, what A writes into B	Eq. 4
Volume ν_B of B	median $\lvert(g-b)\cdot u_B\rvert$ over B's development pairs, with no transfer	§4.2
Own effect of B	the diagonal $T(B{\to}B)$: volume times steepness, the effect per unit of step	§4.2
Centrality of A	mean $\lvert\cos\rvert$ of u_A with the other conditions' directions	§4.4

2 Transfer Matrices and How They Are Read

For conditions A and B, $T(A{\to}B)$ is the effect of an intervention learned on A and applied to B, and an arrow points from A to B when $T(A{\to}B) > T(B{\to}A)$. Under Eq. 1, the log ratio of the two transfers is a difference of one number per condition,

$$\log \frac{T(A{\to}B)}{T(B{\to}A)} = \big(\log s_A - \log r_A\big) - \big(\log s_B - \log r_B\big), \tag{2}$$

a potential that cannot form cycles (Bradley & Terry, 1952; Jiang et al., 2011). The share of the log ratios that one number per condition explains, fitted by least squares, is the *ranking share*. The product of the two transfers over the two diagonals removes s and r entirely,

$$\mathrm{rel}(A,B) = \sqrt{\frac{T(A{\to}B)\,T(B{\to}A)}{T(A{\to}A)\,T(B{\to}B)}} = \frac{K_{AB}}{\sqrt{K_{AA}K_{BB}}}, \tag{3}$$

which is the similarity estimate of Shepard (1957) and Luce (1963). We call it the *relation*. Both readings use pairs whose transfers are positive. It equals 1 for a condition with itself and falls toward 0 as two conditions share less, whatever their own sizes.

In interchange intervention, the running study (Boguraev et al., 2025) uses distributed alignment search (Geiger et al., 2024): for each construction a rank-one direction u is fitted on the construction's own sentence pairs. To transfer A's direction into B, take one of B's sentence pairs, b without a gap and g with one. At one token, the coordinate of b's representation along u_A is replaced by that of g, and the readout is the change in log-odds between the two continuations. Each pair is also used the other way round, and $T(A{\to}B)$ averages over pairs. Table 1 lists the quantities we read.

Table 2 lists the data. *Released:* the original authors released 60 matrices, for three Pythia models (Biderman et al., 2023), ten intervention positions and animate and inanimate constructions, each cell the maximum over layers (layer-max). Their pipeline has two indexing problems affecting four of the seven constructions (Appendix A.4). Pairs untouched by these problems follow the larger diagonal as often as all pairs (79%), so the problems do not create the pattern, and we re-measure with them corrected. *Re-measured:* every layer of 14 model–data combinations in the syntax families of Table 2 (NPI: negative polarity items). The refusal models are Llama-3-8B-Instruct and Gemma-2-9B-it (Gemma Team et al., 2024) with six categories of harmful requests (Arditi et al., 2024). *Other groups:* every square transfer matrix with at least three conditions printed in seven papers (Leng & Xiong, 2025; Tang et al., 2024; Xue et al., 2026; Liu et al., 2026; Nguyen et al., 2026; Ball et al., 2026; Shao et al., 2026), and Taskonomy's task-transfer win rates. Counts over layers use *layer groups*: one layer of one animacy half. Protocols: Appendix A.

3 Arrows Rank

Across every transfer matrix we could find, the arrows follow one number per condition: a pure ranking never forms a cycle among three conditions, and only about 3% of triples do. We show this first on the running study's released matrix and then on all the others.

3.1 A network is a queue

Sorting the seven constructions by their own effect turns the released network into a queue: 19 of its 21 arrows point to the construction with the larger diagonal (Figure 1b), and only 27 of the 5,040

Table 2: **Arrows almost never cycle.** Cycles: triples of conditions whose three arrows form a cycle, among triples with no tied pair. Ranking share (§2): median over matrices, on pairs with both directions positive (in parentheses: after random sign flips, near $2/n$ for n conditions; other groups: 61 matrices with enough such pairs). Taskonomy: win rate over scratch training (16k).

Source	Intervention and conditions	Matrices	Cycles	Ranking share
Seven other papers	neurons, vectors, circuits	68	34/972 = 3.5%	0.82 (0.40)
Boguraev et al. (2025)	interchange; filler–gap	60	78/2100 = 3.7%	0.91 (0.29)
Re-measured	interchange (layer-max)	24	23/894 = 2.6%	0.93 (0.29)
original, single-clause	5 models, two animacy halves	10	14/350 = 4.0%	0.93 (0.29)
original, multi-clause	3 models, two animacy halves	6	5/210 = 2.4%	0.92 (0.28)
independent sentences	Kumon & Yanaka (2026); 2 models	2	0/70 = 0%	0.97 (0.29)
NPI licensing	eight structures; 4 models	4	4/224 = 1.8%	0.78 (0.25)
refusal	six harm categories; 2 models	2	0/40 = 0%	0.99 (0.33)
Taskonomy	frozen encoder; 22 vision tasks	1	52/1540 = 3.4%	0.96 (0.09)
Compare: random arrows			25%	
Compare: pair-specific effects as large as the ranking (simulated)			13%	

orderings of the seven constructions put 19 or more arrows that way. Across the release, 79% of the 1,260 arrows point to the larger diagonal, against 50% with shuffled diagonals, and re-measured with the corrected pipeline, the network has no cyclic triple at all (Appendix B.1). The diagonal stands in here for the recipient's number, which §4 identifies.

3.2 Every matrix, two readouts

The queue is not special to this matrix. Table 2 applies two readouts to every matrix we could find: the share of triples whose arrows cycle, and the ranking share, the fraction of the log ratios that one number per condition explains. Every row of Table 2 cycles in at most 4.0% of triples, and no single study in more than 10.5%, against 25% for random arrows, and one number per condition explains a median 0.78 to 0.99 of the log ratios, against 0.09 to 0.40 after random sign flips. In a simulation, 3% of cyclic triples corresponds to a pair-specific part carrying about a tenth of the variance, and a part as large as the ranking gives 13% (Appendix B.3). The pattern is not specific to language models: in Taskonomy's 22 vision tasks, 3.4% of triples cycle and one number per task explains 0.96 of the log ratios. Eq. 1 also predicts left-out transfers (median R^2 0.78–0.90, and 0.42 in other groups' matrices) and fails on random numbers (Appendix B.3).

3.3 What a ranking cannot say

A ranking describes each condition on its own: B receives a lot, and A pushes hard. To say that B relies on A in particular, the arrows would need a pair-specific part, and that part is small, about a tenth of the variance. A single dependence barely moves the cycle count or the ranking share, but it should show in the pair's residual, the part of its log ratio that the ranking misses, or in its relation. We planted dependences by adding half of a rarer construction's learned direction to a more frequent one's, for three pairs in each of three models. All seven planted dependences whose transfers stayed positive (of nine) raised the residual or the relation, five raised both, and four became their matrix's largest residual (Appendix B.4).

4 What Sets the Ranking: The Recipient's Step

The arrows form a ranking, but a ranking does not say what its number is, and a reliance ordered by one property, such as rare constructions relying on frequent ones, would itself be a ranking. In interchange intervention the number is set by the recipient, through how far the intervention pushes: fixing the push shrinks the arrows, whereas a dependence of one construction on another would not.

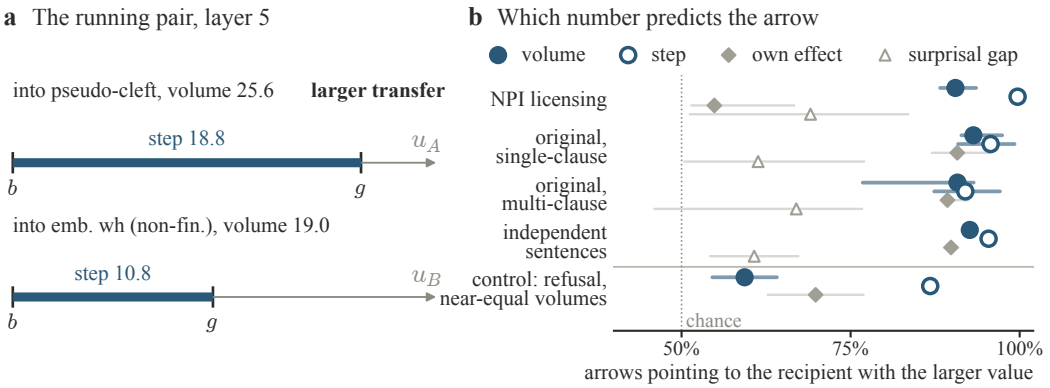


Figure 2: **A recipient's volume, measured before any transfer, predicts the arrows.** (a) The running pair, re-measured at layer 5 of Pythia-1.4B, drawn to scale: each direction writes a step equal to the distance between the recipient's two sentences, b (no gap) and g (gap), along the source's direction (Eq. 4). (b) Among pair–layers whose two transfers differ by more than 0.25 log-odds, the share of arrows pointing to the recipient with the larger volume, step received, own effect or surprisal gap (how much more surprising the gapped sentence is; §4.2), per family: median over models (marker) and range (line). Refusal control: nearly equal volumes, so volume nears chance while the step still names most arrows. Per model: Table 12.

4.1 Interchange intervention is steering with a recipient-set step

Replacing the coordinate of b along the unit direction u_A by that of g is the same as adding u_A with a step (Makelov et al., 2024, Eq. 1):

$$b' \;=\; b + \alpha\, u_A, \qquad \alpha \;=\; (g - b)\cdot u_A. \tag{4}$$

In words, the intervention measures how far apart the recipient's own two sentences lie along the source's direction and pushes by exactly that much. This is steering, except that the step α, the *natural step*, is read off the recipient's own pair instead of being chosen by the experimenter. Readings of transfer have left out that α differs systematically across recipients. When B's pair differs mostly along B's own direction, $\alpha \approx \nu_B \cos(u_A, u_B)$, with ν_B the volume of B. The cosine is the same both ways and largely cancels in the ratio of the two transfers, so each recipient's volume remains (Figure 2a). At layer 5, with each step divided by the source's own, the ratio of the two steps tracks the squared ratio of volumes (log–log R^2 0.79–0.94) and their product the squared cosine (Pearson 0.81–0.89; Appendix B.9).

4.2 Volume predicts the arrows

If the recipient sets the step, its volume should predict the arrows before any transfer is run, and it does: measured on each construction's own development sentences, volume predicts 77–97% of the arrows in the 14 re-measured combinations, and the step itself 87–100% (Figure 2b).[1] No single construction drives this: without the one that most often has the largest volume, the original templates still give 77–94%. Nor is volume a surface difference: pseudo-clefts keep the largest distance with capitalization and spacing matched, and pairs that differ in more tokens have smaller volumes (Appendices B.7 and B.16).

Rival accounts lose where they come apart from the step (Figure 2b; Tables 12 and 13). The own effect is volume times *steepness*, the effect per unit of step, and it fails on NPI licensing, where every structure is read on the same pair of words and steepness runs against volume (Appendix B.6): there the own effect names 51–67% of the arrows and the step 99–100% (Table 12). Structural priming grows with surprisal in people (Jaeger & Snider, 2013) and in language models (Prasad et al., 2019), but where the surprisal gap and the step disagree, 120 of 126 pairs follow the step. The original

[1]We count pair–layers whose two transfers differ by more than 0.25 log-odds, since only closer pairs flip across seeds. With no threshold, volume predicts 68–79% (Appendix B.7).

