CLIMBING THE CAUSAL LADDER: TWENTY LANGUAGE MODELS SHARE A CAUSAL VARIABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Do different language models compute with the same variables? Universality is a founding hypothesis of mechanistic interpretability, yet cross-model evidence (similar representations, transferable steering vectors) shows that models are alike, not that they compute the same causal variable. We give a test that can establish sameness. It grades evidence by Pearl's ladder of causation and requires invariance across the behaviours that read a variable and across models, through maps that never see it. Applied to grammatical number in 20 language models from five families (70M–8B), it finds a causal variable shared across all of them, with evidence at every rung. Counterfactually, interchanging a direction estimated from labels alone at the subject noun mediates a median 0.85 of the full residual state's effect on five held-out agreement behaviours (0.86 in natural sentences), without moving gender. Interventionally, graded edits move every behaviour proportionally, and ablating the direction removes a median 0.53 of the number signal, against 0.01 for the gender direction. Across models, affine maps fitted only on generic text carry it with median efficiency 0.97 within and 0.95 across families (0.99 in natural sentences), while shuffled-pair maps, an untrained source and random directions carry nothing. The test also delimits and discriminates: the shared variable is the noun's own number, not number cued by a determiner (*these sheep*), and gender, causal in every model, is shared less reliably. Rankings of directions by behavioural strength, the field's usual evidence, reverse under calibration, recovery level and fitting cohort in our preregistered experiments. Invariance, not effect size, identifies the variables models share.

1 INTRODUCTION

Mechanistic interpretability rests on two bets: that networks compute with variables we can recover, and that different networks compute with the same ones (Olah et al., 2020; Li et al., 2016; Huh et al., 2024). The second bet, universality, is what would let interpretability scale: a variable identified once could be located in every new model without labels, and findings would accumulate across models instead of restarting with each one. Yet the field has no agreed standard of evidence for either bet. A direction counts as a concept when moving activations along it changes a behaviour, whether it came from probing, class means, dictionary learning or optimization (Turner et al., 2023; Zou et al., 2023; Li et al., 2023; Arditi et al., 2024; Cunningham et al., 2024). That evidence is weaker than it looks: subspace patches can succeed through pathways the model does not use (Makelov et al., 2024), flexible alignment maps find causal abstractions even in untrained networks (Sutter et al., 2025), and in eight of our preregistered studies, rankings of directions by behavioural strength reversed under choices unrelated to the variable (Section 9). Cross-model evidence is weaker still. Similar representations (Kornblith et al., 2019; Huh et al., 2024), correlated neurons and features (Gurnee et al., 2024; Lan et al., 2024) and steering vectors that still steer after transfer (Chen et al., 2025; Oozeer et al., 2025) show that models are alike; none shows that a direction found in one model is the causal variable another model computes with.

We ask what evidence licenses the claim that a direction *is* a variable, and that two models share it, and organize the answer by Pearl's ladder of causation (Pearl, 2009; Bareinboim et al., 2022) (Figure 1, Table 1). Rung 1 is association: a probe or class-mean difference separates the variable's

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.

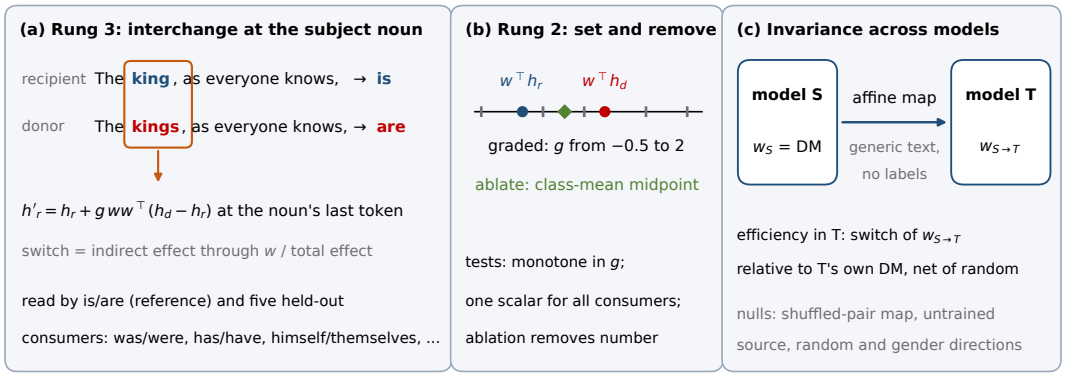


Figure 1: The tests. (a) Rung 3: the coordinate of the subject-noun representation along a direction w is set to its value under a donor that differs only in number, and five consumers that w was never fitted to are read out. (b) Rung 2: the coordinate is set at graded gains and removed. (c) Invariance: a source model's direction is carried into a target model by an affine map fitted on generic text without labels and tested in the target, against a shuffled-pair map and an untrained source.

values. Rung 2 is intervention: setting the coordinate along the direction changes behaviour in a graded way, and removing it removes the behaviour. Rung 3 is counterfactual: what the output would have been for input x had the coordinate taken the value it takes under a counterfactual input x', with everything else at its actual value. An interchange intervention computes exactly this cross-world quantity (Geiger et al., 2021; 2024); normalized by patching the entire state at the same site, it is the share of the site's counterfactual effect that the coordinate mediates (Pearl, 2001; Vig et al., 2020). In a network every rung is computable exactly, so the open question is not identification but abstraction: does the proposed variable reproduce the model's counterfactual behaviour for every behaviour that should read it, and for nothing else? To this we add the criterion that causal inference uses to separate causes from correlates, invariance (Woodward, 2003; Peters et al., 2016): a causal variable's effect is stable across the environments in which it acts. Our environments are the behaviours that consume the variable and the models that compute it; the maps between models never see the variable, and an untrained network must not supply it.

We apply this test to grammatical number, whose values, consumers and confounds are known, in 20 pretrained models from Pythia, Qwen3, SmolLM2, GPT-2 and OLMo-2 and in Pythia-410M training checkpoints, with every study frozen before evaluation. A direction estimated from labels alone at the subject noun earns evidence at every rung (Section 5): it mediates most of the full state's counterfactual effect on five consumers it never saw, in templates and in natural sentences, leaves gender untouched, moves all consumers proportionally under graded edits and, when ablated, removes a large share of the number signal. The same variable is present in every model (Section 6): through maps fitted only on generic text, each model's direction acts in the others, across families, at nearly the strength of the target's own, while shuffled maps, an untrained source and random directions fail. The test also says what the variable is not (Sections 7–8): it carries the noun's own number and almost none of the number cued by a determiner, so what twenty models share is a lexical variable read by every agreement consumer, not abstract phrasal number; and at the same token it returns a different verdict for gender, causal in every model but only partly necessary there and less exactly shared.

Contributions. (i) **A standard of evidence for concept claims**: a preregistered battery that places a claim on Pearl's ladder and tests invariance across consumers and models, with label-free transport and map nulls. (ii) **Causal universality at scale**: across 20 models and five families, noun number is one causal variable, carried between models at nearly native strength, in templates and natural sentences, learned early in training and absent from an untrained network. (iii) **Scope and discrimination**: the test delimits the shared variable (lexical, not phrasal, number) and separates two variables read at the same token. (iv) **Why effect size misleads**: rankings of directions by behavioural strength are unstable, whereas three estimators converge on one variable under the invariance tests. Failed and post hoc outcomes are reported, and every value is bound to a hashed analysis record.

2 RELATED WORK

Concepts as directions, and their evidence. Linear structure for concepts in embeddings and language models (Mikolov et al., 2013; Elhage et al., 2022; Park et al., 2024) underlies probing, steering and dictionary learning. Probes establish association, not use (Hewitt and Liang, 2019; Belinkov, 2022); amnesic probing and linear erasure test use by removing a direction (Elazar et al., 2021; Belrose et al., 2023; Lasri et al., 2022). Steering vectors (Turner et al., 2023; Zou et al., 2023; Li et al., 2023; Arditi et al., 2024) and sparse-autoencoder features (Cunningham et al., 2024) are validated by the behaviour they change; steering evaluations add calibration (Im and Li, 2025; Bao et al., 2026), matched effects (Aparin and Gaintseva, 2026) and output-geometry controls (Wang et al., 2026; Picozzi, 2026; Gao et al., 2026). Yet patches can act through dormant pathways (Makelov et al., 2024), localization need not predict where edits work (Hase et al., 2023), and patching conclusions depend on metrics and corruptions (Zhang and Nanda, 2024). We ask where such evidence sits on the causal ladder, and whether it is invariant.

Causal abstraction. Activation patching localizes behaviour in circuits (Meng et al., 2022; Wang et al., 2023); causal mediation analysis has localized gender bias and agreement mechanisms (Vig et al., 2020; Finlayson et al., 2021); interchange interventions test causal abstractions (Geiger et al., 2021), and distributed alignment search fits subspaces for them (Geiger et al., 2024). Benchmarks score whether a featurization changes its target attribute while sparing others (Huang et al., 2024), compare localization methods (Mueller et al., 2025) and include agreement (Arora et al., 2024); flexible alignment maps make causal abstraction attainable even in untrained networks (Sutter et al., 2025). These methods test a hypothesized variable in one model, usually on the behaviour it was fitted to; we add held-out consumers, isolation, necessity and invariance across models, with untrained and shuffled controls in the role that model randomization plays for saliency maps (Adebayo et al., 2018).

Universality and cross-model maps. Different networks learn related features (Li et al., 2016; Olah et al., 2020; Chughtai et al., 2023), with similar representations (Kornblith et al., 2019; Huh et al., 2024), correlated neurons (Gurnee et al., 2024) and matching dictionary features (Lan et al., 2024); circuits in Pythia persist across training and scale (Tigges et al., 2024), and grammatical features are shared across languages within a model (Brinkmann et al., 2025). Stitching layers and closed-form maps translate between latent spaces (Lenc and Vedaldi, 2015; Bansal et al., 2021; Moschella et al., 2023; Maiorca et al., 2023); affine maps transfer features, probes and steering vectors between model sizes (Chen et al., 2025), and learned maps transfer safety steering across families (Oozeer et al., 2025), though alignment can fail to carry causal effect (Zhang and Xin, 2026). These results measure similarity, or whether a transferred intervention remains useful. We test *sameness*: a transported direction must mediate the target's own counterfactual behaviour about as strongly as the target's native direction, on consumers neither was fitted to, while controls with the same flexibility fail.

Invariance and agreement. Invariance across environments marks causal rather than merely predictive structure (Woodward, 2003; Peters et al., 2016; Arjovsky et al., 2019; Schölkopf et al., 2021); we take behaviours and models as environments. Subject–verb agreement is a classic test of syntax in language models (Linzen et al., 2016; Lakretz et al., 2019), and number is linearly encoded and causally used within a model (Finlayson et al., 2021; Ravfogel et al., 2021; Lasri et al., 2022; Hao and Linzen, 2023). We use number because its values, consumers and confounds are known; our question is whether different models compute the same variable.

3 EVIDENCE ON THE CAUSAL LADDER

Let h_r and h_d be recipient and donor residual activations at a declared site. For a unit direction w and scalar gain g, the rank-one intervention is

$$h'_r = h_r + g\, ww^\top (h_d - h_r). \tag{1}$$

At $g = 1$ it sets the coordinate $w^\top h$ to the donor's value and keeps the orthogonal complement at the recipient's; replacing h_r by h_d is the *full patch*. The number readout at a consumer position is the

Table 1: Evidence for grammatical number on the causal ladder, over 20 models. Criteria and thresholds were frozen before evaluation; bounds are 95% family-bootstrap limits. Ordered pairs: 74 within and 306 across families.

Rung	Test and preregistered criterion	Result
1 Association	DM (class-mean difference) and a logistic probe, from labels only; their cosine (descriptive)	DM–probe 0.96
2 Intervention	Graded: DM switch increases strictly with gain from –0.5 to 2 (in 90% of models)	not supported (15 of 20)
	One variable: consumers-by-gain switch matrix has rank-1 energy share at least 0.95 (90%)	20 of 20 (min 0.998)
	Necessity: mean-ablating DM removes at least half of what setting the full state halfway removes, lower bound (80%)	16 of 20
3 Counterfactual	Consumer generality: normalized switch on five held-out consumers, lower bound at least 0.5; random below 0.1	20 of 20; median 0.85
	Same test for FIT, fitted by interchange on the reference consumer only (descriptive)	median 1.06; cosine with DM 0.80
	Isolation: gender switch at most 0.2 of the full patch; gender direction moves number at most 0.2 of DM	20 of 20
Invariance	Transport efficiency of a source's DM in a target through a label-free map, lower bound at least 0.5 (P16 ridge map)	66 and 283 pairs
	Same, through standardized ridge (conditioning hypothesis: at least as many pairs, and half of the P16 failures rescued) and raw Procrustes (descriptive)	not supported (7 of 31 rescued); 326 of 380
	Nulls: untrained source below 0.25 in every target; shuffled-pair map below 0.25 in 95% of pairs	supported; supported
Natural text	Natural sentence contexts (P18): DM normalized switch lower bound at least 0.5 (80% of models); transported DM lower bound at least 0.5 (80% of pairs), both nulls below 0.25	supported (20 of 20); supported (353 of 380)
Scope	Determiner-cued number: DM normalized switch at least 0.5 (80% of models whose site carries at least 0.25)	not supported
	Cue, not noun class: on animal nouns marked on the noun, DM carries at least half and DM-det under a quarter (80%; post hoc)	not supported (8 of 20)

log-odds of plural over singular continuations (e.g. `are` versus `is`); the *switch* is the pooled change toward the donor divided by the clean donor-minus-recipient difference. Recipient and donor differ only in the noun's number, so the full patch's switch is the share of the total effect of number that passes through the site, the switch of w at $g = 1$ is the natural indirect effect through the coordinate relative to that total effect (Pearl, 2001), and their ratio, the *normalized switch*, is the share of the site's effect that the coordinate mediates. Table 1 lists the criteria, frozen with their thresholds before evaluation (P16–P20; Appendices A–E).

4 SETUP

Stimuli and consumers. Forty-six gendered human-noun families (*king, kings, queen, queens*) are split into fitting (14), development (8) and evaluation (24) families. Six consumers read number at least three tokens after the subject: present copula (`is`/`are`), past copula, perfect auxiliary, reflexive (`himself`/`herself` versus `themselves`), subject pronoun of a new clause and possessive, with three templates each. The present copula is the reference consumer; the other five are held out. Determiner-cued items use 24 animal nouns whose singular and plural are identical (*sheep*, *deer*, *salmon*), in the same templates, with number marked only by *this/these* or *that/those*.

Site and directions. The site is the final token of the subject noun at the deepest block where the full patch still switches the present copula by at least 0.8 on development families. On fitting

families only we estimate DM, the plural-minus-singular mean (labels, no behaviour); FIT, fitted to the present copula alone; a logistic probe; and GDM, the female-minus-male mean. Ten random directions give the baseline.

Maps. For each ordered pair of models, an affine map from source to target activations is fitted at the final token of 27,000 identical WikiText prefixes, each model at its own site; no stimulus, number or gender data enter it. The P16 map is a ridge regression on centred activations; P17 adds a ridge regression and an orthogonal Procrustes map on standardized activations, a raw Procrustes map, and a null map fitted after shuffling the target rows. A source direction is transported as a displacement and applied with the target's own donor activations. *Transport efficiency* is the transported direction's switch as a fraction of the target's own DM, both net of random directions.

Models and statistics. Pythia 70M–6.9B, Qwen3 0.6B–8B, SmolLM2 135M–1.7B, GPT-2 124M–1.5B and OLMo-2 1B and 7B, and Pythia-410M at five checkpoints from random initialization (Biderman et al., 2023). A model–consumer cell is eligible when clean accuracy is at least 0.8. Intervals are 95% percentile bootstraps over the 24 evaluation families (14 evaluation nouns for determiner items), 10,000 draws.

5 A CAUSAL VARIABLE WITHIN EACH MODEL

Rung 3: one direction mediates the counterfactual effect on every consumer. In every one of 20 models, interchanging DM alone switches the five held-out consumers by a median 0.85 of the full patch at the same site (range 0.70–0.98), and every lower bound exceeds 0.5; random directions reach at most 0.003 (Figure 2a). Every consumer is switched alike (per-consumer medians 0.84–0.87). DM moves gender by at most 0.019 of the full patch, whereas GDM switches gender (median 0.77) and moves number by at most 0.031 of DM's effect: a double dissociation in all 20 models. Directions estimated differently converge: FIT, fitted to the present copula only, generalizes to the held-out consumers (median 1.06; above one, a direction optimized for the switch overshoots the donor), as does the probe (0.92), and both lie close to DM (median cosines 0.80 and 0.96).

Natural sentences. Templates fix the syntax, so a preregistered extension (P18; Appendix C) took 266 sentence contexts from WikiText in which a noun-phrase head is followed, after at least one word (median 4), by *is/are*, *was/were* or *has/have*, and substituted an evaluation-family noun in both numbers (532 items, often semantically odd, which tests agreement rather than plausibility). With nothing refitted, DM mediates a median 0.86 of the full patch (range 0.67–0.98; lower bound at least 0.5 in 20 of 20 models), and 0.86 when three or more words intervene, while random directions stay below 0.024 and GDM moves number by at most 0.006.

Rung 2: graded, one-dimensional and necessary. Setting the coordinate to $h_r + g(h_d - h_r)$ along DM moves the held-out switch smoothly with g: from 0.03 at $g = 0.25$ through 0.21 at $g = 0.5$ and 0.80 at $g = 1$ to 0.94 at $g = 2$ (medians; Figure 2b), while a random direction stays at 0.00. The response increases strictly up to the donor's value in 20 of 20 models; beyond it, the switch saturates and dips by at most 0.04 at $g = 2$ in some models, as the full patch also does, so the preregistered strict increase over the whole grid holds in 15 of 20 (criterion not supported). The consumers respond in proportion: the consumers-by-gain matrix of switches has rank-1 energy share at least 0.998 in every model, so a single scalar describes how every consumer reads the edit. Removing the coordinate removes the behaviour: setting DM's coordinate to the midpoint of the fitting-family class means erases a median 0.53 of the signed number log-odds on held-out consumers, a median 0.73 of what moving the entire state halfway to the donor erases (0.76), whereas the same ablation of GDM erases 0.01 and of a random direction 0.00 (Figure 2c; necessity met in 16 of 20 models).