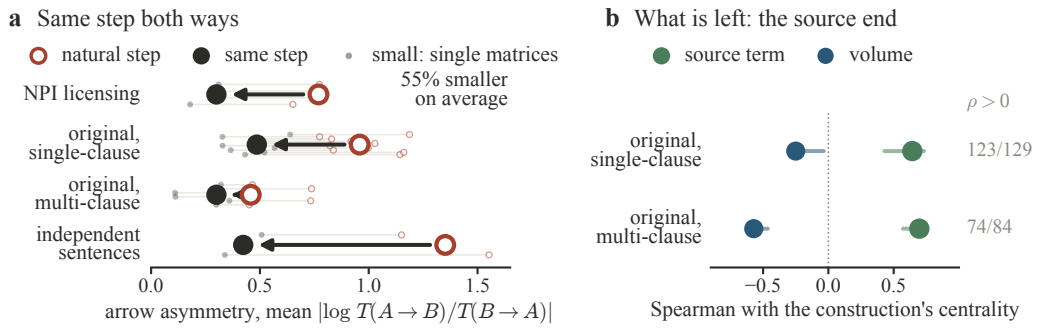


Figure 3: **Fixing one step for both transfers removes more than half of the asymmetry on average, and what is left sits at the source end.** (a) Mean $|\log T(A{\to}B)/T(B{\to}A)|$ on the layer-max matrices at the natural step (open) and at one common step per layer (filled); small: the 22 syntax matrices; large: family medians. (b) At the common step, each model's median over layers of the Spearman correlation of centrality with the fitted source term (green) and with volume (dark blue, below zero: small-volume constructions have central directions); dot: median over models; line: range; right: layer groups with a positive source-term correlation.

authors' frequency account largely coincides with volume, since rare constructions tend to have larger volumes, and where the two disagree, 40 of 52 pairs follow volume.

4.3 Fixing the step

Giving both transfers the same step removes more than half of the asymmetry. At each layer we give both transfers one common step, the median volume, as in ordinary steering. A dependence of one construction on another in particular would leave a pair-specific part, whereas a step would leave arrows that shrink. All 22 syntax matrices become less asymmetric, by 55% on average on the log scale (31–78%; Figure 3a), and so do both refusal models and 391 of 419 layer groups (Appendix B.8). Handing each transfer the source's own step instead moves the ranking to the source: its correlation with the recipient's volume, positive in 21 of 22 syntax matrices at the natural step, turns negative in 16. For the running pair re-measured at layer 5 (Figure 2a), the transfer ratio of 7.8 is 1.9×4.1: the 1.7-fold step gives the 1.9, since the response grows faster than the step, and the 4.1 left at the same step sits at the source (Appendix B.2).

The step distorts how much transfers, not whether: written with the same step as three random directions, a syntactic direction beats all three in 82–96% of (source, target, layer) cells per syntax combination, and at layer 5 the other syntactic sources, on average, also beat directions fitted on the original pipeline's two control tasks for 39 of 42 targets (Table 19). Refusal is the limiting case: its six categories share one direction and nearly one volume, so their transfers differ only 1.07- to 1.09-fold on average, leaving almost only the relation.

4.4 What is left sits at the source end

At the same step, what is left is again one number per construction (ranking share 0.72–0.99), now set at the source. We split each log transfer at the same step into a source term and a recipient term. The source term follows the *centrality* of a construction's direction, its mean $|\cos|$ with the other six, in all eight combinations (197 of 213 layer groups; Figure 3b): a central direction overlaps everyone's, so at a fixed step it pushes everyone. Small-volume constructions learn more central directions, plausibly because their own pairs barely separate and the fitted direction settles on what the constructions share. At a fixed step they therefore push harder, so the remainder mostly points the same way as the arrows at the natural step, which point to large volumes (Appendix B.8). The running pair's remainder is of this kind: non-finite embedded wh-questions learn one of the most central directions, pseudo-clefts one of the least. Volume and centrality are correlated, but at the natural step volume sets the ranking: holding centrality fixed, volume still tracks it (partial rank correlation $+0.82$, positive in 166 of 175 layer groups), while centrality adds almost nothing given volume (Table 16). In

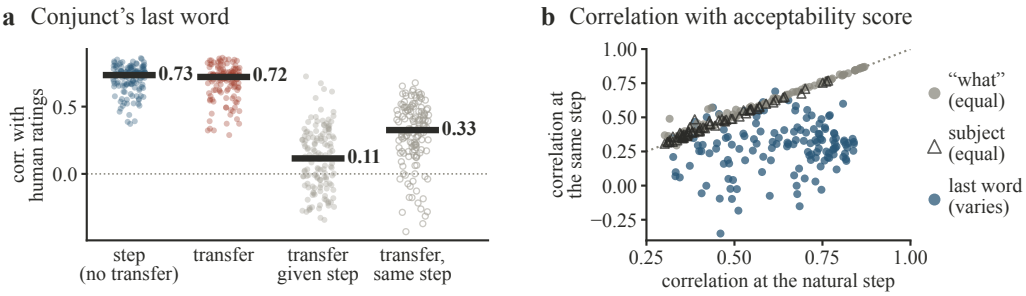


Figure 4: **Where the step varies, transfer's link to acceptability lives mostly in the step.** Coordinated verb phrases; three models, five seeds, all layers. (a) At the conjunct's last word, per responsive cell (transfer correlating with the models' acceptability score at 0.3 or more), the correlation with human ratings of the step (no transfer), the transfer, the transfer given the step (partial), and the transfer at one common step; bars: medians. (b) Per responsive cell, the correlation of transfer with the models' acceptability score at the natural step (x) and with the step matched across conjuncts (y).

Table 3: **Two-way products recover known answers; dividing by the diagonals matters where conditions differ in scale.** Relation: Eq. 3. Undivided: the two-way geometric mean $\sqrt{T(A{\to}B)T(B{\to}A)}$. Bold: the better of the two.

Known answer	Readout	Relation	Undivided
Datasets and vision tasks			
Same-task pairs (Leng & Xiong, 2025)	in top 3 of 15 (worst rank)	**10/12** (6th)	9/12 (15th)
Taskonomy families (Zamir et al., 2018)	AUC, same family first	**0.85**	0.80
A prediction of Eq. 4			
Direction overlap $\lvert\cos\rvert$	median Spearman	**+0.78**	+0.73
Constructions of one study			
Authors' judgments (Boguraev et al., 2025)	pseudo-cleft–wh-question in top 3	49/60	**60/60**
	topicalization nearer the three rarer	59/60	**60/60**
	relative clause lowest on average	48/60	**52/60**

Eq. 1, then, r_B follows the recipient's volume and s_A the centrality of A's direction, and in projection ablation, too, the amount removed follows mainly the source (Appendix B.15).

4.5 The step changes a published reading

The graded mechanism is real, but where the step varies, the step carries most of transfer's link to human judgments. A second study from the same group turns to a coordination island, extraction from one conjunct of a coordinated verb phrase, as in "I know what she hates art and loves _", which people accept to different degrees (Fergus et al., 2025; Boguraev & Mahowald, 2026). Across 33 such phrases, the transfer of a direction learned on ordinary embedded wh-questions tracks how acceptable the models find each extraction, which follows human ratings, and the study reads this as one mechanism blocked to varying degrees. A simpler reading is that the link runs through the step. Before the phrase, every phrase receives the same step, so a link there must come from the mechanism, and there transfer still tracks human ratings (0.69 at "what"). At the conjunct's last word the step varies, and there the step alone, read off each phrase's own sentences without any intervention, correlates with human ratings as highly as transfer does (0.73 and 0.72), while transfer's partial correlation given the step drops to 0.11. Equalizing the step across conjuncts lowers transfer's correlation with human ratings from 0.72 to 0.33 (Figure 4a), yet transfer keeps 45% of its correlation with the models' own acceptability scores (Figure 4b).

5 Products Relate

What two conditions share sits in K, which the products of the two transfers read.

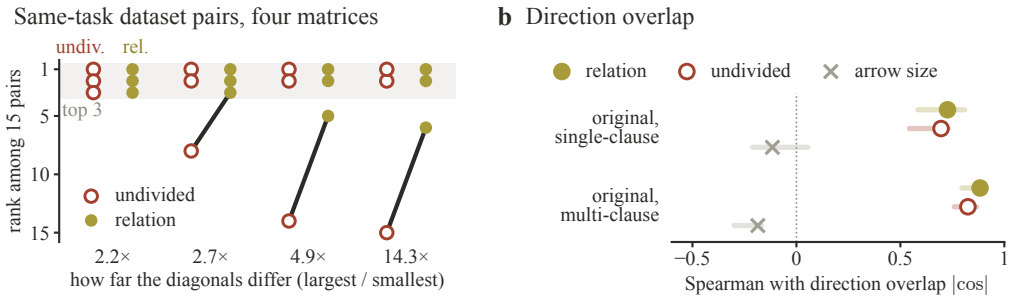


Figure 5: **Two-way products find what conditions share; dividing by the diagonals helps where conditions differ in scale.** (a) The four task-transfer matrices of Leng & Xiong (2025), ordered by how far their diagonals differ (largest over smallest): ranks, among 15 dataset pairs, of the three pairs built for the same task, undivided (open) and by the relation (filled); shaded: the top three; black: pairs that rise five ranks or more. (b) Per combination, the median over layer groups (all transfers positive) of the Spearman correlation of direction overlap $|\cos|$ with the relation, the undivided mean and the arrow's size.

5.1 Normalizing by the source's diagonal sharpens the asymmetry

The original authors normalized each transfer by the source's diagonal, to measure how well a mechanism transfers relative to how well it works at home. The diagonal carries the construction's own volume, and the division multiplies the two-way ratio by the ratio of the two diagonals:

$$\frac{T(A{\to}B)/T(A{\to}A)}{T(B{\to}A)/T(B{\to}B)} = \frac{T(A{\to}B)}{T(B{\to}A)} \cdot \frac{T(B{\to}B)}{T(A{\to}A)}. \tag{5}$$

Since arrows point to the larger diagonal, the strong transfer is divided by the small diagonal and the weak one by the large, and the asymmetry grows. It does for the released running pair, whose ratio of 7.5 becomes 13.0, in all 60 released matrices, by 1.14 to 2.09 times, and in all 14 re-measured combinations. Dividing by the recipient's diagonal instead brings that pair to 4.3 and shrinks all 60. The geometric mean of the two is again 7.5: both divisions give the same two-way product over the diagonals. Its square root is the relation of Eq. 3: the original authors' division distorts the arrows, but read over both directions it gives exactly the shared part, from only four positive transfers.

5.2 Known answers

Two-way products recover three answers known in advance and one prediction of Eq. 4, and dividing by the diagonals helps where conditions differ in scale (Table 3). The dataset pairs that Leng & Xiong (2025) built for the same task rise near the top, where the undivided product had put two of them last (Figure 5a). The gain grows with how far the diagonals differ: the worst same-task pair ranks 3rd either way at a 2.2-fold span, and rises from 15th to 6th at 14.3-fold. In Taskonomy the relation reaches AUC 0.85, against 0.80 undivided and 0.86 for Taskonomy's own normalization, and it beats the undivided product on both win-rate readouts in all three released training settings (Appendix B.13). Across the eight original-template combinations, the relation follows the overlap of the two learned directions, $|\cos(u_A, u_B)|$, which enters both steps, while the size of the arrow barely follows it (medians -0.30 to $+0.06$; Figure 5b). The constructions of one study differ little in scale (own effects 5.8–11.7 in Figure 1b), and on the original authors' three linguistic judgments both readouts recover the answers, the undivided one more often (Appendix B.11).

5.3 Rare constructions do not rely on frequent ones in particular

The ranking and the relation together test the frequency reading of the arrows. Suppose rare constructions did rely on common ones in particular. The planted dependences of §3 show that such a dependence raises the pair's residual or relation. The real matrices do have parts off the ranking, but these point the way frequency predicts no more often than coin flips, in 51% of 3,325 pair–layers and in 50.5% of the 111 residuals larger than 0.84, the residual of a planted dependence (Appendix B.4). Nor do the relations show a reliance on frequent constructions: the most frequent construction, the

relative clause, has the lowest mean relation to the others, and the rare topicalization relates less to the three most frequent constructions than to the other three, in 48 and 59 of the 60 released matrices respectively, and in 52 and 60 undivided (Table 3). The original authors' frequency account thus splits. Its ranking half holds: rare constructions have larger volumes, and more frequent constructions learn more central directions (359 of 424 layer groups; the relative clause is the exception), so at the same step they push everyone harder (§4). Its dependence half appears in neither picture. The authors' two exceptions stop being exceptions. At the same step, the relative clause gives least because its direction is the least central, and topicalization, which receives little, shares little with frequent ones.

6 Related Work

Causal abstraction treats interventions as tests of whether a model implements a shared variable (Geiger et al., 2025), and distributed alignment search learns such variables as directions (Geiger et al., 2024; Wu et al., 2023). Benchmarks compare how well intervention methods find them (Arora et al., 2024; Mueller et al., 2025), and recent work asks what a successful intervention shows (Mueller et al., 2026; Grant et al., 2026; Sutter et al., 2025). Transfer between filler–gap constructions has been measured with these tools (Boguraev et al., 2025; Kumon & Yanaka, 2026). Causal abstraction scores an alignment by interchange intervention accuracy; transfer matrices record the size of the change, and we ask what that size and its direction measure. Patching has its own illusions: subspace patching can act through a dormant pathway instead of the feature it targets (Makelov et al., 2024), and metrics and corruption methods change what patching concludes (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). In interchange intervention, how much a patch writes is set by the recipient. Steering with a chosen step (Tan et al., 2024; Braun et al., 2025) is the fixed-step counterpart of interchange intervention and distributed alignment search (Geiger et al., 2024; Wu et al., 2023), and datasets whose two sides separate better are more steerable (Braun et al., 2025), a dataset-level relative of volume.

Transfer between conditions has been read in several ways. Cross-lingual transfer has been split into donors and recipients (Malkin et al., 2022; Dymkiewicz et al., 2025), and steering transfer can be dominated by its target setting (Nguyen et al., 2026). Task2Vec writes an asymmetric distance as a symmetric one minus the source's complexity (Achille et al., 2019), Task2Box encodes asymmetry as containment (Daroya et al., 2024), and Taskonomy normalizes per target (Zamir et al., 2018). Directed transfer also guides intermediate-task choice (Pruksachatkun et al., 2020; Vu et al., 2020) and multi-task grouping by directed affinities (Standley et al., 2020; Fifty et al., 2021). Donor–recipient splits and per-target normalizations model the one-sided terms. We add the shared term, test the three-way split, identify the recipient's term as a step and read the shared term.

Eq. 1 is the quasi-symmetry model of Caussinus (1965). Its ranking half is the paired comparison of Bradley & Terry (1952), whose violations are Kendall's circular triads (Kendall & Babington Smith, 1940) and HodgeRank's cycles (Jiang et al., 2011). Its relation half, the similarity estimate of Shepard (1957) and Luce (1963), recurs as a product divided by each side's own size in gravity models of trade (Head & Ries, 2001; Novy, 2013), lift in association rules (Brin et al., 1997), coevolution (Dunn et al., 2008) and Hi-C maps (Imakaev et al., 2012). We add a mechanism for the bias, mostly the step, and an intervention that fixes it.

7 Conclusion

Any transfer matrix holds two pictures. Its arrows are almost only a ranking, which a count of cycles can check. In interchange intervention the recipient sets the ranking through the step, so comparisons across recipients should report or fix it. What two conditions share is read from the two-way product, divided by the diagonals where conditions differ in scale. Read this way, the original conclusion that the filler–gap constructions share machinery stands, since syntactic directions beat random ones at the same step, while the frequency reading of the arrows does not. Where psychology could only infer the bias, interchange intervention shows it: mostly the step, fixed before any transfer is run. For fine-tuning and steering with a chosen coefficient, what sets the ranking is the next question. The same two pictures sit in any matrix that records how much one condition moves another: its arrows rank, and its products relate.

AI USE STATEMENT

We used large language models as general-purpose assistants in this work: to search the literature, to discuss ideas, derivations, experiments and results, to help write code and process data, and to draft, translate, edit and polish the text and figures. The authors directed the research and take full responsibility for the content of this paper. All AI-assisted content was reviewed by the authors, and the reported results, references and claims were checked against the underlying data, code and sources.

ETHICS STATEMENT

The refusal experiments use existing harmful requests (Souly et al., 2024; Arditi et al., 2024). We only score the models' next-token preferences between refusing and complying, and we neither generate nor release harmful completions.

REPRODUCIBILITY STATEMENT

Appendix A gives the protocols, data splits and readouts, and Appendix A.4 how the original pipeline was reproduced. All models are open-weight, and the original study's released values, other groups' matrices and Taskonomy's transfer results are public. Code, stored result summaries and the transcribed matrices are included in the supplementary material and will be released; the figure scripts and the analyses behind Tables 2 and 3 run from the stored summaries on a CPU.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6430–6439, 2019. doi: 10.1109/ICCV.2019.00653.

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*, 2024.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663, 2024.

Sarah Ball, Frauke Kreuter, and Nina Panickssery. Understanding jailbreak success: A study of latent space dynamics in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 250–279. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.12. URL https://aclanthology.org/2026.eacl-long.12/.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023.

Sasha Boguraev and Kyle Mahowald. Causal drawbridges: Characterizing gradient blocking of syntactic islands in transformer LMs. arXiv preprint arXiv:2604.13950, 2026.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. Causal interventions reveal shared structure across English filler–gap constructions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25021–25042, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1271. URL https://aclanthology.org/2025.emnlp-main.1271/.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. arXiv preprint arXiv:2505.22637, 2025. ICLR 2025 Workshop on Foundation Models in the Wild.

Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pp. 265–276. ACM, 1997. doi: 10.1145/253260.253327.

Henri Caussinus. Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des sciences de Toulouse : Mathématiques, Série 4*, 29:77–183, 1965. doi: 10.5802/afst.519.

Rangel Daroya, Aaron Sun, and Subhransu Maji. Task2Box: Box embeddings for modeling asymmetric task relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28827–28837, 2024. doi: 10.1109/CVPR52733.2024.02723.

S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008. doi: 10.1093/bioinformatics/btm604.

Kajetan Dymkiewicz, Ivan Vulić, Helen Yannakoudakis, Eilam Shapira, Roi Reichart, and Anna Korhonen. Donors and recipients: On asymmetric transfer across tasks and languages with parameter-efficient fine-tuning. arXiv preprint arXiv:2511.13368, 2025.

Abigail Fergus, Arielle Belluck, Nicole Cuneo, and Adele Goldberg. Islands result from clash of functions: Single-conjunct wh-qs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2303.02536.

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, 2025. URL https://arxiv.org/abs/2301.04709.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118, 2024.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2511.04638.

Aaron Grattafiori et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Keith Head and John Ries. Increasing returns versus national product differentiation as an explanation for the pattern of U.S.–Canada trade. *American Economic Review*, 91(4):858–876, 2001. doi: 10.1257/aer.91.4.858.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv preprint arXiv:2404.15255, 2024.

Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, 2012. doi: 10.1038/nmeth.2148.

T. Florian Jaeger and Neal E. Snider. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83, 2013. doi: 10.1016/j.cognition.2012.10.013.

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.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

M. G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 31(3/4): 324–345, 1940. doi: 10.2307/2332613.

Ryoma Kumon and Hitomi Yanaka. Fine-grained analysis of shared syntactic mechanisms in language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44866–44891. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2078.

Yongqi Leng and Deyi Xiong. Towards understanding multi-task learning (generalization) of LLMs via detecting and exploring task-specific neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2969–2987. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.200/.

Linfeng Liu, Tiffany Zhan, Louie Hong Yao, Saptarshi Ghosh, and Tianyu Jiang. Cross-lingual steering for figurative language generation. arXiv preprint arXiv:2605.30443, 2026.

R. Duncan Luce. Detection and recognition. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter (eds.), *Handbook of Mathematical Psychology*, volume 1, pp. 103–189. John Wiley and Sons, New York, 1963.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.17030.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4903–4915. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.361. URL https://aclanthology.org/2022.naacl-main.361/.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. Mib: A mechanistic interpretability benchmark, 2025. URL https://arxiv.org/abs/2504.13151.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378, 2026. ISSN 1530-9312. doi: 10.1162/coli.a.572.

Matthew Nguyen, Kyle Cox, Austin Meek, and Iván Arcuschin. On the generalization of steering vectors for chain-of-thought faithfulness. arXiv preprint arXiv:2607.29062, 2026.

Robert M. Nosofsky. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23(1):94–140, 1991. doi: 10.1016/0010-0285(91)90004-8.

Dennis Novy. Gravity redux: Measuring international trade costs with panel data. *Economic Inquiry*, 51(1):101–121, 2013. doi: 10.1111/j.1465-7295.2011.00439.x.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 66–76. Association for Computational Linguistics, 2019. doi: 10.18653/v1/K19-1007. URL https://aclanthology.org/K19-1007/.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5231–5247, 2020. doi: 10.18653/v1/2020.acl-main.467.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Shun Shao, Binxu Wang, Shay B. Cohen, Anna Korhonen, and Yonatan Belinkov. Differentiable faithfulness alignment for cross-model circuit transfer. arXiv preprint arXiv:2604.24302, 2026.

Angira Sharma, Christian Schroeder de Witt, Philip Torr, Anisoara Calinescu, and Jialin Yu. A low-rank subspace analysis of LLM interventions. arXiv preprint arXiv:2606.14388, 2026.

Roger N. Shepard. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4):325–345, 1957. doi: 10.1007/BF02288967.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132, 2020.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability?, 2025. URL https://arxiv.org/abs/2507.08802.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.12404.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. doi: 10.1037/0033-295X.84.4.327.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7882–7926, 2020. doi: 10.18653/v1/2020.emnlp-main.635.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.08809.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah D. Goodman, Christopher D. Manning, and Christopher Potts. pyvene: A library for understanding and improving PyTorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, 2024. URL https://arxiv.org/abs/2403.07809.

Jieying Xue, Phuong Minh Nguyen, Minh Le Nguyen, and Shogo Okada. Cross-lingual functional vectors for emotion detection in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026.

An Yang et al. Qwen2 technical report. arXiv preprint arXiv:2407.10671, 2024.

An Yang et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3712–3722, 2018. doi: 10.1109/CVPR.2018.00391.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16042.

A Data and protocols

A.1 The intervention, the readout and three numbers

Directions. For every construction we fit one rank-one direction with distributed alignment search (Geiger et al., 2024), using the original pipeline's trainer (Boguraev et al., 2025; Arora et al., 2024) through pyvene (Wu et al., 2024): 100 optimisation steps of two minimal pairs each, learning rate 0.005, seed 42, on the construction's own training pairs (up to 200, disjoint from the evaluation pairs). Directions are fitted separately at every layer and for animate and inanimate constructions. The intervention site is the residual stream at the article of the gapped clause, the site of the original study; Appendix B.7 moves it to the noun, the readout word and the attention output.

Transfer. To transfer A's direction u_A into construction B, we take one of B's evaluation pairs, replace the coordinate along u_A of one sentence's representation b by that of the other sentence's representation g (Eq. 4), and read the change in the log-odds of the other sentence's continuation against its own, relative to no intervention. Each evaluation pair (200 per construction in the original templates) is used in both orientations, gapless into gapped and gapped into gapless, and $T(A{\to}B)$ is the mean effect; a positive value means the model moves toward the other sentence's continuation.

Three numbers per recipient. The *step* written from A into B is the median of $|(g-b)\cdot u_A|$ over B's evaluation pairs: the amount this intervention actually adds. The *volume* ν_B is the median of $|(g-b)\cdot u_B|$ over B's own development pairs: how far apart B's two sentences lie along its own direction, measured on other sentences and without any transfer. It is also each construction's *own step*, used below when a step is fixed. The *own effect* is the diagonal $T(B{\to}B)$, the only one of the three that the released matrices contain. A number *names* an arrow when the arrow points to the recipient with the larger value. We count pair–layers whose two transfers differ by more than 0.25 log-odds; Table 14 gives other thresholds.

Fixing the step. Because the intervention is steering with the step $\alpha = (g - b) \cdot u_A$, any step can be written instead, keeping the sign of the natural projection. *Same step both ways*: at each layer, one common step, the median of the constructions' own steps, is written for every source and recipient. *Source's step*: $A \to B$ is written with A's own step. *Random directions*: three random unit directions per source, written with the source's own step; a syntactic direction *beats random* in a (source, target, layer) cell when its transfer exceeds that of all three. At layer 5 we also compare them at one common absolute step (Table 19). The same-step comparison uses the layer-max matrix, the statistic of the released figures.