6 THE SAME VARIABLE ACROSS MODELS

Transport through maps that never see number. Carried by the P16 map, a source model's DM aligns with the target's own (median cosine 0.80 within and 0.77 across families) and keeps its causal effect: 66 of 74 within-family and 283 of 306 cross-family ordered pairs meet the invariance criterion, with median efficiencies 0.97 and 0.95 (Figure 3a; Table 2). In 27 within-family pairs the

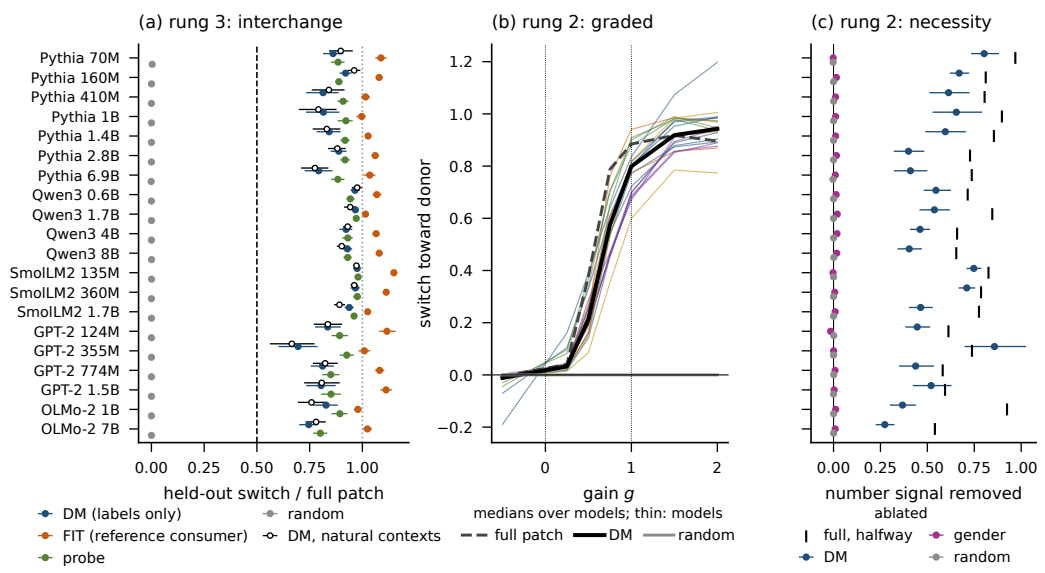


Figure 2: A causal variable within each model. (a) Rung 3: held-out switch of interchange interventions along DM, FIT, the probe and random directions, normalized by the full patch at the same site (95% family intervals; dashed line: preregistered 0.5); open circles: DM in the natural sentence contexts of P18. (b) Rung 2: switch on held-out consumers as the coordinate is set to $h_r + g(h_d - h_r)$; thin lines are models, coloured by family. (c) Rung 2: share of the signed number log-odds removed by mean-ablating DM, the gender direction or a random direction, with the reference of moving the full state halfway to the donor (ticks).

Table 2: Results by target family. Rung 3: models meeting consumer generality and the double dissociation. DM: held-out switch relative to the full patch. Necessity: share of number log-odds removed by mean-ablating DM. Transport: median efficiency of other models' number directions in this family's models, with pairs meeting the criterion, through the P16 ridge map (within and across families) and the raw Procrustes rotation (across). Determiner: DM's switch of determiner-cued number relative to the full patch.

					Transport efficiency			
Family	Models	Rung 3	DM	Necessity	Within	Across	Across, rot.	Determiner
Pythia	7	7/7	0.84	0.61	1.02 (36/42)	1.06 (82/91)	0.90 (85/91)	0.13
Qwen3	4	4/4	0.95	0.50	0.97 (12/12)	0.92 (64/64)	0.53 (29/64)	0.18
SmolLM2	3	3/3	0.97	0.71	0.95 (4/6)	0.90 (37/51)	0.76 (46/51)	0.34
GPT-2	4	4/4	0.81	0.48	0.92 (12/12)	1.00 (64/64)	0.93 (63/64)	0.09
OLMo-2	2	2/2	0.79	0.32	0.83 (2/2)	0.93 (36/36)	0.77 (30/36)	0.09

transported direction outperforms the target's own DM. Directions fitted to behaviour transport like label-derived ones (FIT: 1.03 within, 0.98 across), and the same transported directions carry number in natural sentences (median efficiency 0.99; lower bound at least 0.5 in 353 of 380 pairs, with all 27 failures in Pythia 160M, Pythia 2.8B and SmolLM2 1.7B).

Nulls. Transport requires both a correspondence and a trained source. Maps fitted after shuffling which target activation pairs with which source activation carry a median efficiency of 0.00 (upper bound below 0.25 in 380 of 380 pairs; largest 0.03). The number direction of the randomly initialized Pythia-410M, transported into every target, has median efficiency 0.01 under the P16 map (largest upper bound 0.14) and 0.01 under the standardized map (largest 0.14). Transported gender directions move number by a median 0.003 of the native effect (Figure 3b). In natural sentences both nulls fail again (largest upper bounds 0.03 for shuffled pairs and 0.16 for the untrained source).

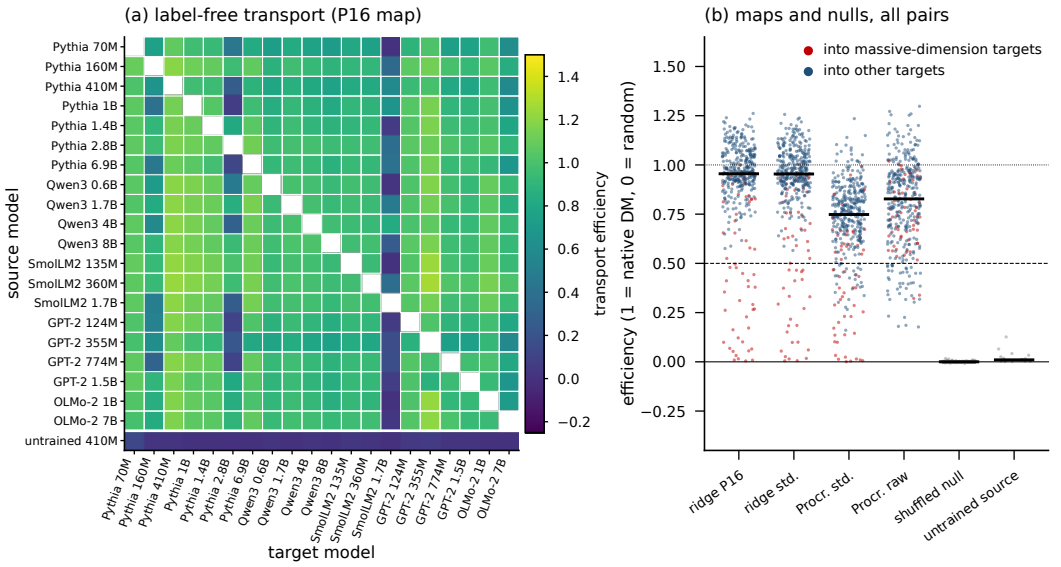


Figure 3: The same variable across models. (a) Transport efficiency of each source's DM (rows) in each target (columns) through the P16 map, fitted only on generic text; 1 equals the target's own DM and 0 equals random directions; outlined cells meet the preregistered criterion. The bottom row is the untrained Pythia-410M. (b) Efficiency over all 380 ordered pairs for four maps and two nulls (bars: medians; red: pairs into the three targets with a massive activation dimension).

Failures are map-specific. The P16 map's failures concentrate in three targets (Figure 3a) whose site activations on generic text are dominated by one massive dimension (Sun et al., 2024): SmolLM2 1.7B, Pythia 2.8B and Pythia 160M (variance share 0.92, 0.45 and 0.45, against at most 0.06 elsewhere), although that coordinate is causally inert for number (zeroing it leaves DM's switch at 1.00, 1.00 and 1.00 of its value). Standardizing activations before fitting, the preregistered remedy, rescues only 7 of the 31 failures (criterion not supported). An orthogonal Procrustes map fitted to the raw activations, which preserves angles and norms instead of predicting activations, rescues 27, lifting the three targets to 15, 17 and 19 of 19 sources, while the untrained source still carries nothing (post hoc control; largest upper bound 0.07); it in turn misses 50 pairs that the P16 map carries, 35 of them into Qwen3 targets. Every target accepts at least 15 of 19 sources under one of the two maps: failures reflect how well a map family fits a target's geometry, not an absent variable.

Sameness is not similarity. Transported directions are only moderately similar to the target's own (median cosine 0.77 across families) yet act with nearly its full effect, and across pairs efficiency is only weakly related to that cosine (Spearman 0.21), to map fit (0.01) or to model size ratio (–0.03). What models share is the causal role, not the geometry, so similarity measures understate it.

7 What the variable is, and is not

Not a suffix detector. DM remains effective for the three evaluation families with irregular plurals (*wives*, *gentlemen*, *businessmen*; median 0.91, against 0.85 for regular families), so the variable abstracts over morphology.

Not phrasal number. Were the variable abstract grammatical number, it would also carry number that the noun inherits from its determiner. Models agree correctly with determiner-cued subjects (*these sheep are*; median clean accuracy at least 0.91 per consumer), but the subject site carries little of that number (full-patch switch median 0.33), and DM carries a median 0.13 of what it does carry. A direction estimated from determiner-cued fitting nouns, DM-det, carries a median 0.82 of the site's effect and is nearly orthogonal to DM (median cosine 0.14), so the preregistered abstraction criterion is not supported: the verdict that separates lexical from phrasal number. A depth sweep (P17d) finds the same at every depth: determiner-cued number peaks at a median switch of 0.36, against 0.99 for

noun-marked number, and no model shares a direction between the two cues at any depth (Appendix Figure 9a,b). Holding noun class fixed shows that the cue decides: for animal nouns marked on the noun (*dog/dogs*), DM-det carries almost nothing (median 0.01) and an animal-noun direction aligns with DM (cosine 0.57) rather than DM-det (0.12); DM from human nouns carries a median 0.59 of it, and the declared criterion holds in 8 of 20 models (not supported; post hoc; Appendix B). Determiner-cued number reaches the consumers mostly through other positions, consistent with subject number spreading across positions when several cues mark it (Hao and Linzen, 2023).

Learned early. In Pythia-410M the variable is absent at initialization (clean agreement 0.49) and meets consumer generality from step 2,000; its transport efficiency into the trained models rises with training (Appendix Figure 9c; exploratory). The shared variable is a learned, lexical variable: the noun's own number, read by every agreement consumer and recoverable in every model we tested.

8 THE TEST DISCRIMINATES: GENDER AND NEGATION

Gender: causal in every model, less exactly shared. A test that passes everything would be worthless. At the same token, the battery returns a different verdict for gender, a variable long studied in embeddings and language models (Bolukbasi et al., 2016; Vig et al., 2020) (P19; Appendix D). GDM, the female-minus-male mean, meets rung 3 and the double dissociation (Section 5); setting its coordinate moves the gender switch smoothly (median 0.75 at $g = 1$; strictly increasing in 20 of 20 models), with a rank-1 consumer response (share at least 0.9997). But ablating GDM removes a median 0.45 of the signed gender log-odds, so necessity holds in only 7 of 20 models (criterion not supported); transported gender directions act with lower efficiency than number (P16 map: median 0.66 within and 0.60 across families), while the untrained source and shuffled-pair maps carry nothing (largest upper bounds 0.07 and 0.003); and the noun-derived direction carries a median 0.50 of the gender of first names (*John, Mary*; criterion not supported, met in 7 of 20). Number is necessary at its site and shared almost perfectly; gender is a graded causal variable in every model that other positions also carry and that models share less exactly.

Negation travels elsewhere. The battery's first requirement is that the site carry the variable, and it reports when one does not. For negation (P20; Appendix E), whose handling by language models has long been questioned (Ettinger, 2020; Kassner and Schütze, 2020), we kept number's site so that the same maps apply: the word that *not* scopes over, at the P16 block, read by negative-polarity licensing (*did not find any*) and explicit polarity (*and that is bad*). Unlike number at its noun, negation mostly bypasses that word at this depth: the full patch there carries at least a quarter of the negation effect in only 5 of 20 models (Pythia 70M, Pythia 160M, GPT-2 124M, GPT-2 774M and GPT-2 1.5B). Where it does, a single direction, NDM, mediates a median 0.62 of the site's effect (rung 3 in 4 of 5) and is isolated from number in both directions; transport could not be tested, since no preregistered source model carried negation at the site. Negation reaches its consumers through other positions, which is where a test of its universality must look.

9 BEHAVIOURAL STRENGTH DOES NOT IDENTIFY A VARIABLE

Eight earlier preregistered studies (P4–P14; Appendices G–Q) ranked rank-one directions for subject–verb agreement at a late site in Pythia-410M and Qwen3-0.6B by what their edits do: U, a normalized difference of means; CE, fitted to the donor's number label; and JS, fitted to the donor's full next-token distribution. P8 and P9 retain the P5/P6 fitted vectors (evaluated off their fitting construction); refitting studies (P12, P14) are reported separately. The rankings did not survive choices unrelated to the variable. Calibrating the gain reversed which direction best reproduces the donor's distribution (P4; Figure 10); the recovery level changed the CE–JS preservation gap (P9; Figure 11); and in a controlled refit the two fitting cohorts disagreed in sign, while an admissible alternative calibration selector, applied post hoc, reversed the sign of another interaction (P14; Figure 12). None of these rankings says what the directions represent. The invariance tests do: DM, FIT and the probe, three estimators that such rankings would separate, each meet consumer generality and isolation in all 20 models (Section 5). A subject-number direction at the late site also accounts for a classic error and, in Pythia, transfers across constructions: replacing only that component of an attractor's difference restores more than half of agreement attraction in all six evaluated conditions, including

Qwen3-1.7B (0.64 [0.59, 0.70]; P13, Appendix R), and a direction and response slope locked on source items predict responses in a new construction within a 0.10 margin in Pythia (0.055 [0.044, 0.066]) but not in Qwen (P15, Appendix S).

10 Discussion

What the rungs establish. Association found a candidate; intervention showed that its coordinate is graded, one-dimensional and necessary; counterfactual interchange showed that it mediates most of the effect of number on behaviours it never saw and almost none of the effect of gender; and invariance showed that it belongs to language models trained on text, not to one network, since label-free maps recover it across families while shuffled maps and an untrained network do not.

What changes for interpretability. *Universality becomes a causal claim.* Representational convergence (Huh et al., 2024) says that models are alike; we show that twenty models from five families compute a variable with the same causal role, recoverable by a map that knows nothing about it, and that similarity measures understate this sharing. *Interpretations can transfer.* A variable identified once, with labels, in one model can be located in another without labels and verified there against nulls; because no single map family suffices for every target, map choice is a model-selection problem decided by the transport criterion and its nulls. *Concept claims can be graded.* A claim should state the highest rung its evidence reaches, the consumers it was tested on, what it was isolated from and the inputs on which it was shown not to apply. The recipe is cheap: class means, held-out consumers, a neighbouring variable, graded and ablative edits, and generic-text maps with two nulls. *Tests delimit what names overreach.* "Number" in these models is the noun's own number, not phrasal number, and gender at the same token is causal but only partly necessary and less shared: distinctions invisible to behavioural strength, and exactly those a mechanistic account must respect.

Limitations. Number is tested with every criterion and gender with most; negation's inherited site carried too little negation to test transport. We study English, in templates and natural sentences with substituted nouns, at one site per model, where the subject position still carries number rather than where number is computed or finally read. Maps are affine and fitted on one corpus; efficiency above one means that a transported direction outperforms the target's class-mean direction, not that it is optimal; and necessity is relative to the site's own information. P16–P20 were each designed after their predecessors' outcomes and frozen before evaluation; later additions (the depth sweep, the Procrustes map's untrained-source control, the noun-class control) are labelled, and unmet preregistered criteria (map conditioning, abstraction, strict increase over the full gain grid, the noun-class reading, gender's necessity and name criteria) are reported with the rest. The ranking studies' absolute preservation gaps are small, and their inference conditions on fixed fits, cohorts and calibration samples.

Outlook. The battery needs only a variable with known consumers, a neighbouring variable and activations on generic text, so it can be run on tense, case, entity attributes or truth, at a site chosen for each, and on any set of open models. Across many variables, it turns universality from a claim about similarity into a map of which causal variables models share, and where each lives.

Reproducibility statement

An evidence registry links every P1–P15 estimate to hashed saved results, and generated macros bind every P16–P20 value to its analysis record. Appendices A–E give designs, preregistration and code hashes, amendments, per-model results and deviations, and Appendices J and T the scopes of verification and replay. The supplement contains the claim verifier, figure and table producers, frozen study sources, receipts, saved measurements and replay code for P1–P6 and P8–P12, and the P16–P20 code, stimuli, records and analyses; to stay under 100 MB it omits the P13–P15 replay arrays and the P16, P17, P19 and P20 raw readouts, listing each with its size and hash. The complete release (about 4 GB; hash recorded) is available to reviewers on request and will be made public after review.

AI USE STATEMENT

The Research Team used generative AI assistance for language refinement and readability; implementation and refinement of code and software architecture; synthetic-dataset generation, preparation, formatting and augmentation; presentation of mathematical formalism; preparation of scientific figures and related research artifacts; and reference formatting. The empirical measurements reported here were produced by the specified model executions and saved-output analyses. Validation included numerical checks, separate implementations of selected calculations and reproducible saved-output replay; their scope and limitations are documented in this paper and its accompanying records. The Research Team retains responsibility for the experimental design, hypotheses, methods, analyses, citations and conclusions.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. *Sanity Checks for Saliency Maps*. NeurIPS, 2018. https://arxiv.org/abs/1810.03292.

Georgii Aparin and Tatiana Gaintseva. *A Geometric Account of Activation Steering through Angle–Norm Decomposition*. arXiv:2606.06735, version 2, 2026. https://arxiv.org/abs/2606.06735v2.

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*. NeurIPS, 2024. https://arxiv.org/abs/2406.11717.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. *Invariant Risk Minimization*. arXiv:1907.02893, 2019. https://arxiv.org/abs/1907.02893.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. *CausalGym: Benchmarking causal interpretability methods on linguistic tasks*. ACL, pp. 14638–14663, 2024. https://aclanthology.org/2024.acl-long.785/.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. *Revisiting Model Stitching to Compare Neural Representations*. NeurIPS, 2021. https://arxiv.org/abs/2106.07682.

Yuntai Bao, Xuhong Zhang, Jintao Chen, Ge Su, Yuxiang Cai, Hao Peng, Bing Sun, Haiqin Weng, Liu Yan, and Jianwei Yin. *Faithful Bi-Directional Model Steering via Distribution Matching and Distributed Interchange Interventions*. ICLR, 2026. Version 2, 15 March 2026. https://arxiv.org/abs/2602.05234v2.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. ACM Books, 2022. https://doi.org/10.1145/3501714.3501743.

Yonatan Belinkov. *Probing Classifiers: Promises, Shortcomings, and Advances*. Computational Linguistics, 48(1):207–219, 2022. https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. *LEACE: Perfect Linear Concept Erasure in Closed Form*. NeurIPS, 2023. https://arxiv.org/abs/2306.03819.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, et al. *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. ICML, 2023. https://arxiv.org/abs/2304.01373.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*. NeurIPS, 2016. https://arxiv.org/abs/1607.06520.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. *Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages*. NAACL, 2025. https://aclanthology.org/2025.naacl-long.312/.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. *Transferring Linear Features Across Language Models With Model Stitching*. NeurIPS, 2025. https://arxiv.org/abs/2506.06609.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. *A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations*. ICML, pp. 6243–6267, 2023. https://proceedings.mlr.press/v202/chughtai23a.html.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. *Sparse Autoencoders Find Highly Interpretable Features in Language Models*. ICLR, 2024. https://arxiv.org/abs/2309.08600.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. *Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals*. Transactions of the Association for Computational Linguistics, 9:160–175, 2021. https://aclanthology.org/2021.tacl-1.10/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, et al. *Toy Models of Superposition*. Transformer Circuits Thread, 2022. https://arxiv.org/abs/2209.10652.

Allyson Ettinger. *What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models*. Transactions of the Association for Computational Linguistics, 8:34–48, 2020. https://aclanthology.org/2020.tacl-1.3/.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. *Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models*. ACL-IJCNLP, pp. 1828–1843, 2021. https://aclanthology.org/2021.acl-long.144/.