Rival numbers. *Surprisal*: how much more surprised the model is, summed over the tokens from the start of the sentence to the intervention word, by the gapped sentence than by the gapless one (for NPIs, by the sentence with the licensor), averaged over evaluation pairs; structural priming in language models grows with the target sentence's surprisal (Prasad et al., 2019). *Frequency*: the corpus counts the original authors report (their Appendix D, Table 8; English Web Treebank): relative clause 504, embedded wh 308 (both kinds), wh-question 82, cleft 20, pseudo-cleft 6, topicalization 6.

A.2 Data

Original templates. The seven filler–gap constructions of Boguraev et al. (2025) (cleft, pseudo-cleft, topicalization, wh-question, restricted relative clause, finite and non-finite embedded wh), each with animate and inanimate fillers, in the single-clause and multi-clause datasets. We re-measure the single-clause templates in Pythia-1.4B, 2.8B and 6.9B (Biderman et al., 2023), Llama-3.2-1B (Grattafiori et al., 2024) and Qwen3-1.7B-Base (Yang et al., 2025), and the multi-clause templates in the three Pythia models, at every layer (24, 32, 32, 16 and 28 layers).

Independent sentences. Kumon & Yanaka (2026) generated their own sentences for seven filler–gap constructions and for eight structures that license negative polarity items. We use their released splits: directions are fitted on their training split (200 pairs per construction), own steps and volumes are measured on their in-distribution test split after removing pairs that share a sentence with training, in complete batches of eight (16–96 pairs for NPIs, 96 for filler–gap), and transfers are evaluated on their lexically disjoint test split (100 pairs per construction). The filler–gap site is the article of the gapped clause, as above, in Pythia-1B and 1.4B. The NPI site is the last word before the NPI, the only position all structures share; directions are fitted on "ever" against other adverbs ("already", "also", "never", "often", "still") as in their training split, and the transfer readout is "any" against "some" for every structure; models are Pythia-1B, 1.4B, 2.8B and 6.9B. Across the four models the NPI structures with the largest volume are yes/no questions ("Had / Is …") and determiner negation ("No …"), and the smallest is the quantifier ("all of …").

Refusal. Six harm categories of StrongREJECT (Souly et al., 2024), in the processed copy of Arditi et al. (2024) (disinformation, hate and harassment, illegal goods, non-violent crimes, sexual content, violence), and Alpaca instructions (Taori et al., 2023) as the harmless side, in Llama-3-8B-Instruct and Gemma-2-9B-it (Gemma Team et al., 2024). Every prompt is wrapped in the model's chat template with the generation prompt. A pair is a harmless prompt and a harmful prompt of one category; the intervention sits on the final token of the generation prompt, and the readout is the refusal token (the most common first token on harmful prompts) against the harmless prompt's own top token. We keep harmful prompts whose top token is the refusal token and harmless prompts whose top token is neither the refusal token nor a special token, split the harmful prompts of each category 40/30/30 into fitting, development and evaluation, and pair them with disjoint pools of harmless prompts: 200 fitting, 100 development (96 used) and 100 evaluation pairs per category. With 50–59 prompts per category, each harmful prompt appears in several pairs with different harmless prompts. Directions are fitted with the same trainer, with the donor-index fix of Appendix A.4. Nine layers per model (Llama: 4, 8, 10, 12, 14, 16, 20, 24, 28; Gemma: 4 to 36 in steps of 4).

Released matrices. The layer-max matrices released by Boguraev et al. (2025) for Pythia-1.4B, 2.8B and 6.9B: in the single-clause dataset three intervention positions (filler, article, noun) and in the multi-clause dataset seven, each for animate and inanimate constructions, 60 matrices of 21 pairs. The position of the readout word is left out: its released values are affected by two indexing problems in the original pipeline (Appendix A.4).

Other groups' matrices. Every square source-by-target matrix with printed numbers in seven papers, 72 in total, transcribed and checked row by row against the papers. Rows are the condition the intervention comes from, columns the condition tested. The effect is the printed change; where a paper prints the score after intervention, we subtract the reference printed in the same table. Matrices with only two conditions are not counted, leaving 68. The seven papers are those our literature search found that print a square transfer matrix, diagonal included, as numbers. Arrows and cycles use the sign of the two-way difference. Because many of these effects are zero or negative, Table 2 computes the ranking share on the pairs with both directions positive (61 matrices have enough of them), and Table 23 also computes it on two-way differences over all 68.

Taskonomy. The official first-order transfer results of Zamir et al. (2018) (the `results/` folder of the public repository): a source task's encoder is frozen and a small readout is trained for the target task. We use the square matrix of the 22 tasks that are both sources and targets; the diagonal is the target's own encoder with the same small readout. The main readout is the win rate against training from scratch; we also use the win rate against the task-specific network and the paper's AHP affinity, each in three released settings: transfer trained on 16k or 1k images, and the files without a size suffix, whose setting the repository does not document and whose task set has object instead of scene classification. We add 0.01 to every win rate before taking logs, since some are zero (0.005 or 0.02 give relation AUCs of 0.80–0.85, all above the undivided mean). Families are the four colored in the paper's Figure 14, the task tree built by clustering transfers, with denoising, drawn apart there, placed with 2D by its output. *2D*: autoencoding, denoising, 2D edges, 2D keypoints, 2D segmentation. *3D*: occlusion edges, 3D keypoints, surface normals, reshading, curvature, depth, distance, 2.5D segmentation. *Low-dimensional geometry*: room layout, vanishing points, the two camera-pose tasks, egomotion, point matching. *Semantic*: object classification, scene classification, semantic segmentation.

Coordinated verb phrases. Boguraev & Mahowald (2026) move the extraction direction into 33 coordinated verb phrases. We run their code with their trained directions in GPT-2 (Radford et al., 2019), Pythia-410M and Qwen2-0.5B (Yang et al., 2024), five seeds, three sites (MLP activation, attention value output, block output), three positions ("what", the subject, the conjunct's last word) and all layers, and add two quantities: the step written into each conjunct, the projection of its pair on the extraction direction, and the transfer with that step set to the source's median step for every conjunct.

A.3 Readouts on a matrix

Cycles. For every triple of conditions whose three pairs all have an arrow (no tie), whether the three arrows form a directed cycle. Random arrows cycle in one triple out of four; a ranking never does.

Ranking share. Let $Y(A,B) = \log T(A{\to}B) - \log T(B{\to}A)$ over the pairs with both directions positive. Least squares finds one number ϕ per condition whose differences $\phi_B - \phi_A$ best reproduce Y; the rest of Y is cyclic (Jiang et al., 2011). The ranking share is $\|\nabla\phi\|^2/\|Y\|^2$. The reference flips the sign of each pair's Y at random (2,000 draws); for n conditions pure noise gives about $(n-1)/\binom{n}{2}$, 0.29 for seven.

Source and recipient terms. Under a fixed step, $\log T(A{\to}B) = \mathrm{src}(A) + \mathrm{rec}(B)$ is fitted by least squares on the off-diagonal entries only. How central a construction's direction is: the mean $|\cos|$ of its direction with the other six.

Relation. Eq. 3 on the layer-max matrix, or on a single layer where stated; logs require all 49 transfers of a group to be positive (175 of 440 layer–animacy groups in the eight original-template combinations).

A.4 Reproducing the original pipeline

We reran the original code, retraining and re-evaluating, and compared it with the released per-layer values (Pythia-1.4B, six source constructions, every position and layer): Pearson 0.96 over 6,720 cells, mean ratio 1.003, and 0.99–1.00 per cell after taking the layer maximum. Our re-measurement

agrees with the released single-clause figures at the article position with Spearman 0.71, 0.75 and 0.72 per layer in the three Pythia models.

Two indexing problems affect four of the seven constructions in both datasets. The donor token is read at the base sentence's index, so in clefts and topicalization every pair reads the wrong word; in these constructions and in pseudo-clefts and wh-questions the base and source sentences have different numbers of words before the article, while in the other three they do not. And the per-span token counts disagree with the whole-sentence tokenization on one side of every pseudo-cleft and wh-question pair. Fixing the two lines in the original code (multi-clause dataset, second article, Pythia-1.4B) brings each construction's curve to ours: the mean per-layer difference drops from 0.79 to 0.10 for clefts and from 1.72 to 0.08 for topicalization (fixing only the line that reads the donor token), and from 3.18 to 0.05 for pseudo-clefts and from 1.65 to 0.10 for wh-questions (fixing both lines; the wh-question "before" uses released values), against 0.07 and 0.06 for the unaffected finite embedded wh. Fixing only the donor line leaves pseudo-clefts at 3.24. With the same direction, sentences and step rule, changing only which word is written changes each item's effect with the same sign as the change in the step written, for 79% of 3,200 items (Spearman 0.51) and all eight cell means. At layer 6, for four fixed directions and the affected targets, the fix changes cell means by factors of 0.79 to 2.92 and flips the sign of up to 32% of items.

At the readout word, the released values split cleanly by these problems: pairs with both ends unaffected point to the larger own effect in 32 of 36 cases, pairs involving an affected construction in 73 of 216. Over all positions pooled, pairs with both ends unaffected point to the larger own effect in 171 of 216 cases (79%), as often as all pairs do. All re-measured results use the fixed pipeline. Before using the sentences of Kumon & Yanaka (2026), we checked their pipeline's operators with donor equal to recipient, which should change nothing: their training operator still changes the preferred continuation on 30–34% of items at attention and MLP components, while the fixed training operator and their evaluation operator change none. We use our own operator throughout.

B Further results

B.1 Released matrices: arrows point to the larger own effect

Table 4 breaks the 1,001 of 1,260 released arrows that point to the larger own effect down by model and position, and Table 5 by how much the two own effects differ: nearly equal diagonals give a coin flip, and diagonals at least twice apart give 96% (340 of 356). The log two-way ratio grows with the log ratio of the own effects with a median slope of 1.23 (0.65–2.02) across the 30 conditions. One number per construction explains a median 0.91 of each matrix's squared log ratios (59 of 60 above sign flips), and that number correlates positively with the own effect in 60 of 60 (median $\rho = +0.75$). Shuffling the own effects within each matrix, 2,000 times, gives 50% on average (at most 56%), against the observed 79%. Re-measured with the corrected pipeline (our Pythia-1.4B single-clause map, article position, animate constructions, layer maximum), the matrix of Figure 1b has no cyclic triple (0 of 35), and 16 of its 21 arrows point to the larger own effect. The figure the original authors drew averages four animacy conditions; there 16 of 21 arrows point to the larger own effect, and the running pair is 2.41 against 0.41. Pairs across animacy (an animate and an inanimate construction) follow too: at the noun position 30, 39 and 33 of 42 in the single-clause matrices and 36, 41 and 31 of 42 at the second noun of the multi-clause matrices.

Splitting each released single-clause matrix into a source term and a recipient term, the recipient term takes a share of 0.70–0.81 at the noun position and alone reproduces 33–37 of 42 arrows; at the article, 0.77–0.83 and 29–36 arrows; at the filler, 0.51–0.69 and 24–30 arrows.

Table 4: Released matrices of Boguraev et al. (2025): arrows pointing to the construction with the larger own effect (42 pairs per position: 21 pairs × animate and inanimate), and the median share of the arrows explained by one number per construction together with the median Spearman correlation of that number with the own effect.

Data	Model	Noun	Article	All but the readout word	Ranking share · ρ with own effect
Single clause	Pythia-1.4B	37/42	32/42	98/126 (3 positions)	0.90 · +0.75
Single clause	Pythia-2.8B	38/42	32/42	99/126 (3 positions)	0.80 · +0.79
Single clause	Pythia-6.9B	33/42	29/42	93/126 (3 positions)	0.86 · +0.48
Multi-clause	Pythia-1.4B	39/42	35/42	232/294 (7 positions)	0.93 · +0.77
Multi-clause	Pythia-2.8B	41/42	38/42	247/294 (7 positions)	0.94 · +0.84
Multi-clause	Pythia-6.9B	32/42	33/42	232/294 (7 positions)	0.90 · +0.70

Table 5: All 1,260 released pairs, binned by how much the two constructions' own effects differ: share of arrows pointing to the larger own effect.