Zhiwei Gao, Shaowen Peng, Shoko Wakamiya, and Eiji Aramaki. *What Does Activation Steering Control? Attribution Across Answer Encodings and Output-Sensitive Subspaces*. arXiv:2608.22985, version 1, 24 August 2026. Preprint. https://arxiv.org/abs/2608.22985v1.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. *Causal Abstractions of Neural Networks*. NeurIPS, 2021. https://arxiv.org/abs/2106.02997.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. *Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations*. Proceedings of Machine Learning Research, 236:160–187, 2024. https://proceedings.mlr.press/v236/geiger24a.html.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. *Universal Neurons in GPT2 Language Models*. Transactions on Machine Learning Research, 2024. https://arxiv.org/abs/2401.12181.

Sophie Hao and Tal Linzen. *Verb Conjugation in Transformers Is Determined by Linear Encodings of Subject Number*. Findings of EMNLP, pp. 4531–4539, 2023. https://aclanthology.org/2023.findings-emnlp.300/.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. *Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models*. NeurIPS, 2023. https://arxiv.org/abs/2301.04213.

John Hewitt and Percy Liang. *Designing and Interpreting Probes with Control Tasks*. EMNLP-IJCNLP, pp. 2733–2743, 2019. https://aclanthology.org/D19-1275/.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. *RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations*. ACL, 2024. https://arxiv.org/abs/2402.17700.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. *The Platonic Representation Hypothesis*. arXiv:2405.07987, 2024. https://arxiv.org/abs/2405.07987.

Shawn Im and Sharon Li. *A Unified Understanding and Evaluation of Steering Methods*. arXiv:2502.02716, 2025; version 2, 9 January 2026. https://arxiv.org/abs/2502.02716v2.

Nora Kassner and Hinrich Schütze. *Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly*. ACL, pp. 7811–7818, 2020. https://aclanthology.org/2020.acl-main.698/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. *Similarity of Neural Network Representations Revisited*. ICML, pp. 3519–3529, 2019. https://proceedings.mlr.press/v97/kornblith19a.html.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. *The Emergence of Number and Syntax Units in LSTM Language Models*. NAACL-HLT, pp. 11–20, 2019. https://aclanthology.org/N19-1002/.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. *Quantifying Feature Space Universality Across Large Language Models via Sparse Autoencoders*. arXiv:2410.06981, 2024. https://arxiv.org/abs/2410.06981.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. *Probing for the Usage of Grammatical Number*. ACL, pp. 8818–8831, 2022. https://aclanthology.org/2022.acl-long.603/.

Karel Lenc and Andrea Vedaldi. *Understanding Image Representations by Measuring Their Equivariance and Equivalence*. CVPR, pp. 991–999, 2015. https://arxiv.org/abs/1411.5908.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. *Convergent Learning: Do Different Neural Networks Learn the Same Representations?* ICLR, 2016. https://arxiv.org/abs/1511.07543.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. *Inference-Time Intervention: Eliciting Truthful Answers from a Language Model*. NeurIPS, 2023. https://arxiv.org/abs/2306.03341.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. *Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies*. Transactions of the Association for Computational Linguistics, 4:521–535, 2016. https://aclanthology.org/Q16-1037/.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. *Latent Space Translation via Semantic Alignment*. NeurIPS, 2023. https://arxiv.org/abs/2311.00664.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. *Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching*. ICLR, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. *Locating and Editing Factual Associations in GPT*. NeurIPS, 2022. https://arxiv.org/abs/2202.05262.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. *Linguistic Regularities in Continuous Space Word Representations*. NAACL-HLT, pp. 746–751, 2013. https://aclanthology.org/N13-1090/.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. *Relative Representations Enable Zero-Shot Latent Space Communication*. ICLR, 2023. https://arxiv.org/abs/2209.15430.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, et al. *MIB: A Mechanistic Interpretability Benchmark*. ICML, 2025. https://arxiv.org/abs/2504.13151.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. *Zoom In: An Introduction to Circuits*. Distill, 2020. https://doi.org/10.23915/distill.00024.001.

Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. *Activation Space Interventions Can Be Transferred Between Large Language Models*. ICML, 2025. https://arxiv.org/abs/2503.04429.

Kiho Park, Yo Joong Choe, and Victor Veitch. *The Linear Representation Hypothesis and the Geometry of Large Language Models*. ICML, 2024. https://arxiv.org/abs/2311.03658.

Judea Pearl. *Direct and Indirect Effects*. UAI, pp. 411–420, 2001. https://arxiv.org/abs/1301.2300.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. *Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals*. Journal of the Royal Statistical Society Series B, 78(5):947–1012, 2016. https://doi.org/10.1111/rssb.12167.

Dario Picozzi. *The information geometry of large language models is shared, learned, and controllable*. arXiv:2609.11063, version 1, 10 September 2026. Preprint. https://arxiv.org/abs/2609.11063v1.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. *Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction*. CoNLL, pp. 194–209, 2021. https://aclanthology.org/2021.conll-1.15/.

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

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. *Massive Activations in Large Language Models*. COLM, 2024. https://arxiv.org/abs/2402.17762.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. *The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability?* NeurIPS, 2025. https://arxiv.org/abs/2507.08802.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. *LLM Circuit Analyses Are Consistent Across Training and Scale*. NeurIPS, 2024. https://arxiv.org/abs/2407.10827.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. *Steering Language Models With Activation Engineering*. arXiv:2308.10248, 2023. https://arxiv.org/abs/2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. *Investigating Gender Bias in Language Models Using Causal Mediation Analysis*. NeurIPS, 2020. https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. *Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small*. ICLR, 2023. https://arxiv.org/abs/2211.00593.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. *FishBack: Pullback Fisher Geometry for Optimal Activation Steering in Transformers*. arXiv:2605.17231, version 2, 18 August 2026. Preprint. https://arxiv.org/abs/2605.17231v2.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

Fred Zhang and Neel Nanda. *Towards Best Practices of Activation Patching in Language Models: Metrics and Methods*. ICLR, 2024. https://arxiv.org/abs/2309.16042.

Peiyan Zhang and Jason Xin. *A Negative Result on Cross-Model Activation Transfer in a Pythia Multi-Hop Setting*. arXiv:2606.03280, 2026. https://arxiv.org/abs/2606.03280.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv:2310.01405, 2023. https://arxiv.org/abs/2310.01405.

A INVARIANT CAUSAL VARIABLES (P16): DESIGN, RECORDS AND PER-MODEL RESULTS

Freeze and scope. The question, stimuli, splits, operator, layer rule, directions, transport map, endpoints and thresholds were frozen in a hashed preregistration before any model was run on the stimuli (preregistration SHA-256 `674c8e49...`, stimuli `1ad19254...`). The pilot used development families only; the code was frozen before any evaluation-family measurement (run code `c5e1b425...`, orchestrator `a34e26fa...`, analysis `7db3c344...`). Four engineering changes were made during the pilot: the pooled switch was signed toward the donor's number, as the preregistration specifies (the first pilot run pooled unsigned changes, whose denominator cancels; that output is preserved); model resolution reads both download manifests; transport evaluates gender items only for the two directions whose gender effect is analysed; and a deprecated loading argument was removed. No stimulus, split, criterion or analysis choice changed. P16 was designed after the outcomes of P1–P15 were known. Its verification standard is lighter than that of the earlier studies: every reported value is generated from the saved per-item readouts by one analysis script and bound to its record, and a separate implementation recomputes every primary endpoint from per-family sums with identical results. A replay in a fresh process with a different batch size recomputed 15 native and transported readout sets for 3 target models (GPT-2 355M, Pythia 70M, Qwen3 0.6B): individual readouts differ by at most 0.0074 nats through floating-point order, and the pooled held-out switches by at most 1.2×10^{-5}. A first replay used a readout-level tolerance of 0.001 nats, which these floating-point differences exceed; the replay was then assessed on the switches the analysis uses (tolerance 0.005). No independent reimplementation of the GPU execution was run.

Stimuli. Forty-six gendered human-noun families (masculine and feminine, singular and plural; nine include irregular or *-men* plurals) were split by a seeded shuffle into 14 fitting, 8 development and 24 evaluation families. Each consumer has three templates in which the readout follows the subject by at least three tokens; for example *The {N}, as everyone knows,* (present copula), *After a while, the {N} started to enjoy* (reflexive) and *When the {N} came home late,* (subject pronoun). Number items pair recipient and donor of the same family, gender and template with opposite number (1,728 per model); gender items pair singular recipient and donor of opposite gender for the reflexive, pronoun and possessive consumers (432 per model). Readout forms carry a leading space; all were single tokens in every tokenizer.

Site, directions and maps. The site is the final subject token at the residual output of the deepest block whose full patch switches the present copula by at least 0.8 on development families; this depends only on where the subject position carries number. DM and GDM are class-mean differences over fitting families; FIT is fitted to the present copula alone (two-way cross-entropy toward the donor's number, 300 Adam steps, learning rate 0.01, initialized at DM); the probe is an L2 logistic regression. Maps are affine ridge regressions between centred activations at the final token of 30,000 word-boundary WikiText-103 prefixes (identical strings across models), 27,000 for fitting and 3,000 for choosing the penalty from six values by held-out R^2. Models run in FP32 up to 4B parameters and in BF16 at 6.9B–8B.

Measures. A switch is a pooled ratio of means over items: the sum of signed patched changes divided by the sum of signed clean donor-minus-recipient differences. Normalized switches divide by the full patch at the same site. A model-consumer cell is eligible when clean number accuracy on evaluation items is at least 0.8. Transport efficiency is $(s_{\text{transported}} - s_{\text{random}})/(s_{\text{DM}} - s_{\text{random}})$ in the target, pooled over its eligible held-out consumers, where s_{random} averages ten native random directions. Intervals are 95% percentile bootstraps over the 24 evaluation families with 10,000 draws; the same draws are used for numerator and denominator.

Secondary and exploratory analyses. The preregistration lists as secondary the gender direction's transport specificity, FIT and probe transport, per-consumer results and tense mass; gender treated as a second variable, site-depth profiles, the Pythia-410M training checkpoints, irregular-plural families and the relation of transport efficiency to map fit, direction cosine and model size are exploratory. All are computed by a separate script written after the code freeze, from the same saved readouts; the frozen primary analysis is unchanged. Gender transport efficiency uses the same

form as number, with the target's gender direction as reference and its native random directions as baseline, pooled over the three gendered consumers.

Execution deviations. Checkpoint downloads were restarted several times after stalled transfers; every checkpoint is pinned by commit. After the native stage, transport was run by a separate launcher with batch size 256 instead of 96, which changes the order of batched floating-point operations but no computation, and with the Pythia-410M training checkpoints as sources only, which is all the analysis uses; the orchestrator was stopped before its own transport stage wrote any record. All 25 planned models and checkpoints completed the native stage; every trained model completed transport; no trained model was ineligible.

Table 3: P16 per-model results on evaluation families. Site: chosen block / blocks. Held-out: eligible held-out consumers (clean accuracy at least 0.8). DM, FIT and probe: switch on eligible held-out consumers normalized by the full patch (DM with 95% family-bootstrap interval). Gender: DM's gender switch relative to the full patch. GDM-num: gender direction's number switch relative to DM. E1: consumer generality supported. E2: double dissociation ("iso.": isolation only). Checkpoint rows are exploratory.

Model	Site	Held-out	DM	FIT	Probe	Gender	GDM-num	E1	E2
Pythia 70M	2/6	BCDEF	0.86 [0.81, 0.91]	1.09	0.88	–0.00	0.03	yes	yes
Pythia 160M	5/12	BCDEF	0.92 [0.89, 0.95]	1.08	0.89	–0.00	0.03	yes	yes
Pythia 410M	5/24	BCDEF	0.81 [0.73, 0.89]	1.02	0.91	0.01	0.01	yes	yes
Pythia 1B	3/16	BCDEF	0.82 [0.73, 0.89]	1.00	0.92	0.00	0.01	yes	yes
Pythia 1.4B	5/24	BCDEF	0.84 [0.78, 0.90]	1.03	0.92	0.00	0.01	yes	yes
Pythia 2.8B	8/32	BCDEF	0.89 [0.85, 0.92]	1.06	0.92	0.00	0.00	yes	yes
Pythia 6.9B	6/32	BCDEF	0.79 [0.73, 0.86]	1.03	0.88	0.00	0.00	yes	yes
Qwen3 0.6B	7/28	BCDEF	0.96 [0.94, 0.99]	1.07	0.94	0.00	0.00	yes	yes
Qwen3 1.7B	5/28	BCDEF	0.97 [0.95, 0.98]	1.01	0.97	0.00	0.00	yes	yes
Qwen3 4B	7/36	BCDEF	0.92 [0.89, 0.95]	1.07	0.93	0.00	0.00	yes	yes
Qwen3 8B	7/36	BCDEF	0.93 [0.91, 0.95]	1.08	0.93	0.00	0.00	yes	yes
SmolLM2 135M	13/30	BCDEF	0.98 [0.96, 0.99]	1.15	0.98	0.00	0.00	yes	yes
SmolLM2 360M	14/32	BCDEF	0.97 [0.94, 0.99]	1.11	0.98	0.00	0.00	yes	yes
SmolLM2 1.7B	5/24	BCDEF	0.94 [0.92, 0.96]	1.03	0.96	–0.00	0.00	yes	yes
GPT-2 124M	4/12	BCDEF	0.84 [0.78, 0.90]	1.12	0.89	–0.00	0.01	yes	yes
GPT-2 355M	4/24	BCDEF	0.70 [0.60, 0.79]	1.01	0.93	0.02	0.01	yes	yes
GPT-2 774M	11/36	BCDEF	0.81 [0.75, 0.87]	1.08	0.85	0.00	0.01	yes	yes
GPT-2 1.5B	13/48	BCDEF	0.80 [0.73, 0.87]	1.11	0.85	0.00	0.01	yes	yes
OLMo-2 1B	2/16	BCDEF	0.83 [0.77, 0.88]	0.98	0.89	0.00	0.00	yes	yes
OLMo-2 7B	6/32	BCDEF	0.75 [0.70, 0.79]	1.02	0.80	–0.00	0.00	yes	yes
410M step 0	1/24	none	–	–	–	–	–	n/a	n/a
410M step 512	13/24	none	–	–	–	–	–	n/a	n/a
410M step 2k	10/24	BCDEF	0.88 [0.81, 0.94]	1.10	0.80	0.01	0.01	yes	yes
410M step 8k	10/24	BCDEF	0.92 [0.89, 0.95]	1.07	0.90	0.01	0.02	yes	yes
410M step 32k	10/24	BCDEF	0.93 [0.90, 0.96]	1.07	0.93	0.00	0.01	yes	yes

Detailed displays. Figures 4–6 and Table 4 give the per-model and per-family P16 results summarized in Sections 5–7.

Table 4: P16 by target family. E1: consumer generality supported. E2: double dissociation. DM: median held-out switch relative to the full patch. Transport: median efficiency of number directions from other models of the same family and of other families into this family's models, with the number of ordered pairs meeting the preregistered criterion. Gender: median transport efficiency of gender directions for gender (secondary).

Family	Models	E1	E2	DM	Transport, within	Transport, across	Gender
Pythia	7	7/7	7/7	0.84	1.02 (36/42)	1.06 (82/91)	0.73
Qwen3	4	4/4	4/4	0.95	0.97 (12/12)	0.92 (64/64)	0.52
SmolLM2	3	3/3	3/3	0.97	0.95 (4/6)	0.90 (37/51)	0.66
GPT-2	4	4/4	4/4	0.81	0.92 (12/12)	1.00 (64/64)	0.69
OLMo-2	2	2/2	2/2	0.79	0.83 (2/2)	0.93 (36/36)	0.27

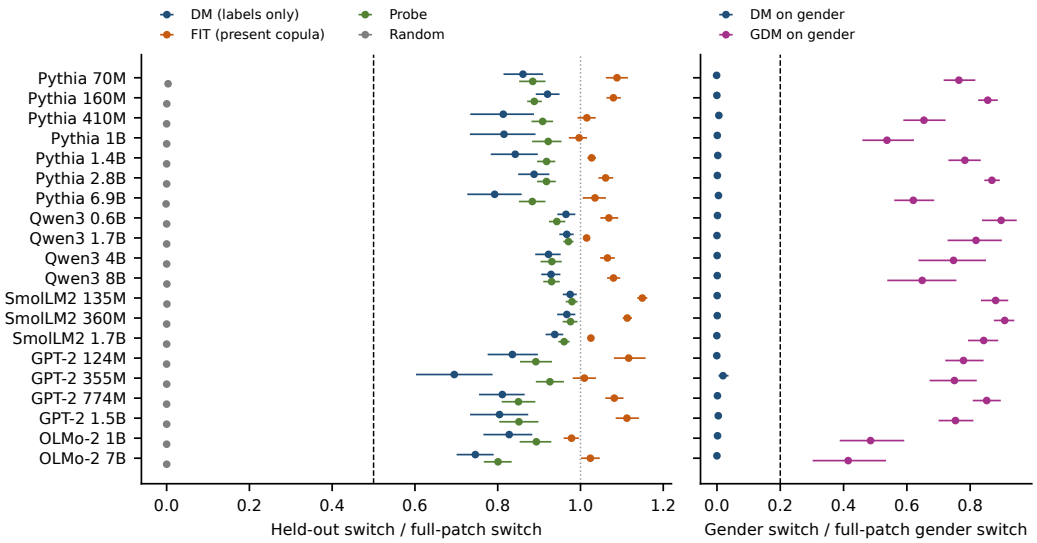


Figure 4: Consumer generality and isolation (P16, evaluation families). Left: switch on the five held-out consumers normalized by the full patch at the same site, for the label-derived direction (DM), the direction fitted to the present copula only (FIT), the probe and random directions; bars are 95% family-bootstrap intervals and the dashed line is the preregistered 0.5 criterion. Right: effect on gender relative to the full patch for DM, and of the gender direction on gender. Models are grouped by family and ordered by size.

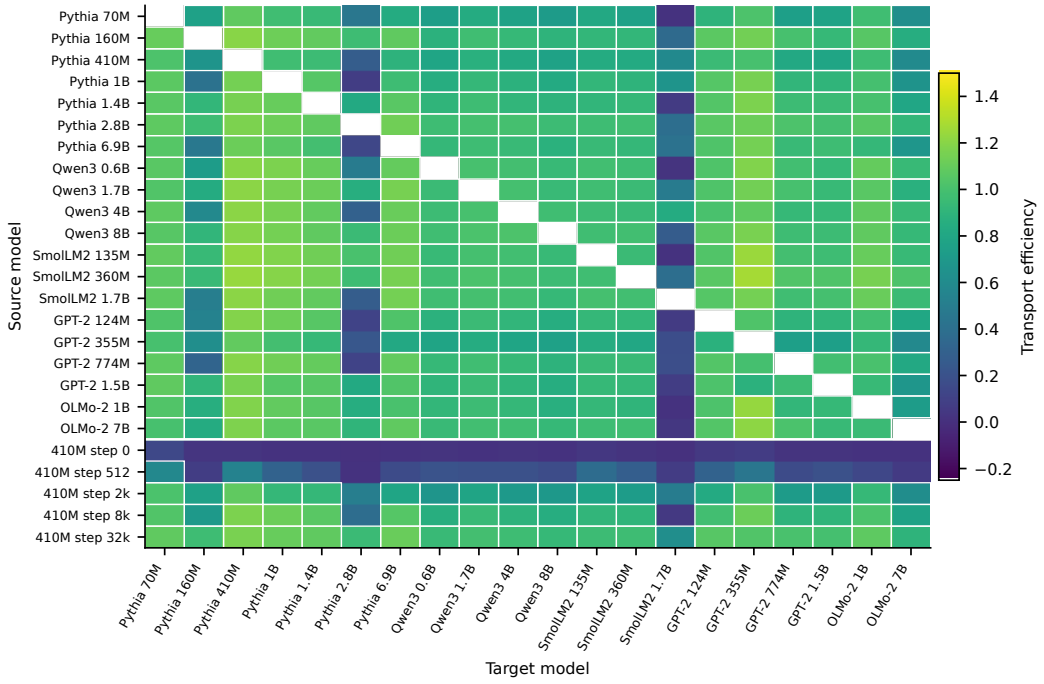


Figure 5: Cross-model invariance (P16). Transport efficiency of each source model's number direction (rows) in each target model (columns), pooled over the target's eligible held-out consumers; 1 equals the target's own direction and 0 equals random directions. Maps are affine and fitted only on generic text. Outlined cells meet the preregistered criterion (lower bound ≥ 0.5). The bottom rows are Pythia-410M training checkpoints used as sources, including random initialization.