Ratio of the two own effects	Animate	Inanimate	Total
1–1.1×	43/76 = 57%	31/54 = 57%	74/130 = 57%
1.1–1.25×	73/106 = 69%	64/93 = 69%	137/199 = 69%
1.25–1.5×	85/125 = 68%	109/139 = 78%	194/264 = 73%
1.5–2×	128/157 = 82%	128/154 = 83%	256/311 = 82%
2–3×	83/89 = 93%	111/118 = 94%	194/207 = 94%
≥3×	74/77 = 96%	72/72 = 100%	146/149 = 98%

B.2 The running pair in numbers

Table 6 collects every number the main text gives for the running pair, the direction learned on non-finite embedded wh-questions and the one learned on pseudo-clefts.

Table 6: The running pair. Released: the original authors' values (Pythia-1.4B, single-clause templates, article position, animate constructions). Re-measured: our corrected pipeline.

Quantity	Value
Released transfers	emb. wh (non-finite) → pseudo-cleft 5.72, reverse 0.76 (ratio 7.5)
Released diagonals, Figure 1b	pseudo-cleft 11.7, wh-question 9.7, relative clause 9.6, cleft 7.5, topicalization 7.2, emb. wh (non-finite) 6.7, emb. wh (finite) 5.8
Orderings of the seven constructions with at least 19 of 21 arrows up	27 of 5,040
Released relations, Figure 1c	running pair 0.24 (fourth of 21); pseudo-cleft and wh-question 0.35 (second); topicalization with the relative clause and both embedded wh-questions 0.06–0.07; mean relation of the relative clause with the other six 0.06 (lowest of seven)
Re-measured, Pythia-1.4B, layer 5	steps written 18.8 into pseudo-cleft and 10.8 back; volumes 25.6 (pseudo-cleft) and 19.0; transfers 4.19 and 0.54 (ratio 7.8); at one common step 3.68 and 0.89 (ratio 4.1)
Re-measured layer-max ratio, natural → common step	Pythia-1.4B 9.4 → 3.7; Pythia-2.8B 4.6 → 2.2; Pythia-6.9B 5.2 → 2.3; Llama-3.2-1B 7.1 → 2.9; Qwen3-1.7B 6.6 → 2.2

B.3 Arrows rank at every size, and beyond seed noise

Small matrices give the cycle count little room: a matrix of three conditions has one triple. Table 8 splits the 152 matrices by size. Cycles stay rare at every size, 1.8% and 3.2% of triples in matrices of 8 to 14 conditions and 3.4% in Taskonomy's 22, and 3.4% overall (95% interval 2.5–4.3%). The matrices are not independent: 60 are one study's release and 24 re-measure its paradigm. Resampling the ten studies first, then matrices within them, gives 2.2–4.6%. The study with the most cycles, cross-model circuit transfer (Shao et al., 2026), cycles in 10.5% of triples, below the 13% of a pair-specific part as large as the ranking. The ranking share is lowest in the largest matrices, the jailbreak matrices of Ball et al. (2026) (0.42 to 0.59, and 0.08 in one), because most of their pairs are tied: only 124 of their 1,156 triples have no tied pair, and one matrix has none. A tied pair is one the ranking must fit at zero, so ties lower the ranking share without adding cycles.

Is the small remainder that does not fit the ranking just noise? Table 9 uses the directions refit with three training seeds (Table 20). Removing seed noise, the ranking accounts for 0.90 to 1.00 of the directed structure that reproduces across seeds (median 0.95). The remainder is small, and in eight

of the ten matrices it is not seed noise: it correlates across seeds (median 0.97). A ranking captures almost all of what the arrows say, and a small pair-specific part is left.

How large a pair-specific part could hide under 3% of cyclic triples? Table 10 simulates matrices whose log two-way ratios are a ranking plus an antisymmetric pair-specific part of a given share of the directed variance. The cycle rate is about a quarter of that share at 4, 7 and 22 conditions: the observed 3% corresponds to a pair-specific part of about a tenth, and a part as large as the ranking itself would produce 13% of cyclic triples. The simulated matrices have no ties, so this correspondence between share and cycles does not apply to heavily tied matrices.

Table 7 fits Eq. 1 directly to the off-diagonal log transfers, next to a symmetric part alone and to one source and one recipient number per condition alone. Eq. 1 explains a median 0.97 of the variance in the released and re-measured matrices, 0.93 in other groups' matrices with all entries positive and 0.98 in Taskonomy; neither half alone comes close except in Taskonomy, whose relation is weak. Because the shared part has one free number per pair, the same fit explains about 0.65 of the variance of random numbers of the same size. The free numbers cannot, however, predict what they have not seen: leaving out each cell in turn and predicting it from the fit to all the others, Eq. 1 still explains 0.78 of the variance in the released matrices, 0.79 in the re-measured ones, 0.42 in other groups' and 0.90 in Taskonomy, while on random numbers the same prediction is worse than the mean (-1.2 to -2.3), and so is predicting each transfer by the reverse one, the symmetric part alone (-0.18 to -1.15).

Table 7: Eq. 1 fitted directly to the off-diagonal log transfers: median share of their variance explained by a symmetric part alone, by one source and one recipient number per condition alone, and by both (quasi-symmetry), over matrices with all off-diagonal entries positive; on noise: the same fit on matrices of independent log-normal entries of the same size (200 draws), which the free parameters alone explain this far. Held out: each cell left out in turn and predicted from the fit to all the others (predictive R^2, which falls below zero when the prediction is worse than the mean; for the symmetric part alone the prediction is the reverse transfer).

Source	Matrices	Symmetric part only	Source and recipient terms only	Eq. 1 (both)	Eq. 1 on noise	Held out: symmetric part only	Held out: Eq. 1	Held out: Eq. 1 on noise
Released	60	0.70	0.59	0.97	0.65	−0.18	0.78	−1.77
Re-measured syntax (layer-max)	16	0.67	0.71	0.97	0.65	−0.33	0.79	−1.77
Other groups (all positive)	32	0.50	0.71	0.93	0.63	−1.00	0.42	−2.29
Taskonomy (16k, win rate +0.01)	1	0.46	0.95	0.98	0.54	−1.15	0.90	−1.21

Table 8: Cycles at every matrix size, over the 152 matrices of Table 2 and, in the last row, Taskonomy; intervals from resampling matrices (10,000 draws).

Matrix size	Matrices	Cyclic triples (95% interval)	Ranking share, median (sign flips)
3–4 conditions	10	0/29 = 0.0% (0.0%–0.0%)	0.94 (0.67)
5–7 conditions	134	127/3589 = 3.5% (2.6%–4.5%)	0.91 (0.29)
8–12 conditions	5	4/224 = 1.8% (0.0%–4.2%)	0.77 (0.25)
13 or more conditions	3	4/124 = 3.2% (0.0%–5.5%)	0.50 (0.15)
Taskonomy (22 conditions)	1	52/1540 = 3.4%	–

Table 9: Noise ceiling from training seeds: directions refit with seeds 42, 43 and 44 on the same evaluation pairs, for the matrices whose 49 transfers are all positive. Seed noise is the variance of each pair's log ratio across seeds; if it were independent across pairs it would add on average 15 times its size to the cyclic energy of a seven-condition matrix and 6 times its size to the ranking part.

Model (single clause)	Layer	Animacy	Ranking share (one seed)	Ranking share, seed noise removed	Seed correlation: ranking part	Seed correlation: cyclic part	Cyclic energy / seed noise
Pythia-1.4B	3	animate	0.94	0.95	0.99	0.98	8.3
Pythia-1.4B	3	inanimate	0.88	0.90	0.97	0.95	4.8
Pythia-1.4B	9	animate	0.94	0.94	1.00	0.97	29.7
Pythia-1.4B	9	inanimate	0.92	0.98	0.99	0.82	1.4
Pythia-2.8B	4	animate	0.96	0.96	1.00	0.98	8.5
Pythia-2.8B	4	inanimate	0.96	0.96	1.00	0.98	11.9
Pythia-2.8B	12	animate	0.96	0.99	0.98	0.61	1.3
Pythia-2.8B	12	inanimate	0.93	1.00	0.82	0.47	0.4
Pythia-6.9B	4	animate	0.93	0.93	1.00	0.98	22.9
Pythia-6.9B	4	inanimate	0.94	0.94	1.00	0.98	36.0

Table 10: Simulated power of the two readouts. Each log two-way ratio is a difference of one number per condition plus an antisymmetric pair-specific part; the share is the part's fraction of the directed variance; 2,000 matrices per cell (300 for 22 conditions).

Pair-specific share of directed variance	4 conditions: cycles · ranking share	7 conditions	22 conditions
0.00	0.0% · 1.00	0.0% · 1.00	0.0% · 1.00
0.05	1.7% · 0.97	1.3% · 0.96	1.4% · 0.95
0.10	2.4% · 0.95	2.7% · 0.92	2.7% · 0.91
0.20	5.3% · 0.90	5.4% · 0.85	5.5% · 0.81
0.30	7.9% · 0.85	7.9% · 0.77	8.0% · 0.72
0.50	13.0% · 0.75	12.9% · 0.63	12.7% · 0.55
0.70	16.8% · 0.66	18.0% · 0.48	17.9% · 0.35
1.00	25.2% · 0.49	25.0% · 0.27	25.0% · 0.09

B.4 A PLANTED DEPENDENCE

Would the readouts see a dependence of B on A in particular if there were one? In Pythia-1.4B (single-clause templates, animate constructions, layer 5) we replaced the relative clause's learned direction u_{rc} by the unit vector along $u_{\mathrm{rc}} + c\,u_{\mathrm{top}}$, so that the relative clause's intervention also carries topicalization's mechanism: topicalization, one of the two rarest constructions, then relies on the relative clause, the most frequent, in particular, which is the frequency reading's case. As c goes from 0 to 0.5 and 1, the planted pair's residual from the ranking grows from 0.46 (third of 21 pairs) to 0.84 and 1.07 (first), and it stays first when both transfers are written with the same step (0.29, 0.42, 0.57). Its relation rises too, from 0.04 (18th of 21) to 0.09 (11th) and 0.16 (fifth). The ranking share moves only from 0.98 to 0.97 and 0.96, and no triple cycles: a single dependence shows in its pair's residual and relation, not in the cycle count or the ranking share.

The same plant in other models and pairs (Table 11): in Pythia-1.4B, 2.8B and 6.9B (single-clause templates, animate, layer 5), a share $c = 0.5$ of a rarer construction's direction is mixed into a more frequent one's, for three pairs: topicalization into the relative clause, pseudo-cleft into the finite embedded wh-question, cleft into the wh-question. In two of the nine plants the planted transfer turns negative, so the pair drops out of the log-ratio readouts. In the other seven the pair's residual from the ranking grows in six and its relation rises in six (both in five); the pair becomes the largest part off the ranking in four, all three in Pythia-1.4B and one in 6.9B. Cycles stay at 0 to 3 of 35 triples and the ranking share at 0.90 to 0.97.

In the real matrices (eight original-template combinations, 175 layer–animacy groups), the largest residual of a group has median 0.64 and exceeds 0.84 in 32% of groups. These residuals do not follow frequency: oriented so that a positive residual means the more frequent construction gives the rarer one more than the ranking predicts, 51.0% of the 3,325 residuals (19 pairs of unequal frequency in 175 groups) are positive, and 50.5% of the 111 larger than 0.84. Nor do they repeat across models: the sign of each pair's mean residual agrees across the eight combinations as often as eight coin flips

would (median 0.625, chance 0.64). Within one model the remainder is stable across training seeds (Appendix B.3), so it belongs to each trained model, not to the pair of constructions.

Table 11: Planted dependences in three models (natural step, layer 5): the planted pair's |residual| from the ranking and its rank among 21 pairs, and the rank of its relation, in the model's own map (base) and after the plant. –: a transfer of the pair is not positive after the plant.