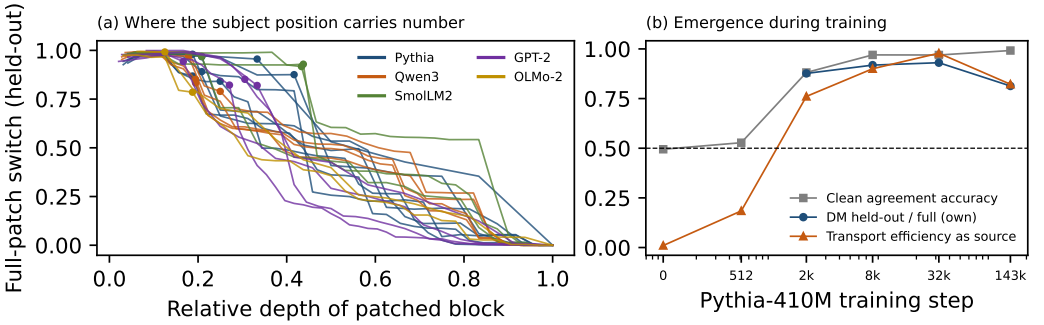


Figure 6: (a) Full-patch switch on held-out consumers at the final subject token as a function of the patched block's relative depth (development families); dots mark each model's preregistered site. (b) Pythia-410M across training (exploratory): clean agreement accuracy, DM's own held-out switch relative to the full patch where eligible, and the median transport efficiency of the checkpoint's number direction into the trained models; the dashed line marks 0.5.

B COUNTERFACTUAL TESTS OF THE NUMBER VARIABLE (P17): DESIGN, RECORDS AND RESULTS

Freeze and scope. P17 was designed after the P16 outcomes were known and after a diagnostic on generic text (no stimulus data) showed that the weakest P16 transport targets are the models whose site activations have a massive dimension. Its preregistration (SHA-256 `0f5ae3a8...`) and determiner-cued stimuli (`ab9a2b09...`) were frozen before any P17 measurement. A development pilot (P16 development families; development nouns *cod, pike, offspring, caribou*) preceded an amendment (`e6dad1fc...`) that added the depth sweep (P17d), a massive-dimension diagnostic, a built-in replay of P16 readouts and the saving of every direction; the run code (`b3a8ab5a...`) and launcher (`c1796c0f...`) were frozen before any evaluation-split measurement. P17 reuses P16's models, sites, directions, generic-text activations and evaluation families without change and does not alter any P16 result. On the evaluation split every run recomputed P16's clean, DM, random and full-patch readouts for its model (20 models). In the models run in FP32, individual readouts differ from P16's saved values by at most 0.0085 nats; in the three BF16 models, whose arithmetic is sensitive to the order of batched operations, by up to 1.57 nats for single items, and DM's normalized held-out switch agrees with P16's to within 0.0014 in every model.

Maps (P17a). All maps are fitted on the same 27,000 generic prefixes as P16, with the same 3,000 held out. *Standardized ridge*: every source and target dimension is centred and divided by its training standard deviation; the penalty (six values, scaled by the mean squared singular value) is chosen by held-out R^2 averaged over standardized target dimensions. *Standardized Procrustes*: the partial isometry $UV^\top$ from the singular value decomposition of the standardized cross-covariance. *Raw Procrustes*: the same between centred raw activations. *Shuffled-pair null*: standardized ridge fitted after permuting the target rows (seed 20260926), with its penalty chosen on held-out rows permuted the same way. Directions map as displacements: $w_T = \mathrm{unit}(\sigma_T \odot ((w_S/\sigma_S)A))$ for standardized maps and $\mathrm{unit}(w_S A)$ for raw maps. The preregistration transports the untrained Pythia-410M source by standardized ridge only; a post hoc control (below) adds the Procrustes maps. Because ridge regression is equivariant to rotations of the target space, standardizing the target changes the transported direction only through the penalty choice; standardizing the source changes the effective penalty on every source direction.

Determiner-cued items (P17b). Twenty-four animal nouns with identical singular and plural forms were split by a seeded shuffle into fitting (6: *salmon, elk, reindeer, swine, carp, mackerel*), development (4) and evaluation (14) nouns. Each appears in the P16 templates with *this/these* or *that/those* in place of *the*; consumers read *is/are*, *was/were*, *has/have*, *itself/themselves*, *it/they* and *its/their*. Recipient and donor share template, noun and determiner pair and differ only in the determiner's number (1,008 evaluation items per model). The noun token is identical in recipient and donor, so any number the site carries was computed from context. DM-det is the plural-minus-singular mean over fitting nouns at the P16 site.

Graded interventions and necessity (P17c). The coordinate along DM, the full state and one random direction are set to $h_r + g(h_d - h_r)$ for $g \in \{-0.5, 0.25, 0.5, 0.75, 1, 1.5, 2\}$ on the 1,728 P16 evaluation items. For necessity, the coordinate of every evaluation text along DM is set to the midpoint of the fitting-family class means of its projection (GDM: midpoint of the singular masculine and feminine means; random: mean projection). The necessity index is one minus the ratio of pooled signed log-odds after and before ablation over eligible held-out consumers; the reference is the same index after setting the full state halfway ($g = 0.5$) toward the opposite-number donor. The rank-1 energy share uses the eligible consumers' switches at the seven gains.

Depth sweep (P17d). At twelve evenly spaced blocks (and the P16 site), DM_L and $\mathrm{DM\text{-}det}_L$ are re-estimated from the fitting families and fitting nouns at block L, and the full patch, DM_L and $\mathrm{DM\text{-}det}_L$ are applied to both item sets. A model shares a direction between the cues at block L if both full patches switch at least 0.25 and both cross-cue normalized switches have lower bound at least 0.5. The final block is on the grid for completeness; patching a non-final position there cannot affect the readout.

Endpoints and outcomes. E17.1 (shuffled-pair null below 0.25 in at least 95% of pairs): supported (380 of 380). E17.2 (standardized ridge yields at least as many invariant pairs as the P16 map and rescues at least half of its failures): not supported (354 against 349; 7 of 31 rescued, 2 lost). E17.3 (descriptive): 312 and 326 invariant pairs under standardized and raw Procrustes; raw Procrustes meets the criterion in 27 of the 31 pairs that fail under the P16 map and fails in 50 that pass under it (median efficiency 0.83, against 0.95 for standardized ridge). E17.4 (untrained source below 0.25 in every target): supported (largest upper bound 0.14). E17.5 (abstraction): not supported (0 of 14 eligible models). E17.6 (strictly increasing dose): not supported (15 of 20; in Qwen3 1.7B, SmolLM2 135M, SmolLM2 360M, SmolLM2 1.7B, OLMo-2 7B the switch rises strictly up to $g = 1.5$ and dips at $g = 2$). E17.7 (rank-1 share at least 0.95): supported (20 of 20; minimum 0.998). E17.8 (necessity ratio lower bound at least 0.5): supported (16 of 20; median ratio 0.73); in the others (Pythia 2.8B, Pythia 6.9B, OLMo-2 1B, OLMo-2 7B) ablating DM removes 0.40–0.56 of what the halfway full state removes, so other directions at their sites carry part of the number signal. E17.9 (descriptive): 0 of 20 models share a direction at some depth.

Post hoc control (amendment 02). After the P17a results of the first six evaluation targets had been seen, and before any of its data existed, a second amendment (`6c705827...`; code `9b033fb4...`) added one control: the untrained source's DM transported into every target by raw Procrustes, standardized Procrustes and standardized ridge, with native DM and ten random directions re-measured in the same run. Over 20 targets its efficiency has median 0.00 under raw Procrustes (largest upper bound 0.07), 0.00 under standardized Procrustes (largest 0.05) and 0.01 under standardized ridge (largest 0.14); the Procrustes maps therefore pass the untrained-source control of E17.4.

Cue versus noun class (P17e, amendment 03). DM is estimated from human nouns and DM-det from animal nouns, so their dissociation could reflect the noun class rather than the cue. A third amendment (`fe21652f...`), declared after a development pilot on one model and before any evaluation data existed, holds the class fixed: 24 animal nouns whose number is marked on the noun (*dog/dogs*, *mouse/mice*), split into 6 fitting, 4 development and 14 evaluation nouns, in the P17b templates with *the*, at the P16 site. Over 20 models, the full patch switches these items by a median 0.89; relative to it, (i) DM switches a median 0.59 (lower bound at least 0.5 in 10 models), (ii) DM-det switches 0.01 (upper bound below 0.25 in 18), and DM-animal, estimated from the animal fitting nouns, switches 0.66. DM-animal's median cosine is 0.57 with DM and 0.12 with DM-det, and on determiner-cued items it switches a median 0.11 of their full patch. The declared cue reading requires (i) and (ii) in 80% of models and a higher cosine with DM: not supported (both in 8 models). Condition (ii) and the cosine ordering hold broadly; (i) fails because DM's transfer from human to animal nouns depends on the family: a median 0.95 in Qwen3 and SmolLM2 (range 0.76–0.97) but 0.47 in Pythia, GPT-2 and OLMo-2 (0.28–0.84), where number on animal nouns is carried partly along other directions.

Execution. All 20 targets ran with batch size 256; small and mid-sized targets ran two at a time on one GPU, which changes only the order of floating-point operations. 20 completed on the first attempt. The post hoc control and P17e ran afterwards, one target at a time on the same GPU; all 20 targets completed in each.

Table 5: P17 per-model results (evaluation split). Invariant sources: number of the 19 other trained models whose transported number direction meets the invariance criterion under P16's ridge map / standardized ridge / raw Procrustes / shuffled null. Dose: DM switch at unit gain. Rank-1: energy share of the first singular value of the consumers-by-gain switch matrix. Necessity: one minus the pooled signed log-odds after mean-ablating DM (95% family interval), GDM, and after setting the full state halfway to the donor. Determiner: switch of determiner-cued number by the full patch at the P16 site, and DM and DM-det normalized by it. Max cos: largest cosine between DM and DM-det over depths.

	Invariant			Necessity			Determiner-cued			
Model	sources	Dose	Rank-1	DM	GDM	Full, half	Full	DM	DM-det	Max cos
Pythia 70M	19/19/19/0	0.84	0.999	0.80 [0.73, 0.88]	–0.00	0.97	0.19	0.13	0.77	0.17
Pythia 160M	14/17/19/0	0.80	0.999	0.67 [0.62, 0.72]	0.02	0.81	0.38	0.25	0.79	0.29
Pythia 410M	19/19/19/0	0.72	1.000	0.61 [0.51, 0.72]	0.01	0.81	0.10	0.06	0.54	0.45
Pythia 1B	19/19/18/0	0.80	1.000	0.65 [0.53, 0.79]	0.01	0.90	0.09	0.05	0.98	0.42
Pythia 1.4B	19/19/18/0	0.82	1.000	0.60 [0.49, 0.70]	0.01	0.85	0.23	0.13	0.84	0.38
Pythia 2.8B	9/9/17/0	0.77	1.000	0.40 [0.32, 0.48]	0.02	0.73	0.39	0.17	0.81	0.41
Pythia 6.9B	19/19/17/0	0.70	1.000	0.41 [0.32, 0.50]	0.01	0.74	0.35	0.07	0.82	0.36
Qwen3 0.6B	19/19/16/0	0.77	1.000	0.55 [0.48, 0.63]	0.01	0.72	0.26	0.38	0.88	0.57
Qwen3 1.7B	19/19/9/0	0.94	1.000	0.54 [0.46, 0.62]	0.02	0.85	0.40	0.18	0.81	0.47
Qwen3 4B	19/19/8/0	0.80	1.000	0.46 [0.41, 0.51]	0.02	0.66	0.30	0.18	0.90	0.48
Qwen3 8B	19/19/8/0	0.80	1.000	0.40 [0.34, 0.47]	0.02	0.65	0.34	0.13	0.90	0.44
SmolLM2 135M	19/19/19/0	0.88	1.000	0.75 [0.71, 0.79]	–0.00	0.83	0.43	0.35	0.76	0.50
SmolLM2 360M	19/19/18/0	0.90	0.999	0.71 [0.67, 0.76]	0.01	0.79	0.63	0.34	0.91	0.57
SmolLM2 1.7B	3/5/15/0	0.91	0.999	0.46 [0.40, 0.53]	0.01	0.78	0.33	0.16	0.82	0.42
GPT-2 124M	19/19/19/0	0.70	0.999	0.45 [0.38, 0.52]	–0.02	0.61	0.24	0.29	0.57	0.40
GPT-2 355M	19/19/19/0	0.67	1.000	0.86 [0.70, 1.03]	0.00	0.74	0.27	0.00	0.86	0.45
GPT-2 774M	19/19/19/0	0.69	1.000	0.44 [0.35, 0.54]	0.01	0.58	0.35	0.13	0.82	0.42
GPT-2 1.5B	19/19/18/0	0.68	1.000	0.52 [0.42, 0.63]	0.00	0.59	0.40	0.05	0.71	0.48
OLMo-2 1B	19/19/19/0	0.82	1.000	0.37 [0.30, 0.44]	0.01	0.92	0.12	0.11	0.82	0.46
OLMo-2 7B	19/19/12/0	0.60	0.998	0.27 [0.22, 0.32]	0.01	0.54	0.35	0.06	0.87	0.32

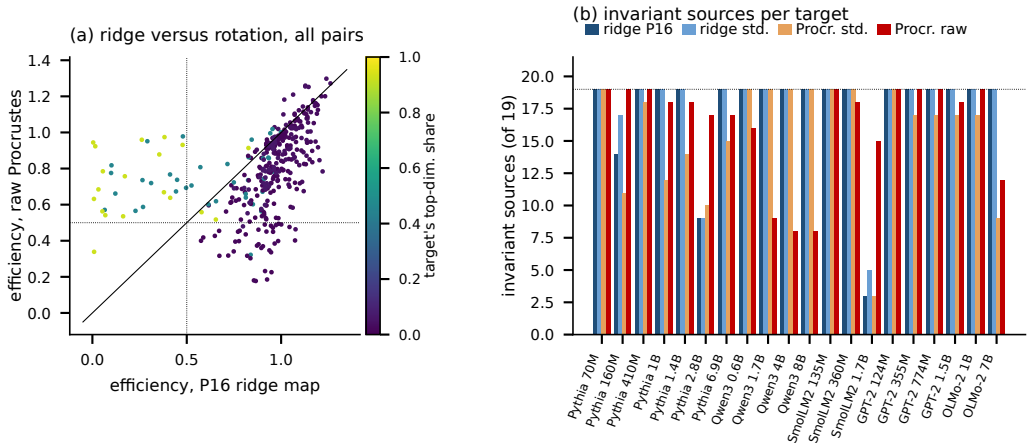


Figure 7: P17a map controls. (a) Transport efficiency of every ordered pair under P16's ridge map and under the raw Procrustes rotation, coloured by the target's largest-dimension variance share on generic text. (b) Number of the 19 other trained models whose transported direction meets the invariance criterion in each target, under each of the four maps.

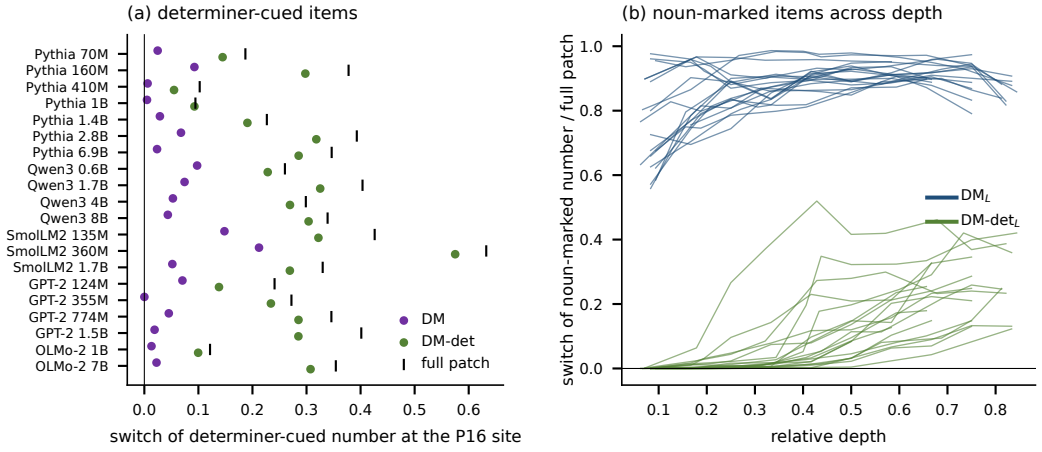


Figure 8: P17b and P17d. (a) Switch of determiner-cued number at the P16 site by DM, DM-det and the full patch, per model. (b) The converse of Figure 9b: on noun-marked items, the switch of DM$_L$ and of DM-det$_L$ relative to the full patch at each depth where that patch switches at least 0.1; one line per model.

C THE NUMBER VARIABLE IN NATURAL SENTENCE CONTEXTS (P18)

Freeze and scope. P18 was designed on 26 September 2026, after all P16 and P17 outcomes were known. A development pilot on three models (Pythia-160M, Qwen3-0.6B and GPT-2; development families and 50 development frames) changed the frame rules twice (verbs read at least one word after the noun, with other rules relaxed to keep enough frames; no verb directly after a relative pronoun) and set the eligibility rule; the preregistration (SHA-256 `1273700e...`), code (`f52f7cf6...`), item extractor (`e49e7ad5...`) and evaluation item file (`0b05772d...`) were then frozen before any evaluation measurement. P18 reuses the P16 sites, DM, FIT, probe, GDM and random directions and the P17 transported directions without change: nothing is fitted on these items.

Items. Frames come from the WikiText-103 validation and test text from which the P16 map prefixes were drawn (without labels). A frame is a sentence prefix containing a noun phrase *the X* (or *his/her/their/its X*; not inside a prepositional phrase) whose head is followed, after at least one and at most 20 words, by an agreeing *is/are*, *was/were* or *has/have*, with no intervening pronoun, other auxiliary or family noun, not directly after a relative pronoun, and with either no comma or one comma pair closing just before the verb; the prefix ends just before the verb. The first rule excludes verbs directly after the noun, for which the full patch at the noun is complete by construction. The head is replaced by an evaluation-family noun (recipient) and by its other-number form (donor). Of the frames the corpus yields, 50 served the pilot and 266 the evaluation, assigned to the 24 evaluation families in rotation with alternating gender; each frame gives a singular and a plural recipient (532 items; median 4 words between noun and verb, 468 items with at least three). The frames are selected by pattern rules rather than a parser, so in some items the verb may not be controlled by the substituted noun; such items lower the full patch and DM alike and leave their ratio unbiased.

Endpoints and outcomes. A model enters when its clean accuracy over recipient and donor texts is at least 0.75 and its full patch switches at least 0.25 (P16 required 0.8 per consumer cell; natural contexts rest on fewer items per cell); 20 of 20 models entered (clean accuracy median 0.88, minimum 0.84; full patch median 0.78, range 0.67–0.96). E18.1 (DM normalized switch lower bound at least 0.5 and every random direction below 0.1, in 80% of entering models): supported (20 of 20; median 0.86, smallest lower bound 0.56). E18.2 (P16-map transport lower bound at least 0.5 in 80% of ordered pairs, untrained source upper bound below 0.25 in every target, shuffled-pair null below 0.25 in 95% of pairs): supported (353 of 380 pairs, median efficiency 0.99; untrained source median 0.01, below 0.25 in 20 targets; shuffled null median 0.00, below 0.25 in 380 pairs). Descriptive: FIT and the probe mediate medians of 1.04 and 0.92 of the full patch; standardized ridge and raw Procrustes transport with medians 0.99 and 0.86 (357 and 333 pairs with lower bound at least 0.5); across models, DM's normalized switch in natural contexts and in the templates (consumers A–C, median 0.86) have Spearman correlation 0.96.

Table 6: P18 natural sentence contexts, per model. Acc.: clean accuracy over recipient and donor texts. Full: full-patch switch. DM/full: DM's normalized switch with 95% family-bootstrap interval; gap ≥ 3: the same for items with at least three words between noun and verb. Transport: median efficiency of the other entering models' DM through the P16 ridge map, with the number of sources whose lower bound is at least 0.5. Untrained: upper bound for the untrained Pythia-410M source.

Model	Acc.	Full	DM/full	Gap 3+	Transport	Untrained
Pythia 70M	0.84	0.89	0.90 [0.84, 0.95]	0.90	1.08 (19/19)	0.16
Pythia 160M	0.88	0.84	0.96 [0.93, 0.99]	0.96	0.85 (16/19)	0.04
Pythia 410M	0.88	0.75	0.84 [0.76, 0.92]	0.84	1.15 (19/19)	0.05
Pythia 1B	0.88	0.96	0.79 [0.70, 0.88]	0.79	1.13 (19/19)	0.01
Pythia 1.4B	0.87	0.85	0.83 [0.76, 0.89]	0.83	1.10 (19/19)	0.01
Pythia 2.8B	0.89	0.85	0.88 [0.84, 0.92]	0.88	0.61 (10/19)	0.00
Pythia 6.9B	0.88	0.77	0.77 [0.71, 0.84]	0.79	1.15 (19/19)	0.01
Qwen3 0.6B	0.86	0.75	0.98 [0.95, 1.00]	0.97	0.93 (19/19)	0.03
Qwen3 1.7B	0.86	0.79	0.94 [0.91, 0.97]	0.95	0.98 (19/19)	0.01
Qwen3 4B	0.87	0.68	0.93 [0.91, 0.95]	0.93	0.96 (19/19)	0.02
Qwen3 8B	0.86	0.67	0.90 [0.88, 0.93]	0.90	0.94 (19/19)	0.02
SmolLM2 135M	0.87	0.90	0.97 [0.96, 0.98]	0.97	0.96 (19/19)	0.06
SmolLM2 360M	0.88	0.77	0.96 [0.94, 0.98]	0.96	0.92 (19/19)	0.03
SmolLM2 1.7B	0.89	0.87	0.89 [0.87, 0.92]	0.90	0.23 (4/19)	0.00
GPT-2 124M	0.89	0.69	0.84 [0.77, 0.90]	0.83	1.04 (19/19)	0.10
GPT-2 355M	0.89	0.95	0.67 [0.56, 0.77]	0.66	1.16 (19/19)	0.11
GPT-2 774M	0.88	0.74	0.82 [0.77, 0.88]	0.83	0.99 (19/19)	0.02
GPT-2 1.5B	0.88	0.70	0.81 [0.72, 0.89]	0.80	0.99 (19/19)	0.02
OLMo-2 1B	0.88	0.94	0.76 [0.69, 0.82]	0.76	1.02 (19/19)	0.01
OLMo-2 7B	0.87	0.68	0.78 [0.73, 0.83]	0.78	1.01 (19/19)	0.01

D GENDER AS A SECOND VARIABLE (P19)

Freeze and scope. P19 was designed on 26 September 2026, after all P16–P18 outcomes were known. A development pilot on three models (Pythia-160M, Qwen3-0.6B and GPT-2; development families and names) changed one thing: name templates begin with *Then* so that a name is never the first token, because in position 0 (an attention sink) the name-derived direction was degenerate in Qwen3-0.6B. The preregistration (SHA-256 `2bb1e297...`) and code (`89f52ef1...`; names and templates `664c9644...`) were then frozen before any evaluation measurement. P19 reuses the P16 sites, GDM, DM and random directions and generic-text activations; only the maps for the transport controls are fitted, exactly as in P17.

Measurements. On the P16 gender items of the evaluation families (reflexive, pronoun and possessive consumers; singular masculine and feminine recipients and donors), the coordinate along GDM, the full state and one random direction are set to $h_r + g(h_d - h_r)$ at the P17 gains, and the coordinate along GDM, DM and a random direction is mean-ablated on every singular gender text. The untrained Pythia-410M source's GDM is transported by the P16 ridge map and standardized ridge, and one trained source's GDM (Pythia-410M, or Qwen3-0.6B for that target) by the P16 ridge map, standardized ridge, raw Procrustes and the shuffled-pair map. Name items use 24 male and 24 female first names (8+8 fitting, 4+4 development, 12+12 evaluation) in the nine gender templates; recipient and donor pair a male and a female evaluation name (216 items), and GDM-name is the female-minus-male mean over fitting names at the site. Bootstraps resample the 24 evaluation families or the 12 name pairs (10,000 draws, seed 20260929).

Endpoints and outcomes. E19.1 (GDM's switch strictly increasing up to $g = 1$ in 90% of models): supported (20 of 20; over the whole grid in 20; medians 0.21 at $g = 0.5$ and 1.02 at $g = 2$). E19.2 (rank-1 share at least 0.95 in 90%): supported (20; minimum 0.9997). E19.3 (necessity ratio lower bound at least 0.5 in 80%): not supported (7; median share removed 0.45, median ratio 0.49 to the halfway full state, which removes 0.96). E19.4 (untrained source below 0.25 in every model; shuffled-pair map below 0.25 in 95%): supported (largest upper bounds 0.07 and 0.003). The trained source transports GDM with median efficiency 0.41 through the P16 ridge map, 0.42 through standardized ridge and 0.28 through raw Procrustes. E19.5 (GDM's normalized switch of name gender lower bound at least 0.5 in 80% of entering models): not supported (7 of 20; median 0.50, range 0.09–0.80); GDM-name carries a median 0.62, DM at most 0.002, and the name full patch switches a median 0.96 (clean accuracy median 1.00).

Table 7: P19 gender as a second variable, per model. GDM/full: gender switch of GDM relative to the full patch on the P16 gender items. Rank-1: energy share of the consumers-by-gain matrix. Necessity: share of signed gender log-odds removed by mean-ablating GDM, with the lower bound of its ratio to the halfway full state in parentheses. Transport: efficiency of one trained source's GDM through the P16 ridge map. Untrained: upper bound for the untrained Pythia-410M source. Names: GDM's normalized switch of first-name gender (lower bound in parentheses). Cosine: GDM with the name-derived direction.

Model	GDM/full	Rank-1	Necessity	Transport	Untrained	Names	Cosine
Pythia 70M	0.76	0.9999	0.75 (0.69)	0.81	0.02	0.54 (0.49)	0.41
Pythia 160M	0.86	0.9999	0.77 (0.74)	0.39	0.01	0.80 (0.77)	0.63
Pythia 410M	0.65	0.9997	0.38 (0.30)	0.95	0.01	0.46 (0.37)	0.38
Pythia 1B	0.54	0.9998	0.23 (0.17)	0.46	0.01	0.16 (0.11)	0.30
Pythia 1.4B	0.78	0.9999	0.33 (0.26)	0.51	0.01	0.58 (0.51)	0.43
Pythia 2.8B	0.87	1.0000	0.51 (0.52)	0.12	0.01	0.63 (0.59)	0.52
Pythia 6.9B	0.62	0.9999	0.21 (0.19)	0.30	0.01	0.35 (0.30)	0.36
Qwen3 0.6B	0.90	1.0000	0.53 (0.46)	0.43	0.01	0.79 (0.71)	0.56
Qwen3 1.7B	0.82	0.9999	0.37 (0.28)	0.24	0.01	0.34 (0.27)	0.43
Qwen3 4B	0.75	0.9999	0.28 (0.22)	0.20	0.01	0.33 (0.25)	0.39
Qwen3 8B	0.65	0.9999	0.17 (0.14)	0.18	0.01	0.30 (0.20)	0.32
SmolLM2 135M	0.88	1.0000	0.70 (0.64)	0.67	0.02	0.73 (0.66)	0.54
SmolLM2 360M	0.91	0.9999	0.67 (0.60)	0.47	0.01	0.70 (0.58)	0.43
SmolLM2 1.7B	0.84	0.9999	0.41 (0.33)	0.01	0.01	0.50 (0.37)	0.38
GPT-2 124M	0.78	0.9998	0.61 (0.62)	0.68	0.04	0.57 (0.47)	0.50
GPT-2 355M	0.75	0.9998	0.54 (0.49)	0.52	0.07	0.50 (0.40)	0.38
GPT-2 774M	0.85	0.9998	0.59 (0.57)	0.49	0.02	0.72 (0.67)	0.50
GPT-2 1.5B	0.75	0.9999	0.49 (0.48)	0.39	0.02	0.51 (0.44)	0.42
OLMo-2 1B	0.49	0.9999	0.22 (0.14)	0.19	0.01	0.10 (0.08)	0.23
OLMo-2 7B	0.41	0.9999	0.11 (0.09)	0.09	0.01	0.09 (0.07)	0.26

E NEGATION AS A THIRD VARIABLE (P20)

Freeze and scope. P20 was designed on 26 September 2026, after all P16–P19 outcomes were known, from the negation contrast pairs of an earlier design pilot by the Research Team. A phase-1 development pilot on two models (Pythia-160M and GPT-2; development templates) changed nothing; the preregistration (SHA-256 `0ff12bb2...`), code (`6e11bac0...`), launcher (`0eb79288...`) and item loader (`82d0ac65...`) were then frozen before any evaluation measurement. The site is the P16 block of each model, so the P16 generic-text activations serve for the maps; this choice, made for the maps, proved to be the study's limitation.

Items and measurements. Four hundred contrast pairs per behaviour (negative-polarity licensing: *some/any*; explicit polarity: *good/bad*, valence-balanced; consequence inference: six verb pairs) in auxiliary, modal, copula and embedded constructions; the negated text is the affirmative text with one *not* inserted, and the negation log-odds compares the continuation licensed by the negated text with the one licensed by the affirmative text. Templates are split per behaviour into fitting (25%), development (10%) and evaluation (65%; 780 pairs) sets. On evaluation pairs in both directions the recipient is patched with the donor's site state and along NDM, DM and the ten P16 random directions; NDM, the full state and one random direction are applied at the P17 gains; NDM, DM and one random direction are mean-ablated; NDM is applied on the P16 number items at the noun site; and NDM of five trained sources (Pythia-410M, Qwen3-0.6B, SmolLM2-360M, GPT-2 355M, OLMo-2 1B) and of the untrained Pythia-410M is transported by the P16 ridge map (one source also by the shuffled-pair map). A behaviour enters a model's pool when its clean accuracy is at least 0.8 (median accuracies: licensing 0.96, polarity 0.80, consequence 0.59; consequence inference is at chance in every model and never enters), and a model enters when its full patch switches the pooled pairs by at least 0.25 (5 of 20; not entering: Pythia 410M, Pythia 1B, Pythia 1.4B, Pythia 2.8B, Pythia 6.9B, Qwen3 0.6B, Qwen3 1.7B, Qwen3 4B, Qwen3 8B, SmolLM2 135M, SmolLM2 360M, SmolLM2 1.7B, GPT-2 355M, OLMo-2 1B, OLMo-2 7B). Bootstraps resample evaluation templates (10,000 draws, seed 20260930).

Endpoints and outcomes. E20.1 (NDM normalized switch lower bound at least 0.5, random below 0.1, in 80% of entering models): supported (4 of 5; median 0.62, range 0.51–0.88; random at most 0.009; full patch median 0.33; per behaviour, medians 0.66 for licensing and 0.45 for polarity; 2 entering models pool both). E20.2 (isolation in both directions at most 0.2, in 80%): supported (5; largest 0.002 and 0.001; median cosine of NDM and DM 0.03). E20.3 (transport lower bound at least 0.5 in 80% of pairs; untrained source below 0.25 in every target; shuffled map below 0.25 in 95%): not supported: none of the five preregistered sources entered, so no ordered pair could be evaluated; in the entering targets the untrained source and the shuffled-pair map carried nothing (largest upper bounds 0.09 and 0.01). E20.4: graded supported (5; median switch 0.31 at $g = 1$), necessity not supported (0; median share removed 0.08, median ratio 0.29).

Table 8: P20 negation as a third variable, per model. Behaviours: eligible consumers (N: NPI licensing, P: explicit polarity, C: consequence inference; clean accuracy at least 0.8). Full: full-patch switch at the word after *not*. NDM/full: normalized switch (lower bound in parentheses). Isolation: DM's effect on negation relative to NDM, and NDM's effect on number relative to DM. Transport: median efficiency of the other entering sources' NDM through the P16 ridge map (sources meeting the criterion). Necessity: share of signed negation log-odds removed by mean-ablating NDM. Untrained: upper bound for the untrained source.

Model	Behaviours	Full	NDM/full	Isolation	Transport	Necessity	Untrained
Pythia 70M	N	0.26	0.62 (0.52)	0.001 / 0.001	– (0/0)	0.19	0.09
Pythia 160M	N	0.33	0.88 (0.84)	0.002 / –0.001	– (0/0)	0.28	0.02
Pythia 410M	N	0.02	–	–	–	–	–
Pythia 1B	N	0.07	–	–	–	–	–
Pythia 1.4B	N	0.17	–	–	–	–	–
Pythia 2.8B	N	0.24	–	–	–	–	–
Pythia 6.9B	N	0.17	–	–	–	–	–
Qwen3 0.6B	N	0.04	–	–	–	–	–
Qwen3 1.7B	NP	0.03	–	–	–	–	–
Qwen3 4B	NP	0.09	–	–	–	–	–
Qwen3 8B	NP	0.10	–	–	–	–	–
SmolLM2 135M	N	0.17	–	–	–	–	–
SmolLM2 360M	N	0.14	–	–	–	–	–
SmolLM2 1.7B	NP	0.02	–	–	–	–	–
GPT-2 124M	N	0.25	0.71 (0.64)	0.001 / 0.000	– (0/0)	0.08	0.01
GPT-2 355M	NP	0.15	–	–	–	–	–
GPT-2 774M	NP	0.37	0.59 (0.52)	–0.000 / 0.000	– (0/0)	0.07	0.00
GPT-2 1.5B	NP	0.45	0.51 (0.43)	0.001 / –0.000	– (0/0)	0.07	0.01
OLMo-2 1B	NP	0.02	–	–	–	–	–
OLMo-2 7B	NP	0.20	–	–	–	–	–

F DETAILED DISPLAYS FOR THE RANKING, MECHANISM AND TRANSFER STUDIES

This appendix collects the detailed displays behind Sections 9 and 9: the controlled-refit interactions with cohort-specific contrasts (P14 and P14-S), the controller-mechanism families (P13 and P13-X), the source-frozen transfer (P15), and the recovery-response and refit figures summarized in Section 9; the calibration and objective contrasts (P4–P6), moved here from the main text, are shown in Figure 10; the depth and training panels of Section 7 are in Figure 9.

Table 9: Fitting-construction interaction K (Section 9), in 10^{-6} nats. Original intervals are the prespecified two-target simultaneous 95% family; the alternative selector (P14-S) was chosen after original outcomes were known, and its intervals are descriptive. Cohort columns give the two fixed-cohort contrasts averaged by K. "Matched" means the within-policy held-out recovery family passes; it does not establish recovery equivalence between policies.

Checkpoint	t	Policy	K [interval]	Cohort A	Cohort B	Held-out recovery
Pythia-410m	.50	Original	–3.98 [–9.61, 0.26]	–4.96	–3.01	Matched
	.90	Original	–4.18 [–15.28, 6.61]	–6.78	–1.59	Matched
	Alternative		Learned-operator selections and all K values unchanged			
Qwen3-0.6B	.50	Original	–1358.25 [–2606.43, –470.49]	–2718.18	1.67	Withheld
	.90	Original	–145.89 [–323.15, –35.11]	–485.67	193.89	Matched
	.50	Alternative	3.66 [0.73, 6.97]	5.65	1.67	Matched
	.90	Alternative	–136.07 [–299.32, –32.81]	–92.34	–179.81	Matched

Table 10: Controller-mechanism family statuses and three quantitative cells: restoration ϕ_{CTRL} (C1) with its family-adjusted interval, and the parameter-free ratios ρ_{B2} (B2) and ρ_{C3} (C3). Statuses: *all* = all cells supported; *part.* = partial. C1 tests restoration above 0.5; ratio cells are *accurate* within [0.8,1.25], *same order* within [0.5,2], *not supp.* outside [0.5,2]. The original Pythia-410m cohort failed its numerical gate (B/C withheld). Cross-row differences are descriptive.

Checkpoint (post-block)	R	B	C	C1 restoration	B2 ratio	C3 ratio
Qwen3-0.6B (19), original	part.	all	part.	0.85 [0.78, 0.92]	0.80 same order	0.58 same order
Pythia-410m (19), confirm.	part.	all	part.	1.07 [0.98, 1.17]	0.80 same order	0.84 accurate
Qwen3-0.6B (19), matched	part.	all	part.	0.89 [0.83, 0.96]	0.75 same order	0.59 same order
Qwen3-1.7B (19), primary	part.	part.	part.	0.64 [0.59, 0.70]	0.79 same order	0.42 not supp.
Pythia-1.4B (14), primary	part.	part.	part.	0.76 [0.71, 0.82]	0.34 not supp.	0.34 not supp.
Pythia-1.4B (19), sensitivity	part.	all	part.	1.17 [1.10, 1.24]	0.81 same order	0.96 accurate

Table 11: Source-frozen OBJECT_RELATIVE response prediction (P15). Normalized MAE and improvement over the zero-response predictor are normalized by the fixed source response scale and use the four-component adjusted interval; the FRONTED reference MAE and the paired difference (OBJECT_RELATIVE minus FRONTED MAE) are descriptive 95% intervals. All eligibility gates pass in both checkpoints.