Model	Planted (rarer into more frequent)	|Residual|: base	planted	Relation rank: base	planted	Cycles
Pythia-1.4B	topicalization into relative clause	0.46 (#3)	0.84 (#1)	18	11	0
Pythia-1.4B	pseudo-cleft into emb. wh (finite)	0.02 (#20)	1.04 (#1)	2	9	0
Pythia-1.4B	cleft into wh-question	0.24 (#5)	0.71 (#1)	15	7	0
Pythia-2.8B	topicalization into relative clause	0.23 (#8)	–	18	–	0
Pythia-2.8B	pseudo-cleft into emb. wh (finite)	0.06 (#19)	0.32 (#5)	3	2	0
Pythia-2.8B	cleft into wh-question	0.65 (#2)	0.23 (#8)	17	8	0
Pythia-6.9B	topicalization into relative clause	0.26 (#13)	1.15 (#1)	19	12	2
Pythia-6.9B	pseudo-cleft into emb. wh (finite)	0.24 (#14)	0.33 (#7)	7	4	0
Pythia-6.9B	cleft into wh-question	0.43 (#1)	–	15	–	3

B.5 Which number names the arrow

Table 12 gives, for every re-measured combination, how often each per-recipient number names the arrow. The step written is the quantity the intervention actually adds, measured on the same sentences as the arrow; volume uses other sentences. Counting each pair of constructions once (majority over its layers where the two transfers differ by more than 0.25; ties dropped), the step names 231, volume 215 and the own effect 215 of 251 pairs in the ten filler–gap combinations. In the two refusal models the six categories share one direction and are of almost equal volume, so both one-number-per-category stand-ins fall behind the step: 55% and 64% for volume, 63% and 77% for the own effect, against 86% and 87%.

Surprisal names 46–84% of the arrows. Per pair, the step names 303 of 324 and surprisal 190; where they disagree, 120 of 126 pairs follow the step. Rare constructions do surprise the model more (18 of 18 groups, median $\rho = -0.65$ with frequency), but surprisal is only weakly related to volume (median $+0.20$).

Frequency is also one number per construction, and it runs with volume: rarer constructions tend to have larger volume (18 of 18 groups, median $\rho = -0.41$ between frequency and layer-max volume). Table 13 compares them where they disagree.

Table 12: Which number names the arrow. Pair–layers whose two transfers differ by more than 0.25 log-odds; share of arrows pointing to the recipient with the larger step written, volume, own effect or surprisal gap, or, for frequency, to the rarer construction (pair–layers whose constructions differ in frequency).

Data	Model	Pair–layers	Step	Volume	Own effect	Surprisal	Frequency
Reference: the 60 published matrices (diagonal only)	3 models	1260 pairs	–	–	1001/1260 = 79%	–	842/1140 = 74%
NPI licensing	Pythia-1B	85	85/85 = 100%	75/85 = 88%	44/85 = 52%	71/85 = 84%	–
NPI licensing	Pythia-1.4B	140	139/140 = 99%	131/140 = 94%	72/140 = 51%	94/140 = 67%	–
NPI licensing	Pythia-2.8B	162	162/162 = 100%	149/162 = 92%	108/162 = 67%	115/162 = 71%	–
NPI licensing	Pythia-6.9B	164	163/164 = 99%	146/164 = 89%	95/164 = 58%	84/164 = 51%	–
Original, single clause	Pythia-1.4B	336	323/336 = 96%	313/336 = 93%	305/336 = 91%	216/336 = 64%	238/279 = 85%
Original, single clause	Pythia-2.8B	349	334/349 = 96%	340/349 = 97%	332/349 = 95%	214/349 = 61%	266/306 = 87%
Original, single clause	Pythia-6.9B	362	329/362 = 91%	333/362 = 92%	315/362 = 87%	219/362 = 60%	250/304 = 82%
Original, single clause	Llama-3.2-1B	133	132/133 = 99%	126/133 = 95%	127/133 = 95%	67/133 = 50%	97/117 = 83%
Original, single clause	Qwen3-1.7B	174	161/174 = 93%	159/174 = 91%	153/174 = 88%	134/174 = 77%	129/140 = 92%
Original, multi-clause	Pythia-1.4B	103	100/103 = 97%	96/103 = 93%	92/103 = 89%	79/103 = 77%	70/100 = 70%
Original, multi-clause	Pythia-2.8B	87	76/87 = 87%	79/87 = 91%	77/87 = 89%	40/87 = 46%	52/76 = 68%
Original, multi-clause	Pythia-6.9B	224	206/224 = 92%	172/224 = 77%	207/224 = 92%	150/224 = 67%	138/214 = 64%
Independent filler–gap	Pythia-1B	55	52/55 = 95%	51/55 = 93%	50/55 = 91%	37/55 = 67%	35/55 = 64%
Independent filler–gap	Pythia-1.4B	107	103/107 = 96%	99/107 = 93%	95/107 = 89%	58/107 = 54%	87/100 = 87%
Refusal (six harm categories)	Llama-3-8B-Instruct	110	95/110 = 86%	60/110 = 55%	69/110 = 63%	–	–
Refusal (six harm categories)	Gemma-2-9B-it	78	68/78 = 87%	50/78 = 64%	60/78 = 77%	–	–

Table 13: The authors' frequency account (arrows point to the rarer construction), with frequencies from their Appendix D, Table 8 (English Web Treebank counts: relative clause 504, embedded wh 308, wh-question 82, cleft 20, pseudo-cleft 6, topicalization 6).

Pairs differing in frequency	Pairs	Our number names the arrow	Frequency names the arrow	Where they disagree, arrows follow
All released pairs (30 conditions × 2 sentence sets)	1140	901/1140 = 79% (own effect)	842/1140 = 74%	433: own effect 246 · frequency 187
Of these, the network of Figure 1	19	17/19 = 89% (own effect)	14/19 = 74%	7: own effect 5 · frequency 2
Our ten re-measured filler–gap combinations (pair–layers, two-way difference > 0.25)	1691	1596/1691 = 94% (step)	1362/1691 = 81%	318: step 276 · frequency 42
Same, one vote per pair (majority over layers)	228	209/228 = 92% (step)	154/228 = 68%	73: step 64 · frequency 9
Same, one number per construction on both sides: layer-max volume vs. frequency (layer-max matrix, two-way difference > 0.25)	169	152/169 = 90% (volume)	124/169 = 73%	52: volume 40 · frequency 12

B.6 Why the diagonal loses on NPIs: steepness

The own effect is the effect of the step a construction writes into itself, so it is volume times *steepness*: how much effect each unit of step buys on the construction's own sentences. In the ten filler–gap combinations steepness rises with volume (median per-layer Spearman $+0.11$ to $+0.77$, all positive), so the diagonal ranks like volume, and arrows point to the steeper end in 74–92% of pair–layers. In the four NPI combinations steepness runs against volume (-0.32 to -0.06), and arrows point to the steeper end in only 10–21%. There the diagonal comes apart from the step: where the step and the own effect disagree, all 230 pair–layers follow the step; where volume and the own effect disagree, 190 of 198 follow volume. Per pair, the step names 72 of 73 NPI pairs, volume 65 and the own effect 38.

B.7 Robustness across layers, thresholds and sites

The threshold of 0.25 log-odds separates near-ties from arrows. When the directions of Pythia-1.4B, 2.8B and 6.9B at layer 5 are refit with three further training seeds, 8 of the 126 pairs flip their arrow across the four seeds, all of them with a difference below 0.25; none of the 72 pairs at or above 0.25 flips (median seed-to-seed spread of the difference 0.015 to 0.040). Table 14 repeats the volume law layer by layer at three thresholds on the two-way difference. Without a threshold, volume names 68–79% of the arrows, and the threshold makes it more accurate in all 14 combinations; at 0.5 it names 80–100%. The law does not hang on one construction with a large volume. Leaving out, in each original-template combination, the construction that most often has the largest volume (pseudo-clefts in the single-clause templates, topicalization in the multi-clause ones), volume still names 77–94% of the arrows, against 77–97% with all seven. The authors' division by the source's diagonal enlarges the asymmetry in all 14 combinations, as in the 60 released matrices. Table 15 moves the intervention to the noun, the readout word and the attention output.

Table 14: Every combination, layer by layer: how often the recipient's volume names the arrow at three thresholds on the two-way difference, and the authors' division by the source's diagonal.

Data	Model	Layers	Volume names arrow (>0.25)	No threshold	>0.5	| log | undivided → source-normalized	One-ranking share
Original, single clause	Pythia-1.4B	24	313/336 = 93%	760/966 = 79%	214/221 = 97%	1.15 → 1.60	0.95 / 0.93
Original, single clause	Pythia-2.8B	32	340/349 = 97%	1005/1302 = 77%	209/211 = 99%	0.92 → 1.24	0.96 / 0.92
Original, single clause	Pythia-6.9B	32	333/362 = 92%	999/1302 = 77%	198/204 = 97%	0.92 → 1.27	0.91 / 0.80
Original, single clause	Llama-3.2-1B	16	126/133 = 95%	498/630 = 79%	66/66 = 100%	0.87 → 1.29	0.95 / 0.87
Original, single clause	Qwen3-1.7B	28	159/174 = 91%	836/1134 = 74%	81/86 = 94%	0.98 → 1.23	0.98 / 0.90
Original, multi-clause	Pythia-1.4B	24	96/103 = 93%	736/966 = 76%	18/18 = 100%	0.59 → 1.01	0.88 / 0.97
Original, multi-clause	Pythia-2.8B	32	79/87 = 91%	970/1302 = 75%	11/11 = 100%	0.31 → 0.47	0.72 / 0.95
Original, multi-clause	Pythia-6.9B	32	172/224 = 77%	889/1302 = 68%	40/50 = 80%	0.60 → 0.99	0.93 / 0.91
Independent filler–gap	Pythia-1B	16	51/55 = 93%	229/315 = 73%	27/28 = 96%	1.55 → 2.64	0.98
Independent filler–gap	Pythia-1.4B	24	99/107 = 93%	356/483 = 74%	26/28 = 93%	1.15 → 2.29	0.95
NPI licensing	Pythia-1B	16	75/85 = 88%	349/448 = 78%	20/25 = 80%	0.65 → 0.91	0.75
NPI licensing	Pythia-1.4B	24	131/140 = 94%	498/672 = 74%	24/24 = 100%	0.77 → 0.98	0.79
NPI licensing	Pythia-2.8B	32	149/162 = 92%	661/896 = 74%	37/37 = 100%	0.77 → 1.06	0.92
NPI licensing	Pythia-6.9B	32	146/164 = 89%	615/896 = 69%	33/33 = 100%	0.77 → 1.02	0.77

Table 15: Other intervention positions and sites.

Pythia-1.4B, single clause, all layers	Volume names arrow	One-ranking share	Beats random
position: noun	342/372 = 92%	0.99 / 0.99	95%
position: readout word	631/752 = 84%	0.96 / 0.96	100%
site: attention output (article)	41/44 = 93%	0.88 / 0.82	80%

B.8 FIXING THE STEP

Table 17 gives the same step both ways for every combination. The log asymmetry shrinks in 22 of 22 syntax matrices (mean 55%, 31–78%) and in both refusal models, and the relative asymmetry $|T(A{\to}B) - T(B{\to}A)|/(T(A{\to}B) + T(B{\to}A))$ shrinks in 22 of 22 (mean 51%). The shrinkage is not a product of taking the maximum over layers: layer by layer, in the eight original-template combinations, the log asymmetry shrinks in 391 of 419 layer–animacy groups (median 45%, mean 40%). What is left is still mostly a ranking (share 0.72–0.99), now at the source end: the source term varies more than the recipient term in 21 of 22 matrices and falls with volume in 22 of 22, since directions learned on small-volume constructions push harder at an equal step. Section 4 identifies it with how central a direction is (197 of 213 layer–animacy groups) and relates centrality to frequency: across the seven constructions, centrality correlates positively with frequency in 359 of 424 layer groups (median $+0.25$), and in 404 of 424 without the relative clause (median $+0.79$). Mean centrality ranks (1 = most central) are finite embedded wh 1.8, non-finite embedded wh 2.2, wh-question 3.7, cleft 4.2, pseudo-cleft 5.1, topicalization 5.3 and relative clause 5.7.

Handing the step to the source instead (Table 18) shrinks the relative asymmetry in all 14 combinations, by 56% on average (28–76%), and the ranking moves to the source's end: its correlation with volume is positive in 21 of 22 groups at the natural step and negative in 16 of 22 at the source's step. Writing one absolute step into every cell at layer 5 (Table 19) shrinks the relative asymmetry by 40–51%, what is left again falls with the source's volume, and syntactic directions still beat every random direction of the same step in 84 of 84 cells.