Checkpoint	Primary	Normalized MAE	Improvement	FRONTED MAE	Difference
Pythia-410m	Supported	0.055 [0.044, 0.066]	0.557 [0.530, 0.583]	0.076 [0.072, 0.081]	–0.022 [–0.030, –0.012]
Qwen3-0.6B	Inconclusive	0.102 [0.089, 0.115]	0.380 [0.346, 0.416]	0.108 [0.101, 0.115]	–0.006 [–0.019, 0.008]

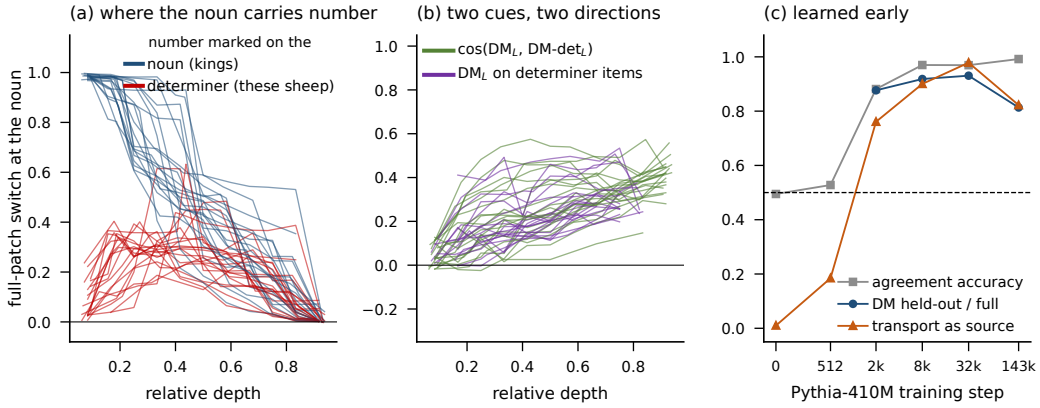


Figure 9: What the variable is. (a) Full-patch switch at the subject noun as a function of depth, for number marked on the noun (blue) and number marked only on the determiner (red); one line per model. (b) Cosine between DM$_L$ and DM-det$_L$ at each depth (green) and DM$_L$'s switch of determiner-cued number relative to the full patch, where that patch switches at least 0.1 (purple). (c) Pythia-410M across training (exploratory): clean accuracy, DM's held-out switch relative to the full patch, and median transport efficiency into the trained models.

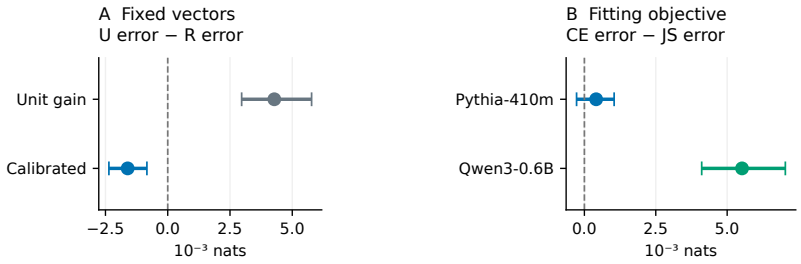


Figure 10: Same vectors, opposite rankings. (A) P4 keeps a mean-difference direction U and task-fitted directions R fixed: at unit gain R fits the donor's distribution better; after both are calibrated to minimize donor divergence, U does. (B) After independent full-divergence calibration, JS fitting improves donor fidelity in Qwen but not in Pythia (P5, P6). Bars: 95% lexical-group intervals; P4 unit gain is secondary, the other contrasts primary.

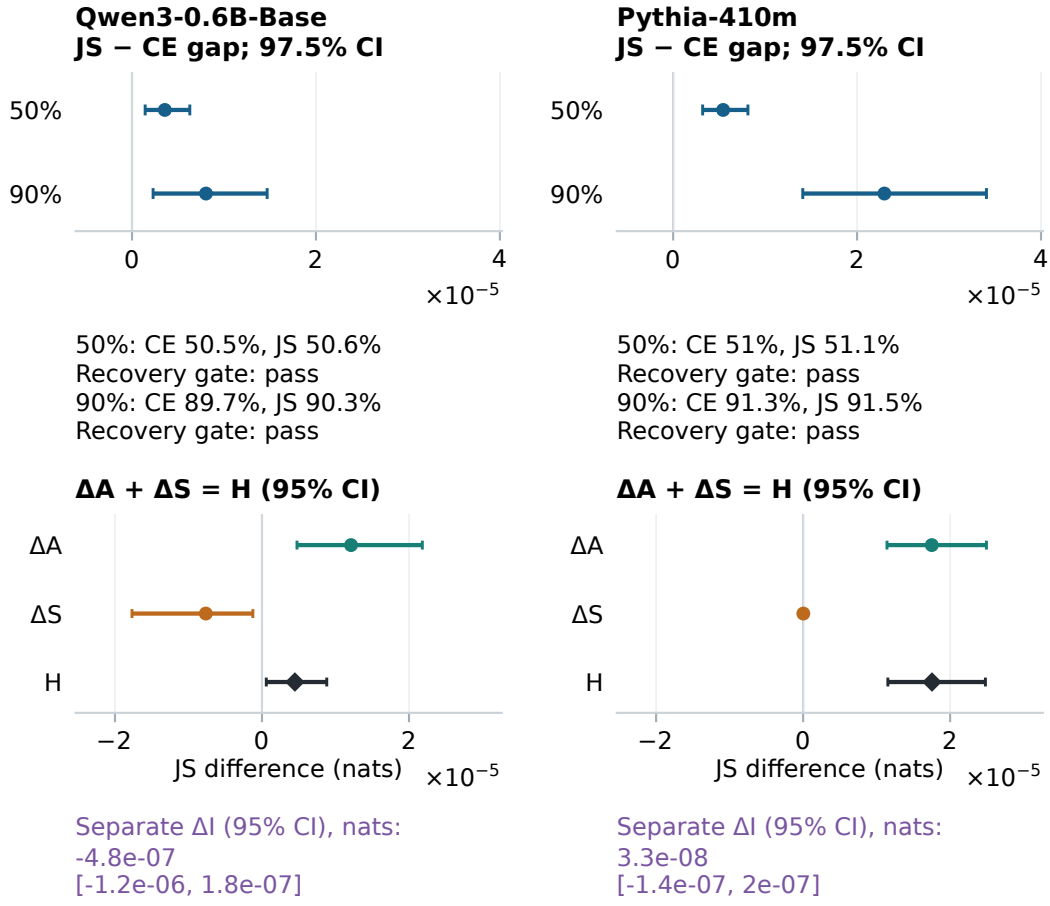


Figure 11: Prospective recovery response and executed crossed controls (P9). Top: JS-minus-CE attractor disturbance with 97.5% marginal sign intervals; positive values favor CE. Both targets pass the recovery gate in each checkpoint; neither shows the predicted reversal. Bottom: across-target changes in amplitude and signed-axis contributions with their total H and 95% intervals. Only Qwen H is primary. All intervals resample 48 lexical groups and condition on fixed fits and calibrated gains; hybrid controls do not inherit subject-recovery matching.

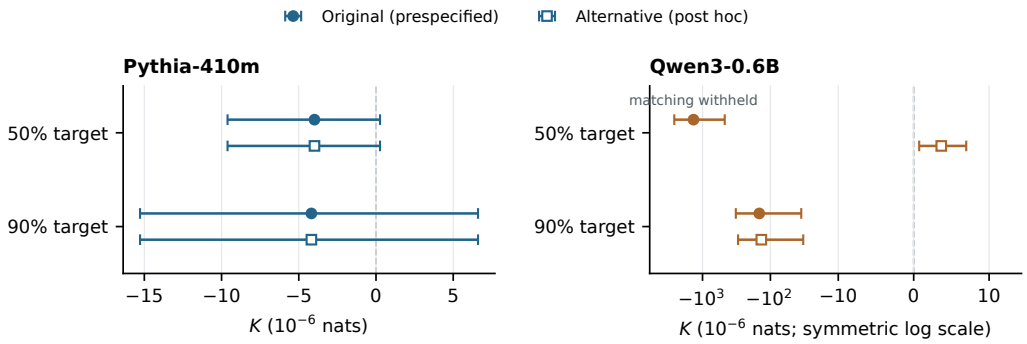


Figure 12: Fitting-construction interaction K under the original selector and the post hoc alternative (P14-S). Original intervals are the two-target simultaneous 95% family; alternative intervals are descriptive. Panels use different horizontal scales; the Qwen panel is symmetric-logarithmic beyond ± 10. Original Qwen 50% held-out recovery matching is withheld; recovery equivalence between policies is not established. Cohort-specific contrasts are in Table 9.

G STUDY LINEAGE AND POPULATION BOUNDARIES

Negation targets and operators. For negation, $y_+ = \arg\max p(\cdot \mid x_+)$ and $y_- = \arg\max p(\cdot \mid x_-)$ are frozen clean affirmative/negated targets. A pair is gradeable exactly when $y_+ \neq y_-$. Its signed score uses $z = \log p(y_-) - \log p(y_+)$. These are model-derived operational labels, not independent semantic correctness judgments. Signed donor-target log-odds movement measures direction and magnitude of change; it does not measure distance to the actual donor. For P4–P6, absolute recovery is the error reduction $D(\mathrm{noop}) - D(w, g)$ in nats; P8 onward uses the dimensionless fraction of Section 4. Complete anchor-residual replacement is an auxiliary reference, not a rank-matched ceiling or a reconstruction of every donor state. P1 also audits distinct additive-steering and coordinate-erasure operators; these are not Equation 1.

P1 reanalyzes 51 gradeable historical GPT-2 items in 44 anchor groups, preserving frozen directions and historical folds. Its 10,000 whole-anchor bootstrap intervals are exploratory. Its scalar profile uses fitting rows only, but historical U is global and the fitted direction is fold scoped. The signed audit does not invalidate the reproduced unsigned measurements; it limits their interpretation.

The retained additive operator has form $h' = h + aw$, and coordinate erasure has form $h' = h - bww^\top h$, using their archived direction conventions, fold-specific scalars and position scopes. Scalars match the original affirmative-training energy targets; they are not recalibrated on negative or held-out contexts. Signed steering asks whether the affirmative recipient moves toward its frozen negative target. Discrimination reduction and reduction in absolute target margins ask different erasure questions: crossing and then amplifying the opposite margin can reduce signed discrimination while failing neutralization.

P2 evaluates five frozen operator pairs on a fresh inventory selected by deterministic whole-group review rules. The Research Team reviewed 480 candidate/reserve pairs while blinded to assignments and fresh outcomes, accepting 462 and rejecting 18. Four candidate groups were replaced. The resulting 320 pairs comprise 80 groups, equally divided across AUX, COP, EMB and MOD. The attestation applies only to this review. The evaluation creates neither fresh fits nor five independent model replications.

P3 uses 768 agreement prefixes, 288 PP fitting rows and a held-out relative-clause primary with 192 rows in 24 lexical groups. P4 retains P3's U and three fitted directions, calibrates on those fitting rows, and evaluates 1,024 fresh prefixes across four environments, including 256 primary rows in 32 groups. Calibration in these stages is not independent of fitting.

P5 introduces the disjoint 32/16/32-group split and six fits described in Section 4. P6 repeats that exact corpus with a new U and six new fits in Qwen; its full corpus hash and deterministic example selection are recorded in the accompanying evidence package. Separate Qwen development groups select the locus; another disjoint development population supplies preflight. The P4-to-P5 change also changes corpus, directions and calibration population, so differences between their results cannot identify a mediated contribution of fitting loss.

Later populations. P8 and P9 use fresh tokenizer-balanced inventories (Appendices L and N). P10 uses a further disjoint 16/16/48-group development/calibration/evaluation inventory in Qwen. P11 reanalyzes the retained P9 evaluation populations. P12 fits and evaluates a new cohort with a disjoint lexicon; its 48 study groups later supply the exposed source groups for the P13-X and P15 controllers. P13 introduces 64 fresh groups (16 development, 48 evaluation) in three constructions for Qwen and the original Pythia run. After the original Pythia run failed its numerical gate, a separate Pythia-410m confirmation used the next 128 nouns of the same registered shuffle, disjoint from the P13 corpus; P13-X and the P15 destination reuse this confirmation lexicon. P14 uses two fixed 32-group fitting cohorts and a 16-group calibration population with prior exposure, and a 48-group evaluation lexicon disjoint from the recorded P1–P13 forms. Studies sharing a lexicon are not independent lexical replications.

H FRESH NEGATION: PRESERVED NON-ADJUDICATION

The 56 gradeable P2 pairs comprise AUX 8/80, COP 17/80, EMB 9/80 and MOD 22/80. The minimum observed counts pass, but the construction-stratified group bootstrap can still contain no

gradeable AUX item. The primary statistic equally weights four construction means after averaging all five operator contrasts within pair. Omitting empty groups or undefined draws would change the declared population or conditioning; the original rule withholds the primary interval. The interval among the 19,777 defined draws, $[-0.041218, -0.012251]$, is diagnostic only.

All-pairs JS advantage, free minus constrained, is -0.000607 $[-0.000782, -0.000439]$; the negative-target log-probability contrast is $+0.024485$ [0.019861, 0.029289]. These prespecified secondaries describe different properties. They neither rescue the primary nor provide independent semantic labels.

The model saved all 12 full-vocabulary arrays before a normalization check failed. A separately declared offline derivation cast each row to FP64 and subtracted its FP64 logsumexp. All clean argmax pairs were unchanged; within-row log-odds are algebraically invariant, while CE/JS were recomputed. The original gate remains failed. Independent numerical reconstruction and exact original-module replay validate the derivation, not the failed model execution.

I METRIC AND ENERGY INTERPRETATION

Let T be the target-token pair, $a = p(T)$, $b = q(T)$, and $p_T = p(\cdot \mid T)$, $q_T = q(\cdot \mid T)$ when defined. The full-JS coordinate sum over T is not ordinary $\mathrm{JS}(p_T, q_T)$. With $c = (a + b)/2$ and mixture $m_T = (ap_T + bq_T)/(a + b)$ when $a + b > 0$, the target-coordinate contribution is

$$\tfrac{1}{2}a \log(a/c) + \tfrac{1}{2}b \log(b/c) + \tfrac{1}{2}a\mathrm{KL}(p_T \| m_T) + \tfrac{1}{2}b\mathrm{KL}(q_T \| m_T). \tag{2}$$

Zero-mass terms use their continuous limits. The complement has the analogous expression, and the two coordinate contributions sum to full JS. Ordinary equal-weight conditional JS discards mass disagreement and uses a different conditional mixture when $a \neq b$. Therefore these scores can rank interventions differently without a numerical contradiction. Tokens outside T may still encode grammatical alternatives.

The energy control selects a calibration subject-RMS multiplier relative to U at unit gain, using FP64 geometry and a fixed empirical agreement tolerance. Off-grid gains are permitted by that declared policy. A zero U reference withholds all energy inference. An unattainable fitted seed withholds the complete-objective comparison requiring it; seeds are never dropped. Evaluation pools squared norms before taking the square root and reports subject and attractor perturbations separately. Energy matching is not outcome matching and does not establish equal useful recovery.

J FITTING, COMPUTATIONAL IDENTITY AND RELEASE SCOPE

The objective studies use paired seeds 2111, 2112 and 2113, batch size 24 with a final batch of 16, four epochs, and 44 updates per fit. The common gain grid is $\{-4, -2, -1, 0, .25, .5, .75, .875, 1, 1.125, 1.25, 1.5, 1.75, 2, 3, 4, 8, 16\}$. Qwen's full-JS gains are 1.5 for U and 1.125 for all six fits. No selected Qwen operator has zero gain or lies at a grid boundary. P12 and P14 refit with their own seeds (3111–3113 and 14111–14113) and the same 44-update budget. P13 and P15 fit no directions by optimization: their controller is a normalized sum of mean subject flips, and P15 additionally fits one source response slope.

The checkpoint identities are:

`Qwen/Qwen3-0.6B-Base`
Revision: `da87bfb608c14b7cf20ba1ce41287e8de496c0cd`

`EleutherAI/pythia-410m-deduped`
Revision: `c4fc8d586d62df497f1f9b69d66d3ca419992d3e`

`Qwen/Qwen3-1.7B-Base`
Revision: `ea980cb0a6c2ae4b936e82123acc929f1cec04c1`

`EleutherAI/pythia-1.4b-deduped`
Revision: `554d9c1bae3877f740aece41feb90f912cf9fedc`

Qwen computation uses its native eager attention in FP32 on CPU; stored BF16 checkpoint precision is not restored by conversion. P13-X runs every condition on CPU in FP32 with five threads and

batch 24, because both larger GPU routes failed an unchanged CPU-envelope certification criterion. These sampled implementation checks are not a second study.

The release separates three operations. Claim verification checks every reported value against exact saved-result snapshots by hash and JSON pointer. Compact replay re-runs each frozen statistical analysis from included small records and selected arrays, without model weights. Full raw reconstruction and fresh model execution require large arrays and pinned external weights. Large full-vocabulary arrays are omitted from the compact records; every omission is inventoried with its path, size and locked SHA-256. For example, Qwen's three P6 raw locations contain 512 files totaling 69,857,051,648 bytes, and the P9 compact archive omits 358 logit files totaling 27,052,192,512 bytes. Frozen study sources are included unchanged; corrections and amendments are separate files. The repository does not silently merge experiment packages or rewrite their frozen sources.

Historical SAE, geometric-angle and attention-position analyses motivated the original programme but do not establish the reported results. The original records and retractions remain preserved separately. A claim to broader representation invariance would require a defined cross-space map, admissibility assumptions, an invariance estimand, an equivalence tolerance and held-out environments. None is established by the present checkpoints.

K Objective-comparison design and absolute results

P5 and P6 use 80 lexical groups partitioned into 32 fitting, 16 calibration and 32 evaluation groups: 256, 128 and 1,280 rows, respectively. Fitting and calibration use POST_OBJECT only. Evaluation crosses prepositional-phrase placement (pre/post) with noun role (object/animate subject) and adds a paired PRE_ANIMATE discourse-prefix challenge. Each environment has 256 rows. The prefix changes content, capitalization and length; the role manipulation exchanges lexical roles and does not isolate animacy. For example, the recipient "Near the bracelet, the lecturer probably" has subject donor "Near the bracelet, the lecturers probably" and attractor donor "Near the bracelets, the lecturer probably." The former targets the clean donor distribution; the latter is compared with the unchanged recipient. The complete environment examples are in Table 16.

Three paired seeds per objective initialize from that checkpoint's newly estimated U. Both procedures receive 44 Adam updates at unit gain, learning rate 0.02, the same epoch permutations, and frozen model weights. CE uses the donor's grammatical number label, never the donor model's argmax. JS fits the clean donor's full distribution. Final-update directions are retained without early stopping or seed selection. Equal opportunities do not imply equal convergence. After direction locking, independent calibration considers 18 gains from -4 to 16, with the same grid for each direction and ties resolved toward the smaller numerical gain. Conditional CE, conditional JS and full JS are separate policies; unit gain and calibration-RMS energy matching are additional declared controls. Operators are locked before evaluation. The sole primary is CE-minus-JS donor full-JS error on PRE_ANIMATE with each direction independently full-JS calibrated. Positive values favor JS fitting.

Pythia uses the inherited post-block L19 locus. Qwen selects a common locus on a separate development population: the earliest of all 28 post-block locations reaching half of the no-op gap under literal full-subject replacement. It selects zero-based layer 19 (block 20), before fitting or calibration. This rule uses no CE-minus-JS contrast and does not establish homologous depth or rank-one capacity. Qwen uses its native sequence computation, attention, normalization and tokenizer; hidden-space vectors are not transferred between models. For P4–P6, percentile intervals use 20,000 whole-group bootstrap draws retaining paired templates, number cells, environments and seeds together; the objective comparison uses bootstrap seed 90512, and P3 uses 20,000 draws over its 24 relative-clause groups with seed 90210. The P1 and P2 resampling rules differ and are retained separately.

In P4, conditional-CE calibration chooses gain 16 for every direction. Primary donor accuracy remains 100%, yet full-JS errors increase to 0.5705090 and 0.5933571, exceeding no-op. Clean probability mass on the two specified target tokens is only 0.7197%. Perfect conditional accuracy consequently does not establish faithful next-token generation. Conditional and full-JS calibration select identical operators there, so P4 identifies no vocabulary-coverage advantage for scalar selection on its grid. The P4 preservation comparison is not matched on useful recovery or energy.

Table 12: PRE_ANIMATE under full-JS calibration. Seed-dependent entries average measurements, not distributions or vectors. All methods have 100% conditional donor accuracy. Disturbance refers to attractor intervention against the unchanged recipient. Values are not a common cross-model difficulty scale.

Checkpoint	Procedure	Donor JS	Attractor JS	Target mass
Pythia-410m	U	0.0375203	0.0003917	0.0082144
	CE	0.0380399	0.0001258	0.0081483
	JS	0.0376281	0.0002366	0.0076647
Qwen3-0.6B-Base	U	0.0475288	0.0006340	0.0268484
	CE	0.0510029	0.0002318	0.0277484
	JS	0.0454821	0.0003003	0.0259475

No-op donor JS: Pythia 0.1368024; Qwen 0.2657998.

In Qwen, U-minus-JS error is $+0.0020467$ [0.0008806, 0.0032929], while U-minus-CE is -0.0034740 [-0.0046944,-0.0022938]. Thus JS fitting improves fidelity beyond calibrated U in this setting; CE does not. All three improve over no-op. These are secondary comparisons, not additional primaries. Ordinary conditional donor JS favors CE over JS among the fitted procedures: U, CE and JS score 0.0039503, 0.0055334 and 0.0069320. By their sample means, the three criteria each favor a different procedure: full donor JS favors JS fitting, conditional donor JS favors U, and attractor preservation favors CE. These point orderings do not imply that every pairwise difference is resolved. The full-JS target-coordinate contribution differs from ordinary conditional JS because it also contains target-mass disagreement; the remaining vocabulary is not necessarily unrelated to agreement (Appendix I).

Context, scalar policy and checkpoint change the relative effect, while calibration energy matching does not guarantee matched evaluation energy or recovery. These dependent secondary comparisons and the paired checkpoint interaction remain in Appendix M; they identify no model-family or training-scale cause.

Constructed populations and optimization. P6 repeats P5's exact texts, giving checkpoint replication rather than independent stimulus replication. Equal-length number counterparts hold in P5 under Pythia, but Qwen has 288 directed subject and 272 directed attractor swaps with unequal prefix lengths; all are retained. P4 also has number-linked length changes. Matching counts would still leave lexical identity, morphology and context. The prior blinded review of 480 negation candidates does not attest to any agreement corpus. Earlier agreement stimuli have deterministic structural checks without a separate linguistic-review attestation. Later inventories add fresh tokenizer-balanced lexicons with documented lexical/template screening and targeted dictionary checks. Their scene-based synthetic prefixes are neither frequency matched nor sampled from natural text, and none constitutes an independent linguistic-panel or naturalness-rating study. Both Qwen fitting objectives improve their own full-fitting-set metric and both metrics over U. This supports objective-sensitive learning, not convergence. Three seeds, 44 updates and small across-seed variation characterize the declared procedures; they do not establish global optima or uncertainty over new fitting/calibration samples.

L Fresh recovery-targeting extension

The exact corpus, rejected lexical candidates, prospective specification, gain profiles, numerical checks and saved-output reconstruction accompany P8. FRONTED prefixes have form "In the scene, next to the attractor, the subject probably"; POST places the same prepositional phrase after the subject. "In this picture," supplies the second context. These constructions and concrete noun categories do not isolate animacy or token position. Matching token counts removes one observed number-linked difference, not every lexical or contextual confound. All comparison intervals below are conditional on the frozen fits and calibrated gains. Every comparison except Qwen FRONTED target-policy JS-minus-CE preservation is secondary and descriptive; there is no family-wise correction, pooled checkpoint endpoint or full-procedure uncertainty. The narrower gain grid differs from the earlier P5/P6 best-full-JS policy. Unattainability is limited to the declared finite gain range and search; it is not a mathematical impossibility result. If any fitted seed is unavailable, its ob-

Table 13: All declared P8 preservation contrasts. Positive values favor the procedure subtracted. "Pass" denotes the pair-specific comparable-recovery gate; unavailable arms remain unavailable.

Model	Policy	Context	Contrast	Nats	95% interval	Gate
Pythia	target	Fronted	JS-CE	+0.0000111	[+0.0000063,+0.0000168]	Pass
Pythia	target	Fronted	U-CE	+0.0000222	[+0.0000120,+0.0000344]	Pass
Pythia	target	Fronted	U-JS	+0.0000111	[+0.0000037,+0.0000195]	Pass
Pythia	target	Post	JS-CE	+0.0000583	[+0.0000230,+0.0000956]	Pass
Pythia	target	Post	U-CE	+0.0003068	[+0.0002094,+0.0004139]	Pass
Pythia	target	Post	U-JS	+0.0002485	[+0.0001649,+0.0003399]	Pass
Pythia	best-grid	Fronted	JS-CE	+0.0000513	[+0.0000254,+0.0000812]	Pass
Pythia	best-grid	Fronted	U-CE	+0.0001579	[+0.0000848,+0.0002456]	Pass
Pythia	best-grid	Fronted	U-JS	+0.0001067	[+0.0000490,+0.0001758]	Pass
Pythia	best-grid	Post	JS-CE	-0.0001683	[-0.0003398,+0.0000054]	Pass
Pythia	best-grid	Post	U-CE	+0.0022315	[+0.0015968,+0.0029301]	Pass
Pythia	best-grid	Post	U-JS	+0.0023998	[+0.0017751,+0.0030882]	Pass
Qwen	target	Fronted	JS-CE	+0.0000033	[+0.0000011,+0.0000062]	Pass
Qwen	target	Fronted	U-CE	+0.0000159	[+0.0000064,+0.0000296]	Pass
Qwen	target	Fronted	U-JS	+0.0000126	[+0.0000044,+0.0000252]	Pass
Qwen	target	Post	JS-CE	+0.0000096	[-0.0000073,+0.0000219]	Pass
Qwen	target	Post	U-CE	+0.0000406	[+0.0000186,+0.0000584]	Pass
Qwen	target	Post	U-JS	+0.0000309	[+0.0000208,+0.0000418]	Pass
Qwen	best-grid	Fronted	JS-CE	-0.0002210	[-0.0004651,-0.0000715]	Pass
Qwen	best-grid	Fronted	U-CE	-0.0000364	[-0.0001256,+0.0000253]	Pass
Qwen	best-grid	Fronted	U-JS	+0.0001846	[+0.0000756,+0.0003497]	No
Qwen	best-grid	Post	JS-CE	-0.0015406	[-0.0020887,-0.0011031]	No
Qwen	best-grid	Post	U-CE	-0.0006603	[-0.0010720,-0.0003613]	Pass
Qwen	best-grid	Post	U-JS	+0.0008802	[+0.0006798,+0.0011080]	Pass

jective aggregate is withheld. The equivalence gate checks mean normalized recovery, not equality of rowwise output distributions, perturbation energy or all unrelated behavior. The source records also retain achieved recovery, nominal-target displacement, conditional metrics and per-direction quantities.

M Additional historical and calibration results

M.1 Exact agreement examples

For the PRE_ANIMATE example in Table 16, the subject donor is "Near the bracelet, the lecturers probably" and the attractor donor is "Near the bracelets, the lecturer probably." Subject interchange targets the clean subject-donor distribution; attractor interchange is evaluated against the unchanged recipient. This counterfactual definition is distinct from testing whether the model's most likely unrestricted continuation is the grammatical target.

M.2 Historical movement results in detail

The original GPT-2 negation comparison assigned 92.3% of its sample-mean improvement over raw U to the fitted direction after a learned scalar control. The exploratory P1 reanalysis strengthens that control using a training-only finite full-vocabulary cross-entropy search. U's held-out signed effect rises from 0.1172 to 0.2614 nats, against R's 0.7960. R-minus-profiled-U remains 0.5346 [0.2125, 0.8590], but the descriptive sample-mean remainder becomes 72.7%. This comparison across different gains is not a unique attribution to geometry. The chosen gains, 6.5–7.5, are not established global optima. Historical global-U versus fold-local-R scope also remains asymmetric.

The signed editing audit changes the practical interpretation. Paired affirmative-to-negated steering-advantage intervals span zero at all three tested energies. R reduces pair discrimination more, but its advantage in reducing average absolute target margins is supported only at the lowest energy. At the highest energy, pair-discrimination inversion occurs on 9/51 items for R and 1/51 for U. Matching energy on affirmative training contexts does not match negative held-out energy. Thus the repro-

Table 14: Achieved P8 evaluation recovery and its 95% lexical-group interval for every declared method, checkpoint, policy and environment. Recovery is `1 − E/B`, with equal-group mean subject-donor error E and the shared clean-gap mean B. CE and JS average the three seed-specific measurements before grouping; U has one fixed direction. The target policy calibrates to 0.5 on separate FRONTED groups, which does not force evaluation recovery to equal 0.5. Best-grid instead minimizes donor error on the ten calibration gains. These values support the pair-specific gates in Table 15; they add no primary endpoint.

Model	Policy	Context	Method	Recovery	95% interval	Status
Pythia	Target	Fronted	U	0.50601	[0.49179, 0.52051]	Est.
Pythia	Target	Fronted	CE	0.50715	[0.49608, 0.51823]	Est.
Pythia	Target	Fronted	JS	0.50513	[0.49471, 0.51551]	Est.
Pythia	Target	Post	U	0.56453	[0.55160, 0.57682]	Est.
Pythia	Target	Post	CE	0.55266	[0.54182, 0.56309]	Est.
Pythia	Target	Post	JS	0.56307	[0.55252, 0.57318]	Est.
Pythia	Best-grid	Fronted	U	0.91863	[0.90873, 0.92716]	Est.
Pythia	Best-grid	Fronted	CE	0.93156	[0.92212, 0.93987]	Est.
Pythia	Best-grid	Fronted	JS	0.92545	[0.91568, 0.93382]	Est.
Pythia	Best-grid	Post	U	0.93456	[0.92597, 0.94254]	Est.
Pythia	Best-grid	Post	CE	0.94055	[0.93235, 0.94811]	Est.
Pythia	Best-grid	Post	JS	0.93879	[0.93047, 0.94647]	Est.
Qwen	Target	Fronted	U	0.51564	[0.49684, 0.53543]	Est.
Qwen	Target	Fronted	CE	0.52365	[0.50567, 0.54316]	Est.
Qwen	Target	Fronted	JS	0.52207	[0.50413, 0.54140]	Est.
Qwen	Target	Post	U	0.62782	[0.61178, 0.64345]	Est.
Qwen	Target	Post	CE	0.62973	[0.61641, 0.64276]	Est.
Qwen	Target	Post	JS	0.62685	[0.61348, 0.63991]	Est.
Qwen	Best-grid	Fronted	U	0.91741	[0.90755, 0.92641]	Est.
Qwen	Best-grid	Fronted	CE	0.92868	[0.91815, 0.93803]	Est.
Qwen	Best-grid	Fronted	JS	0.93628	[0.92614, 0.94529]	Est.
Qwen	Best-grid	Post	U	0.93699	[0.92897, 0.94434]	Est.
Qwen	Best-grid	Post	CE	0.93209	[0.92426, 0.93924]	Est.
Qwen	Best-grid	Post	JS	0.94900	[0.94160, 0.95575]	Est.

"Est." denotes the saved status `ESTIMABLE`. All 24 method estimates and intervals are available in this run. Unavailable methods or undefined normalized intervals would be retained as such, never assigned zero or omitted. Intervals use 20,000 shared whole-lexical-group resamples of 48 evaluation groups (seed 38123), retaining paired cells, contexts, directions and checkpoints and recomputing the denominator in every draw. They condition on the frozen fits, loci, calibration data and selected gains. No denominator draw or fitted seed is discarded. These are mean-recovery quantities, not equality of individual predictions or perturbation energy.

Table 15: All 24 declared P8 paired recovery differences and comparable-recovery gates. Each difference is first method minus second method. The asterisk marks the pair supporting the sole primary preservation contrast: Qwen FRONTED target-policy JS-minus-CE attractor divergence. Every other preservation comparison is secondary and descriptive, with no family-wise correction. A gate is an interpretation condition, not a separate test of preservation.

Model	Policy	Context	Contrast	Recovery diff.	95% interval	Gate	Status
Pythia	Target	Fronted	JS–CE	-0.00201	[-0.00485, +0.00089]	Pass	Est.
Pythia	Target	Fronted	U–CE	-0.00113	[-0.00746, +0.00550]	Pass	Est.
Pythia	Target	Fronted	U–JS	+0.00088	[-0.00534, +0.00730]	Pass	Est.
Pythia	Target	Post	JS–CE	+0.01042	[+0.00808, +0.01259]	Pass	Est.
Pythia	Target	Post	U–CE	+0.01187	[+0.00758, +0.01617]	Pass	Est.
Pythia	Target	Post	U–JS	+0.00145	[-0.00315, +0.00588]	Pass	Est.
Pythia	Best-grid	Fronted	JS–CE	-0.00611	[-0.00788, -0.00434]	Pass	Est.
Pythia	Best-grid	Fronted	U–CE	-0.01293	[-0.01514, -0.01077]	Pass	Est.
Pythia	Best-grid	Fronted	U–JS	-0.00682	[-0.00833, -0.00523]	Pass	Est.
Pythia	Best-grid	Post	JS–CE	-0.00176	[-0.00352, +0.00002]	Pass	Est.
Pythia	Best-grid	Post	U–CE	-0.00599	[-0.00808, -0.00393]	Pass	Est.
Pythia	Best-grid	Post	U–JS	-0.00423	[-0.00560, -0.00290]	Pass	Est.
Qwen	Target	Fronted	JS–CE*	-0.00158	[-0.00358, +0.00048]	Pass	Est.
Qwen	Target	Fronted	U–CE	-0.00800	[-0.01360, -0.00204]	Pass	Est.
Qwen	Target	Fronted	U–JS	-0.00643	[-0.01138, -0.00110]	Pass	Est.
Qwen	Target	Post	JS–CE	-0.00288	[-0.00491, -0.00080]	Pass	Est.
Qwen	Target	Post	U–CE	-0.00191	[-0.00676, +0.00327]	Pass	Est.
Qwen	Target	Post	U–JS	+0.00097	[-0.00364, +0.00571]	Pass	Est.
Qwen	Best-grid	Fronted	JS–CE	+0.00760	[+0.00431, +0.01109]	Pass	Est.
Qwen	Best-grid	Fronted	U–CE	-0.01127	[-0.01571, -0.00652]	Pass	Est.
Qwen	Best-grid	Fronted	U–JS	-0.01887	[-0.02130, -0.01641]	No	Est.
Qwen	Best-grid	Post	JS–CE	+0.01691	[+0.01346, +0.02060]	No	Est.
Qwen	Best-grid	Post	U–CE	+0.00490	[+0.00087, +0.00889]	Pass	Est.
Qwen	Best-grid	Post	U–JS	-0.01201	[-0.01413, -0.00990]	Pass	Est.

"Pass" requires both complete methods' normalized-recovery 95% lower bounds to exceed zero and the paired recovery-difference 95% interval to lie strictly inside (-0.02, 0.02). Equality at either margin fails the gate. "No" means the comparable-recovery interpretation is not established; its descriptive preservation contrast remains reported. "Est." denotes `ESTIMABLE`. All 24 paired differences are available here; any unavailable method or undefined normalized interval would retain its saved status and could not pass the gate. A missing fitted seed withholds the complete-objective aggregate rather than being dropped. The margin is two percentage points of the common clean gap and was fixed before outcomes. Gate decisions use full-precision saved values; displayed values are rounded. The two failed gates are Qwen FRONTED best-grid U-minus-JS and Qwen POST best-grid JS-minus-CE. They fail on recovery equivalence, despite positive mean useful recovery in each method. Within-policy passage does not equate the recovery levels of different policies, nor establish rowwise distributional equality or preservation of all unrelated behavior.

Table 16: Exact examples from the shared P5/P6 corpus, using the first evaluation group in lexical-ID order, preposition *near*, and singular subject/attractor. Selection uses no model outcome. Prefixes end at the intervention anchor; the target alternatives follow them. All number cells and both prepositions remain in evaluation.

Environment	Recipient prefix
POST_OBJECT	The bracelet near the lecturer probably
PRE_OBJECT	Near the lecturer, the bracelet probably
POST_ANIMATE	The lecturer near the bracelet probably
PRE_ANIMATE	Near the bracelet, the lecturer probably
PRE_ANIMATE_PREFIX	Remarkably, near the bracelet, the lecturer probably

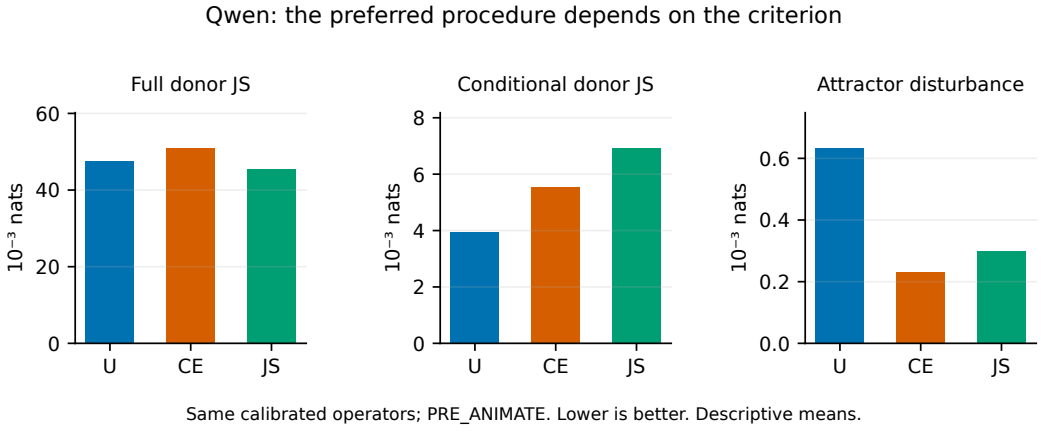


Figure 13: Different criteria favor different procedures under the same Qwen full-JS-calibrated operators. These are descriptive means over the same primary-environment groups, with seed-specific measurements averaged before grouping. The full-distribution, conditional-distribution and attractor-reference criteria must not be combined into one score. Intervals for the prespecified primary and relevant secondary contrasts are reported in the text; this plot adds no interval or endpoint.

duced unsigned editing premium cannot be reported as reliable directional steering or uniformly better neutralization.

The fresh Pythia negation evaluation does not repair that limitation. Its equal-construction R-minus-constrained estimate is -0.024646 nats, with 56/320 gradeable pairs, but 223/20,000 construction-stratified bootstrap draws contain no gradeable AUX observations. The prespecified rule withholds the primary interval and adjudication. Its all-pairs JS contrast favors the constrained comparator, whereas its target-log-probability contrast favors R (Appendix H). The original model run also failed a normalization gate; a disclosed offline derivation does not relabel it as a successful confirmation.

Agreement makes the intended grammatical target explicit, but movement alone is still insufficient. P3's held-out relative-clause primary favors R by 4.2279 nats [4.1073, 4.3471]. All CE profiles select boundary gain 16. Donor log-odds absolute error is 43.8126 for R versus 39.5847 for U, both worse than no intervention. The positive primary establishes greater signed movement, not faithful transfer. This motivated the subsequent distributional assay; it was not retrospectively replaced by its endpoint.