Is the natural ranking set by the recipient's volume or by the source's centrality, given that small-volume constructions learn more central directions? Table 16 separates the two. Holding centrality fixed, volume still tracks the ranking (partial rank correlation $+0.82$, positive in 166 of 175 groups); holding volume fixed, centrality adds nothing (-0.05). Writing the same step both ways flips the direction of only 76 of 826 arrows that are not near-ties: what is left sits at the source end and runs the same way, since small-volume constructions learn more central directions that push harder.

At the same step, the remainder has the direction of the natural arrows. Volume and centrality correlate negatively or not at all in 177 of 213 layer groups (combination medians -0.61 to -0.04), so at a fixed step small-volume constructions push harder, while at the natural step large-volume ones receive larger steps. The remainder at the same step therefore points the same way as the arrows at the natural step in 750 of 826 arrows that are not near-ties. Across the original templates, non-finite embedded wh-questions have mean centrality rank 2.2 of 7 and pseudo-clefts 5.1, so the running pair is of this kind; the relative clause is the least central (5.7). Refusal is the limiting case: its six categories share one direction and nearly one volume, so their two transfers differ only 1.07- to 1.09-fold on average.

Table 16: Volume, not the centrality of the learned direction, sets the natural ranking: eight original-template combinations, layer–animacy groups with all 42 off-diagonal transfers positive; correlations over the seven constructions; partial rank correlations residualize the ranks on the other variable.

Ranking potential at the natural step against	Median Spearman / partial	Positive groups
Volume (log)	+0.79	174/175
Centrality of the direction	−0.36	39/175
Volume, centrality held fixed (partial)	+0.82	166/175
Centrality, volume held fixed (partial)	−0.05	74/175

Table 17: Writing the same step both ways (layer-max matrix; animate / inanimate halves separated by a slash). Same step: the median of the constructions' own steps at each layer. Beats random: share of (source, target, layer) cells where the syntactic direction, written with the source's own step, beats three random directions written with the same step.

Combination	Log asymmetry: natural → same step	One-ranking share: natural → same step	Ranking vs. volume, same step	Source term · recipient term vs. volume, same step	Beats random
Single clause, Pythia-1.4B	1.14 → 0.43 / 1.16 → 0.52	0.95 → 0.95 / 0.93 → 0.87	+0.54 / +0.89	−0.54 · +0.32 / −0.86 · +0.18	90%
Single clause, Pythia-2.8B	0.84 → 0.37 / 1.00 → 0.57	0.96 → 0.88 / 0.92 → 0.94	+0.54 / +0.86	−0.57 · +0.11 / −0.82 · +0.21	90%
Single clause, Pythia-6.9B	0.82 → 0.33 / 1.03 → 0.49	0.91 → 0.90 / 0.80 → 0.72	+0.21 / +0.82	−0.43 · −0.46 / −0.68 · −0.04	90%
Single clause, Llama-3.2-1B	0.92 → 0.52 / 0.83 → 0.48	0.95 → 0.86 / 0.87 → 0.90	+0.54 / +0.71	−0.61 · +0.57 / −0.75 · −0.39	88%
Single clause, Qwen3-1.7B	0.77 → 0.33 / 1.19 → 0.64	0.98 → 0.90 / 0.90 → 0.96	+0.29 / +0.54	−0.46 · −0.32 / −0.61 · +0.39	92%
Multi-clause, Pythia-1.4B	0.45 → 0.30 / 0.73 → 0.36	0.88 → 0.97 / 0.97 → 0.96	+0.82 / +0.57	−0.82 · +0.54 / −0.43 · +0.57	88%
Multi-clause, Pythia-2.8B	0.30 → 0.11 / 0.32 → 0.11	0.72 → 0.89 / 0.95 → 0.89	+0.00 / −0.11	−0.61 · −0.93 / −0.32 · −0.21	90%
Multi-clause, Pythia-6.9B	0.74 → 0.30 / 0.47 → 0.32	0.93 → 0.99 / 0.91 → 0.97	−0.43 / +0.00	−0.21 · −0.50 / −0.04 · −0.29	89%
Indep. filler–gap, Pythia-1B	1.55 → 0.34	0.98 → 0.83	+0.71	−0.82 · −0.61	82%
Indep. filler–gap, Pythia-1.4B	1.15 → 0.51	0.95 → 0.92	+0.46	−0.86 · −0.39	90%
NPI, Pythia-1B	0.65 → 0.18	0.75 → 0.77	+0.62	−0.29 · +0.19	93%
NPI, Pythia-1.4B	0.77 → 0.30	0.79 → 0.72	+0.74	−0.10 · +0.55	93%
NPI, Pythia-2.8B	0.77 → 0.30	0.92 → 0.87	+0.88	−0.79 · +0.38	96%
NPI, Pythia-6.9B	0.77 → 0.31	0.77 → 0.79	+0.88	−0.31 · +0.71	95%
Refusal, Llama-3-8B-Instruct	0.07 → 0.03	0.99 → 0.98	– (six categories of almost equal volume)	–	100%
Refusal, Gemma-2-9B-it	0.09 → 0.02	0.99 → 0.99	– (six categories of almost equal volume)	–	100%

Table 18: Handing the step to the source: $A{\to}B$ written with A's own step on its own sentences. RA: relative asymmetry $|A{\to}B - B{\to}A|/(A{\to}B + B{\to}A)$, averaged over the animate and inanimate halves; the ranking column gives the two halves separated by a slash. Beats random as in Table 17.

Data	Model	|RA|: natural → source's step	Ranking vs. volume: natural → source's step	Beats random
Original, single clause	Pythia-1.4B	0.46 → 0.17	+0.96 → −0.54 / +0.89 → −0.18	90%
Original, single clause	Pythia-2.8B	0.38 → 0.12	+0.89 → +0.11 / +0.82 → +0.68	90%
Original, single clause	Pythia-6.9B	0.39 → 0.16	+0.82 → −0.57 / +0.54 → −0.54	90%
Original, single clause	Llama-3.2-1B	0.38 → 0.13	+0.89 → −0.36 / +0.96 → −0.07	88%
Original, single clause	Qwen3-1.7B	0.40 → 0.28	+0.75 → +0.07 / +0.82 → +0.64	92%
Original, multi-clause	Pythia-1.4B	0.28 → 0.20	+0.57 → −0.61 / +0.71 → −0.68	88%
Original, multi-clause	Pythia-2.8B	0.15 → 0.11	+0.32 → −0.54 / +0.11 → −0.71	90%
Original, multi-clause	Pythia-6.9B	0.28 → 0.12	+0.00 → −0.82 / +0.14 → −0.71	89%
Independent filler–gap	Pythia-1B	0.58 → 0.14	+0.64 → +0.07	82%
Independent filler–gap	Pythia-1.4B	0.49 → 0.18	+0.71 → +0.11	90%
NPI licensing	Pythia-1B	0.30 → 0.11	+0.88 → −0.43	93%
NPI licensing	Pythia-1.4B	0.34 → 0.17	+0.86 → −0.24	93%
NPI licensing	Pythia-2.8B	0.34 → 0.11	+0.86 → −0.48	96%
NPI licensing	Pythia-6.9B	0.35 → 0.12	+0.86 → −0.24	95%

Table 19: One absolute step for every source–recipient cell, and control directions (layer 5, the layer the original pipeline selects). The two control directions are fitted, like the syntactic ones, on the original pipeline's two control tasks: a grammatical control without a filler–gap dependency and a lexical control.

Layer 5, Pythia	1.4B	2.8B	6.9B
|RA|: natural → one absolute step (shrink)	0.50 → 0.25 (51%)	0.40 → 0.24 (40%)	0.49 → 0.25 (50%)
Arrows to the larger-volume end, natural → one absolute step (42 pairs)	38 → 29 (69%)	38 → 30 (71%)	34 → 27 (64%)
Same, pairs with a clear residual asymmetry	19/20	20/20	16/25
Residual in the source term: Spearman of source term with the source's volume	−0.60	−0.65	−0.54
Syntactic direction beats every random direction of the same step	84/84	84/84	84/84
Syntactic directions (mean of six other sources) beat two control directions of the same step: targets · cells	13/14 · 74/84	14/14 · 73/84	12/14 · 53/84

B.9 STEP ACCOUNTING AND TRAINING SEEDS

Table 20 accounts for the step at layer 5, the layer the original pipeline selects, in the three Pythia models, and checks stability over training seeds. Volume, on the input side, names more arrows than the margin between the two labels on the output side. With each written step divided by the source's own step, Eq. 4 predicts that the two-way ratio is the squared ratio of the volumes and the product the squared cosine of the two directions; the fitted exponent is 2.4–2.6 (R^2 0.79–0.94) and the Pearson correlation with the squared cosine 0.81–0.89. The response to the step is superlinear: the full-step effect exceeds twice the half-step effect in 123–142 of the 144 cross-construction cells at layer 5 (Table 20), and in a separate three-point map (zero, half and full step; four source directions, nine targets) in 36 of 36 cells, where proportional extrapolation from the half step misses by 0.18 on average, against 0.02 for a piecewise fit.

Table 20: Step accounting at layer 5 and stability over training seeds.

Layer 5, Pythia	1.4B	2.8B	6.9B
Arrows named by volume (input side) · by the output label margin (42 pairs)	38 · 21	38 · 23	34 · 18
Ratio of the two written steps (each over the source's own step) against the ratio of volumes, log–log: R^2 · fitted exponent	0.94 · 2.6	0.79 · 2.4	0.91 · 2.6
Product of the two written steps (each over the source's own step) against the squared cosine of the two directions: Pearson	0.85	0.89	0.81
Transfer = source term + recipient term: recipient share, natural → source's step → one absolute step	0.64 → 0.29 → 0.21	0.65 → 0.44 → 0.24	0.69 → 0.37 → 0.25
Spearman of the recipient term with the recipient's volume, natural step	+0.72	+0.68	+0.52
Superlinear response: full-step effect $>$ 2 $\times$ half-step effect (144 cross-construction cells)	142/144	123/144	136/144
Refit with 3 training seeds: lowest Spearman over 84 cells · pairs whose arrow flips	0.996 · 0	0.982 · 5	0.996 · 3
Two more seeds (43, 44), four layers per model (1.4B: 3, 9, 13, 19; 2.8B, 6.9B: 4, 12, 18, 26): arrows clear under both seeds with the same sign · arrows named by the step · Spearman of transfers with the original seed	100/102 · 102/110 · 0.60–1.00	89/89 · 106/108 · 0.71–1.00	103/103 · 98/105 · 0.90–1.00

B.10 THE TWO DIVISIONS

Dividing by the source's diagonal enlarges the log asymmetry in all 30 released conditions, and dividing by the recipient's diagonal shrinks it in all 30, to 0.27–0.91 of its size (median 0.69; Table 21). The authors' division also spreads the ranking in 60 of 60 matrices; the recipient's compresses it in 59 of 60. Per matrix, dividing by the source's diagonal raises the asymmetry by 1.14 to 2.09 times and dividing by the recipient's lowers it to 0.21–0.98 of its size. For the running pair the two diagonals are 11.70 and 6.72, a ratio of 1.74.

Table 21: The two divisions on the released matrices: range over positions of the factor by which the mean $|\log(A{\to}B/B{\to}A)|$ changes, with animate and inanimate constructions pooled (per matrix, with the two halves apart, the source division gives 1.14–2.09×).

Data	Model	Positions	| log | asymmetry, over source (authors) / undivided	Over recipient / undivided
Single clause	Pythia-1.4B	3	1.22–1.86×	0.44–0.84×
Single clause	Pythia-2.8B	3	1.40–1.98×	0.37–0.81×
Single clause	Pythia-6.9B	3	1.33–1.78×	0.77–0.85×
Multi-clause	Pythia-1.4B	7	1.19–1.64×	0.45–0.91×
Multi-clause	Pythia-2.8B	7	1.32–1.83×	0.27–0.78×
Multi-clause	Pythia-6.9B	7	1.34–1.78×	0.46–0.84×

B.11 RELATIONS: FURTHER READOUTS

In the matrix of Figure 1b, the five strongest relations are finite with non-finite embedded wh (0.73), pseudo-cleft with wh-question (0.35), finite embedded wh with pseudo-cleft (0.25), non-finite

embedded wh with pseudo-cleft (0.24) and cleft with topicalization (0.23). The running pair, 5.72 against 0.76, has relation 0.24, fourth of 21.

Over the 60 released matrices, the original authors' three observations are counted as holding when pseudo-cleft and wh-question are among the three most related of the 21 pairs; when topicalization's mean relation to the three most frequent constructions (relative clause and the two embedded wh) is below its mean relation to the other three; and when the relative clause has the lowest mean relation to the other six constructions.

In the eight original-template combinations, the relation increases with the overlap of the two learned directions, $|\cos(u_A, u_B)|$, with per-combination medians of $+0.58$ to $+0.90$ and positive in 173 of 175 layer–animacy groups; against shuffled construction labels, 155 of the 175 are significant at $p < 0.05$. The size of the arrow does not follow the overlap (medians -0.30 to $+0.06$). The undivided two-way mean also follows it (medians $+0.54$ to $+0.87$), and the relation follows it more closely in 117 of 175 groups.

Table 22 sets the relation against four other symmetric readings on the four known answers. The undivided readings recover the original authors' judgments at least as well, since those do not depend on each construction's scale, but they put two of Leng & Xiong's designed pairs last and trail on Taskonomy and on direction overlap. How much the division helps follows how far the diagonals differ: ordered by the ratio of their largest to their smallest diagonal (2.2, 2.7, 4.9 and 14.3), Leng & Xiong's four matrices put their worst designed pair 3rd, 8th, 14th and 15th undivided and 3rd, 3rd, 5th and 6th by the relation. A fit that removes one source and one recipient effect from the off-diagonal entries alone does best on the task answers, 0.98 on Taskonomy, but by absorbing each condition's overall sharing into its own effects it asks a different question, what a pair shares beyond its members' averages: it no longer sees that the relative clause shares little with anyone (5 of 60 matrices) and follows direction overlap only weakly ($+0.41$). Dividing by the diagonals matters most where conditions differ in scale, as Leng & Xiong's datasets do.

Table 22: The relation against four other symmetric readings on the four known answers. Intervals from resampling tasks (Taskonomy, 2,000 draws) and layer–animacy groups (direction overlap, 175 groups with all 49 transfers positive). The last reading fits one source and one recipient effect to the logs of the off-diagonal entries and averages the two residuals of a pair.

Symmetric reading	Leng & Xiong pairs: mean rank · top three	Taskonomy AUC: vs. scratch (95% interval) · vs. task-specific	Authors' three observations hold (of 60)	Spearman with $\lvert\cos\rvert$, median (95% interval)
Relation (product over the diagonals)	2.4 · 10/12	0.85 (0.64–0.98) · 0.82	49 · 59 · 48	+0.78 (+0.74–+0.80)
Undivided geometric mean	4.3 · 9/12	0.80 (0.49–0.96) · 0.71	60 · 60 · 52	+0.73 (+0.69–+0.75)
Arithmetic mean	4.3 · 9/12	0.80 (0.50–0.96) · 0.71	60 · 60 · 53	+0.68 (+0.65–+0.70)
Smaller direction	3.5 · 9/12	0.84 (0.64–0.96) · 0.71	59 · 57 · 52	+0.75 (+0.71–+0.78)
Off-diagonal source + recipient fit, symmetric residual	2.0 · 12/12	0.98 (0.92–1.00) · 0.97	37 · 54 · 5	+0.41 (+0.37–+0.46)

In refusal the six categories largely share one direction: mean $|\cos|$ between category directions is 0.80–0.93 across layers in Llama-3-8B-Instruct and 0.58–0.90 in Gemma-2-9B-it. Their volumes are nearly equal (mean log volume ratio 0.006–0.25 and 0.040–0.22) and so are the two directions of each pair (relative asymmetry 0.03–0.16 and 0.04–0.24, against per-layer medians of 0.41–0.65 in the 14 syntax combinations). The relation follows $|\cos|$ in 9 of 9 layers ($+0.19$ to $+0.88$) and 8 of 9 layers (-0.22 to $+0.95$).

B.12 Other groups' matrices

Table 23 gives cycles and ranking shares per paper; 49 of the 68 matrices contain no cycle at all. Table 24 gives the ranks of the dataset pairs that Leng & Xiong (2025) built for the same task, in their four matrices (Llama-2-7B; sentiment AmazonFood / SST-2, paraphrase QQP / PAWS, inference MNLI / GPTNLI; summarization CNN / XSum, question generation SciQA / TweetQA, data-to-text E2E / CommonGen).

Table 23: Other groups' published matrices. Cycles: cyclic triples among triples with no tie. Ranking share computed on two-way differences (many effects are zero or negative).

Paper	Intervention	Matrices	Cycles	No cycle	Ranking share (flips)
Leng & Xiong 2025 (tasks)	ablate task neurons; fine-tune only them	4	1/64	3/4	0.86 (0.33)
Tang et al. 2024 (languages)	ablate language neurons or regions	13	3/244	11/13	0.64 (0.29)
Xue et al. 2026 (languages)	inject function vectors; switch demonstration language	7	0/70	7/7	0.97 (0.40)
Liu et al. 2026 (languages × figures of speech)	add steering vectors	16	4/246	12/16	0.82 (0.33)
Nguyen et al. 2026 (prompt cues, datasets)	add steering vectors	6	0/15	6/6	0.97 (0.58)
Ball et al. 2026 (jailbreak types)	subtract jailbreak vectors	4	4/124	3/4	0.48 (0.15)
Shao et al. 2026 (tasks)	transfer circuits across models (learned alignment)	18	22/209	7/18	0.63 (0.36)
Total		68	34/972 = 3.5% (random 25%)	49/68	0.74 (0.33)

Table 24: Dataset pairs that Leng & Xiong (2025) built for the same task (classification: sentiment, paraphrase, inference; generation: summarization, question generation, data-to-text).

Matrix (Llama-2-7B)	Relation: ranks of the 3 designed pairs among 15	Undivided geometric mean: same
Classification, neuron ablation	1, 2, 3	1, 2, 8
Classification, fine-tune only these neurons	1, 2, 6	1, 2, 15
Generation, neuron ablation	1, 2, 3	1, 2, 3
Generation, fine-tune only these neurons	1, 2, 5	1, 2, 14

B.13 TASKONOMY

Table 25 applies both readings to three readouts in three released settings. Four tasks of the 2D family (autoencoding, denoising, 2D edges, 2D keypoints) never beat the task-specific network with their own encoder, so their diagonal is zero in that readout and the relation is undefined for them; the last two columns therefore also leave out the whole 2D family (17 tasks). The AHP affinity is already normalized per target by the paper's pairwise comparisons, and its undivided two-way mean ranks same-family pairs first by itself; on the raw win rates, one division by the diagonals reaches the same level.

Table 25: Taskonomy (Zamir et al., 2018): official first-order transfer results (22 tasks both source and target). AUC: probability that a same-family pair has a larger relation than a different-family pair; families as colored in the paper's Figure 14 (denoising placed with 2D). Settings: transfer trained on 16k or 1k images, and the repository's files without a size suffix, whose setting is not documented.

Training size	Readout	Cycles	One-number share (flips)	AUC relation	AUC undivided	Without 2D: cycles · share	Without 2D: relation vs. undivided
16k (paper's setting)	win rate vs. scratch	52/1540 = 3.4%	0.96 (0.09)	0.85	0.80	7.6% · 0.79 (0.12)	0.97 vs. 0.91
16k (paper's setting)	win rate vs. task-specific	7/1540 = 0.5%	0.99 (0.09)	0.82	0.71	0.9% · 0.98 (0.12)	0.89 vs. 0.73
16k (paper's setting)	AHP affinity (first order, renormalized)	18/1540 = 1.2%	0.96 (0.09)	0.83	0.86	2.6% · 0.93 (0.12)	0.85 vs. 0.95
1k	win rate vs. scratch	45/1540 = 2.9%	0.99 (0.09)	0.82	0.74	6.6% · 0.69 (0.12)	0.95 vs. 0.79
1k	win rate vs. task-specific	3/1540 = 0.2%	0.99 (0.09)	0.80	0.63	0.3% · 0.99 (0.12)	0.87 vs. 0.58
1k	AHP affinity (first order, renormalized)	12/1540 = 0.8%	0.97 (0.09)	0.77	0.81	1.5% · 0.95 (0.12)	0.77 vs. 0.86
no suffix	win rate vs. scratch	45/1500 = 3.0%	0.96 (0.09)	0.85	0.81	6.9% · 0.75 (0.12)	0.97 vs. 0.92
no suffix	win rate vs. task-specific	11/1520 = 0.7%	0.99 (0.09)	0.82	0.69	1.3% · 0.97 (0.12)	0.89 vs. 0.70
no suffix	AHP affinity (first order, renormalized)	21/1540 = 1.4%	0.95 (0.09)	0.84	0.86	2.2% · 0.91 (0.12)	0.86 vs. 0.94

B.14 Coordinated verb phrases

Table 26: Coordinated verb phrases of Boguraev & Mahowald (2026), Experiment 2 (GPT-2, Pythia-410M, Qwen2-0.5B × 5 seeds × 3 sites × all layers). Responsive cells: (model, site, layer), seed-averaged, whose transfer correlates with the wh-licensing score at 0.3 or more. The licensing score is the model behaviour of the original Experiment 1 (correlation with human acceptability 0.54–0.80).

Intervention position	Responsive cells	Transfer–licensing correlation, matched / natural	Transfer vs. human ratings (structure level): natural → matched	Step into the conjunct vs. human ratings (no transfer)
"what"/"that" (step equal by construction)	84	1.01 (Qwen2-0.5B 1.01; GPT-2 1.00; Pythia-410M 1.02)	0.69 → 0.69	– (same step for every construction)
Subject (step equal by construction)	67	1.01 (Qwen2-0.5B 1.01; GPT-2 1.05; Pythia-410M 1.01)	0.56 → 0.57	– (same step for every construction)
Last word of the conjunct (step varies)	138	0.45 (Qwen2-0.5B 0.40; GPT-2 0.45; Pythia-410M 0.46)	0.72 → 0.33	0.73 (Qwen2-0.5B 0.71; GPT-2 0.70; Pythia-410M 0.77); partial correlation of transfer given the step 0.11, ΔR^2 0.02

Table 26 gives the three intervention positions of Figure 4. Of the 180 (model, site, layer) cells at each position, 84, 67 and 138 are responsive. Where the step is equal by construction, matching it changes nothing and transfer keeps its correlation with human ratings. At the conjunct's last word, matching the step leaves 0.45 of the correlation between transfer and the licensing score and lowers the correlation with human ratings from 0.72 to 0.33, and the step written into each conjunct correlates with human ratings at 0.73 without any transfer.

B.15 PROJECTION ABLATION: THE SOURCE SETS THE AMOUNT

In projection ablation the removed amount belongs to the source. We reimplemented the refusal-category cross-effect maps of Sharma et al. (2026) following their method, eighth layer from the end, in Gemma-2-9B-it, Llama-3-8B-Instruct and Mistral-7B (Jiang et al., 2023). The amount removed varies mainly with the source subspace (source share of its variance 0.99, 0.56 and 0.90), and so does the map (source share 0.99, 0.64 and 0.96). Interchange intervention is the mirror image: the recipient sets the step, and the recipient term carries 0.64–0.69 of the map (Table 20). Our cell values do not match their published ones (Pearson -0.50 to $+0.05$), consistent with the instability of category subspaces under resampling that they report.

B.16 CHECKS ON THE VOLUME

Not a tokenization artifact. Editing the templates so that capitalisation and spacing at the intervention word form true minimal pairs, pseudo-clefts still have the largest representational distance between their two sentences of the seven constructions in 100% of layer–animacy groups.

Not the literal difference. The number of tokens by which a construction's two sentences differ before the intervention word runs against its volume: per-layer median Spearman -0.55 to -0.44 in the five single-clause combinations, negative in 127 of 132 combination–layers. Clefts differ by 5.5 tokens and have among the smallest volumes; pseudo-clefts differ by one and are the largest volume at layer 5 in all five models, and the largest by layer-max volume in 10 of 10 single-clause groups and in both of the independent filler–gap combinations.