Context, calibration and energy. Qwen's full-JS-calibrated CE-minus-JS contrast is positive in all five environments. Its paired prefix-minus-base contrast is -0.0023327 $[-0.0035826, -0.0012293]$, indicating a smaller relative JS advantage with the added discourse prefix. All five scalar policies have positive PRE_ANIMATE contrasts, but the sizes differ. These comparisons share fits and observations and are not independent replications.

All seven Qwen energy arms match calibration subject RMS. PRE_ANIMATE subject RMS is 23.15978, 23.10797 and 23.16468 for U, CE and JS; attractor RMS is 2.71476, 2.00487 and 2.24895.

Pooled RMS takes the square root after averaging squared norms over rows and seeds. Calibration energy matching guarantees neither matched test energy nor matched recovery. No energy comparison is withheld here; the protocol would withhold inference for a zero reference or any fitted-objective comparison requiring an unattainable seed.

Direct paired checkpoint comparison. The secondary Qwen-minus-Pythia difference of CE-minus-JS primary contrasts is 0.0051090 nats [0.0035424, 0.0067680]. The paired comparison resamples the same 32 lexical groups jointly. Dividing each contrast by its own mean no-op gap, recomputed in every bootstrap draw, gives a dimensionless paired difference of 0.0177602 [0.0111458, 0.0243734]. All required denominators exceed the declared floor. There is no JS between differently indexed model vocabularies.

This interaction is policy dependent. Under CE calibration, the corresponding absolute difference is -0.0109276 [-0.0133826,-0.0082453]. That does not establish better absolute Pythia behavior: its CE-calibrated donor errors exceed its no-op gap. Model, tokenizer, pretraining, scale, runtime and locus rule co-vary; the paired contrast identifies none of these as its cause.

Numerical acceptance is separate from scientific outcome. P1/P3 audits reconstruct retained target-level quantities and use sampled forward checks; they do not have the same full-vocabulary audit scope as P4–P6. P2 preserves its failed model status and a separately specified offline normalization derivation. P4 preserves an initial normalization failure and a new accepted attempt. P5's accepted recovery inherits its primary rank-one outputs and corrects only an auxiliary full-replacement reference. P6 completes a fresh model run; its initial analysis fails on a catalog lookup, followed by an explicitly bound validation-only adapter that leaves all 21 frozen sources and model outputs unchanged. The unchanged independent P6 auditor passes 18,815 comparisons, and separate amended replay reproduces results exactly. P8 preserves full-vocabulary outputs and separately reconstructs all saved calibration and evaluation metrics. Archived norm fields are cross-checked, not independently reconstructed from activation deltas. A separate frequency-weight calculation checks method-aggregate intervals and gates. Additional ordinary-forward checks cover 29 conditions on 32 fixed rows per checkpoint, spanning both constructions and contexts; they do not exhaust evaluation rows. No successful audit promotes a secondary endpoint or converts P2's withheld primary into a result.

M.3 P8 DESIGN AND COMPLETE RECOVERY-TARGETING NARRATIVE

The previous objective comparison minimizes each direction's donor error independently. Its preservation difference could therefore accompany different amounts of useful recovery. We test this ambiguity prospectively on a fresh corpus, reusing all seven accepted directions and the inherited locus in each checkpoint. No direction is refitted or selected by its fresh outcome; seed-specific measurements are averaged after intervention.

Let B be mean no-op subject-donor JS and $E_d(g)$ the intervened subject-donor error, with equal lexical-group weights. We calibrate the fraction $R_d(g) = 1 - E_d(g)/B$ to 0.5. This common clean-gap denominator differs from normalizing by each direction's own best performance. Ten gains in $[0, 4]$ are evaluated; the smallest sampled acceptable gain is used, otherwise the first sampled ascending bracket receives at most 12 actual midpoint evaluations, with recovery tolerance 0.001. No preservation outcome selects gains. Every policy and unavailable status is locked before evaluation computation. The secondary best-grid policy minimizes calibration donor JS on these ten points.

The new inventory has 128 distinct concrete count-noun paradigms absent from P3–P5. Sixty-four lexical groups are split into 16 calibration and 48 evaluation groups. All four subject/attractor number cells cross two scene prefixes. Calibration uses FRONTED only (128 rows); evaluation uses FRONTED and POST (768 rows). Both pinned tokenizers preserve complete prefix length for every number counterpart. The primary tests fresh lexical groups within the calibration construction; POST tests secondary transfer of the calibrated gains. The directions themselves were fitted on P5 POST_OBJECT rows, in which, as in POST, the attractor intervenes. Targeted lexical review excludes one variable plural before output access; it is not the earlier negation-review attestation.

The sole primary is Qwen FRONTED target-policy JS-minus-CE attractor divergence. The stronger comparable-recovery interpretation additionally requires the paired 95% interval for mean evalua-

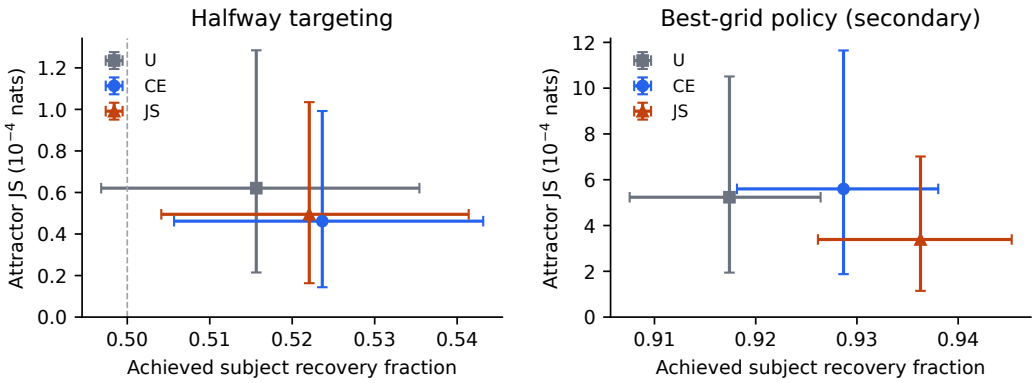


Figure 14: Qwen FRONTED preservation at two operating points on the same fresh evaluation groups and frozen directions. Left: prospective halfway targeting. Right: the prespecified secondary best-grid policy. Bars are marginal 95% group intervals, not joint regions. The CE–JS paired recovery gate passes within each panel; recovery is not matched across panels. Axes differ. The U–JS best-grid recovery gate fails, so that comparison has no matched-recovery interpretation.

tion recovery difference strictly inside $[-0.02, 0.02]$, and each complete objective's recovery lower bound above zero. The margin is two percentage points of the clean gap, fixed before outcomes. Twenty thousand paired whole-group resamples preserve all seeds and cells and recompute the denominator. These intervals condition on the frozen fits and calibration choices. Calibration matching alone cannot establish evaluation matching.

The sole Qwen FRONTED primary, JS-minus-CE attractor-reference divergence, is $+0.0000033[+0.0000011, +0.0000062]$ nats. Its interval supports lower CE disturbance. The paired JS-minus-CE recovery difference is $-0.00158[-0.00358, +0.00048]$. Both recovery lower bounds are positive and the paired interval is strictly within the prespecified equivalence margin; the comparable-mean-recovery interpretation passes. CE achieves recovery 0.5236; JS achieves 0.5221.

Pythia's corresponding secondary contrast is $+0.0000111[+0.0000063, +0.0000168]$ nats, with its comparable-recovery gate passing. Under target calibration, both learned procedures disturb the attractor reference less than U in both checkpoints and both constructions; every corresponding secondary interval is positive and every pair-specific recovery gate passes (Appendix L). These are conditional preservation benefits beyond the mean-difference baseline.

A secondary policy reverses the preservation ranking. On the same Qwen FRONTED examples and directions, best-grid calibration favors JS preservation: JS-minus-CE is $-0.0002210[-0.0004651, -0.0000715]$ nats. CE and JS recover 0.9287 and 0.9363 of the clean gap; their paired recovery difference is $+0.00760[+0.00431, +0.01109]$, so this policy's comparable-recovery gate also passes. Thus opposite preservation orderings occur at approximately 0.52 and 0.93 recovery (Figure 14). This is a secondary observed reversal, not a new primary, a formal cross-policy interaction test or matching across the two recovery levels. No evaluation output retunes gains.

Clean FRONTED conditional agreement accuracy is 100.00% in Pythia and 100.00% in Qwen; the specified target pair receives 5.756% and 8.792% of full probability mass. These are descriptive corpus diagnostics, not additional inferential endpoints.

N Recovery-response protocol and controlled perturbations

Prospective scope and inherited directions. This follow-up was specified after inspection of P8 and before P9 model execution. It tests two recovery targets under one calibration rule on new lexical groups, retaining all seven accepted directions in each checkpoint: U and the three paired CE/JS fits with seeds 2111, 2112 and 2113. The checkpoints are Pythia-410m-deduped and Qwen3-

0.6B-Base at their previously pinned revisions; both retain their own zero-based post-block locus 19. No direction, seed or locus is selected again. The inherited fits use 44 Adam updates at unit gain, learning rate 0.02, with the same initialization and data permutations within each CE/JS pair. Their finite budgets do not establish convergence. The two loci have different selection histories and are not assumed homologous; vectors are not transported between checkpoints.

Corpus and population. The saved preparation inventory contains 1,283 ordered, explicitly written singular/plural paradigms. Screening requires single-word lowercase alphabetic forms, distinct singular/plural strings and a declared regular written plural rule; prospective exclusions remove flagged variable, invariant, mass or otherwise unsuitable paradigms. Any form occurring in the declared lexical fields of P3, P4, P5 or accepted P8 is excluded. The first 128 eligible paradigms are paired consecutively into 64 groups and shuffled with seed 48119. The first 16 groups calibrate gains; the other 48 are reserved for evaluation. No model output, grammatical accuracy or intervention measurement selects the inventory.

The executed construction is FRONTED:

```
{context}next to the {attractor}, the {subject} probably
```

The contexts are `In the scene,` and `In this picture,` . Each group crosses both subject numbers, both attractor numbers and both contexts, yielding eight rows: 128 calibration and 384 evaluation rows. The prepared POST construction is unused in this experiment. Subject and attractor counterpart maps are closed, involutive and commuting. Both pinned tokenizers must preserve complete prefix length under each number swap; `is` and `are` must each append exactly one target token, and `probably` supplies the final anchor. Equal token counts do not equate token identities or semantic content. Structural and targeted lexical checks do not constitute an independent linguistic-panel review; the earlier 480-pair negation attestation does not cover this corpus. The prefixes are synthetic, not a frequency-balanced or naturalness-rated sample of English.

Common calibration rule. For one population, let B denote equally weighted lexical-group mean clean recipient-to-subject-donor JS, and let $E_w(g)$ denote the corresponding error after subject interchange along direction w. Recovery is $R_w(g) = 1 - E_w(g)/B$, a ratio of group means rather than a mean of rowwise ratios or normalization by each direction's best outcome. Each direction receives the coarse gains $\{0, .25, .5, .75, 1, 1.25, 1.5, 2, 3, 4\}$ for both targets $t \in \{0.50, 0.90\}$. The smallest coarse gain within absolute recovery tolerance 0.001 is accepted. Otherwise, the first adjacent ascending coarse bracket crossing t receives at most 12 actually evaluated midpoint steps, stopping at the first acceptable point. Both searches start from the same coarse grid; an exactly repeated gain reuses its saved measurement. There is no interpolation, extrapolation, global monotonicity assumption or attractor-based selection. A recorded coarse-grid minimum is a diagnostic, not a third P9 policy.

A clean gap $B \leq 10^{-8}$ is marked baseline-degenerate. An absent acceptable point or ascending bracket is unattainable under the declared search; budget exhaustion is precision-not-reached. These statuses and complete profiles are retained. Any missing fitted seed withholds its objective aggregate; missing U does not block CE–JS. All original operators and statuses are locked before evaluation prefix or suffix computation. Evaluation outcomes never retune gains.

Primary and interpretation gates. Let $\Delta_P(t)$ be mean-seed JS-fitted minus CE-fitted attractor-reference divergence at target t, with the unchanged clean recipient as reference. Positive values favor CE preservation. The sole primary is Qwen $H = \Delta_P(0.90) - \Delta_P(0.50)$ with a two-sided 95% interval. A change in this contrast need not be a ranking reversal. Stronger reversal interpretation requires complete arms, supported opposite signs under two 97.5% marginal gap intervals, and comparable-recovery gates at both targets. The expected pattern is $\Delta_P(0.50) > 0$ and $\Delta_P(0.90) < 0$; the reverse pattern is reported distinctly. Bonferroni provides nominal 95% joint sign coverage per checkpoint, conditional on approximate bootstrap coverage.

At each target, the paired 95% interval for JS-minus-CE achieved recovery must lie strictly within $[-0.02, 0.02]$, and each complete objective's 95% recovery lower bound must exceed zero. This gate concerns mean recovery, not equal recovery on every row or exact attainment of the nominal target. Absolute donor error, preservation, achieved recovery and displacement from the target remain re-

portable when matching fails. U comparisons have their own pairwise gates. Pythia replication and all component comparisons are secondary; they neither replace the Qwen primary nor pool checkpoint effects. Component, U-comparison and replication intervals are descriptive and unadjusted. The separate Bonferroni sign rule applies only to the two target-specific preservation gaps within one checkpoint, not jointly across checkpoints or recovery gates.

Measurements are computed separately for each seed, then averaged over seeds and the eight rows within each lexical group. Inference uses 20,000 shared whole-group bootstrap draws with seed 48123, preserving paired targets, contexts, number cells and checkpoint texts. Each draw recomputes the clean-gap denominator. If any denominator is at most 10^{-8}, normalized interval inference is withheld without discarding draws. Intervals condition on the inherited fits, locus, calibration groups, selected gains and lexicon; they do not propagate uncertainty from refitting or recalibration.

Crossed amplitude and signed axis. For each recipient h, attractor counterpart h_a, target and paired seed, the original edit for method $m \in \{C, J\}$ is $\delta_m = g_m w_m w_m^\top (h_a - h)$, with the parent engine's direction normalization. Write $a_m = \|\delta_m\|_2$ and $v_m = \delta_m / a_m$ when $a_m > 0$. The signed axis follows the actual edit; reversing a fitted vector's sign leaves its projector and edit unchanged. The two crossed conditions request $a_C v_J$ and $a_J v_C$. Across three paired seeds and two targets, these specify 12 potential evaluation conditions per checkpoint. Each uses the same recipient and unchanged clean reference. They are diagnostic hybrid edits, not new fitted methods or replacements with matched useful subject recovery.

Let F_{ab} be full-vocabulary disturbance with amplitude source a and signed-axis source b, so F_{CC} and F_{JJ} are the original methods. The symmetric factorial quantities are

$$A = \tfrac{1}{2}[(F_{JC} - F_{CC}) + (F_{JJ} - F_{CJ})], \tag{3}$$

$$S = \tfrac{1}{2}[(F_{CJ} - F_{CC}) + (F_{JJ} - F_{JC})], \tag{4}$$

$$I = F_{JJ} - F_{JC} - F_{CJ} + F_{CC}. \tag{5}$$

The algebraic identity $A + S = F_{JJ} - F_{CC}$ also holds for across-target changes, whose total is H. All four cells and I remain visible. The symmetric attribution splits interaction equally; it does not identify unique mediation. A includes gain and nuisance projection, while S includes signed orientation and downstream response. No local quadratic approximation, percentage explained or circuit identification is inferred.

A zero source amplitude gives a zero crossed edit regardless of axis. A positive source amplitude with exactly zero axis amplitude makes the entire crossed condition undefined; its row diagnostics are retained and the corresponding complete factorial aggregate is withheld. No row or seed is dropped. Finite near-zero projections are retained, with minimum-positive amplitudes recorded. Requested FP32 shifts and actual residual differences $(h+\delta)-h$ are archived separately. Requested amplitudes define the controlled factor; residual-addition rounding prevents a claim of exact applied-energy matching.

Execution and numerical acceptance. The design fixes sequential FP32 CPU runs on five threads, batch size 24, with FP32 full-vocabulary outputs and FP64 measurements. Sampled process working/private memory must remain within 12×10^9 bytes, available system memory above 128×10^6 bytes, and disk reserve at least 10×10^9 bytes. Sampling does not certify an unsampled hard bound.

Before fresh calibration, acceptance requires exposed-example checks of the cached suffix against ordinary native-hook forwards, original diagonal agreement, both crossed conditions, sign invariance and zero/self controls. Final ordinary checks score the first 32 evaluation rows for each executed original or crossed condition: four complete groups, both contexts and all number cells. The logit/target-odds tolerance is 0.001, and probability normalization tolerance is 10^{-12}; these sampled checks do not exhaust evaluation rows. Crossed signed-axis unit error and positive-amplitude relative norm error must be at most 2×10^{-6}. Ordinary-hook shift reconstruction permits 0.001 absolute discrepancy. Parent-diagonal agreement uses the declared numerical gates; bitwise equality is descriptive.

The saved-output analyzer authenticates source, calibration and output locks before and after analysis and independently replays scalar selection. It reconstructs all saved metrics in the probability

Table 17: Qwen3-0.6B-Base: all original procedures at both nominal recovery targets. Donor error E and attractor disturbance P are full-vocabulary JS in nats; $R = 1 - E/B$ and $R - t$ use the shared clean-gap denominator. Recovery intervals are 95%; CE/JS average measurements over all three seeds. These are original operators, not crossed controls. Unavailable means a required complete arm is missing; a retained point with "CI withheld" has invalid normalized interval inference.

Target t	Method	E	P	R [95% interval]	$R - t$ [95% interval]
0.50	U	0.21771	5.9582×10^{-5}	0.50903 [0.49044, 0.528]	0.009035 [−0.0095605, 0.027999]
0.50	CE	0.21941	4.6479×10^{-5}	0.50519 [0.48893, 0.52225]	0.0051903 [−0.01107, 0.022253]
0.50	JS	0.21889	5.0046×10^{-5}	0.50637 [0.49077, 0.52286]	0.0063654 [−0.0092337, 0.022863]
0.90	U	0.04389	0.00030426	0.90102 [0.8882, 0.91302]	0.0010219 [−0.011805, 0.013023]
0.90	CE	0.045602	0.00017834	0.89716 [0.88496, 0.90875]	−0.002839 [−0.015037, 0.0087531]
0.90	JS	0.042885	0.00018638	0.90329 [0.89143, 0.91434]	0.0032867 [−0.0085668, 0.014337]

domain with absolute tolerance 10^{-10}. Separate NumPy reconstruction checks all original requested norms and the crossed requested shifts and norms, using 2×10^{-5} absolute plus 2×10^{-5} relative tolerance for FP32 reduction-order differences. Replay of the archived crossed applied shifts requires exact FP32 array equality. A count-frequency calculation separately reconstructs group-summary estimates, ratios and intervals. The combined adapter authenticates each checkpoint through its own frozen verifier and uses unchanged statistical and raw-reconstruction code; acceptance also requires an exact saved-output replay. These checks have distinct scopes; passing them does not make a secondary result primary or establish a native causal concept. Acceptance receipts, including any failed attempts, remain part of the record.

Preserved implementation history. The original Pythia exposed preflight failed: the two crossed controls had maximum full-vocabulary logit discrepancies of approximately 1.214×10^{-3} and 1.136×10^{-3}, exceeding the fixed 10^{-3} tolerance. A separate diagnosis on the same 32 exposed rows isolated an FP32 operation-order discrepancy in the ordinary checker, which normalized the signed fitted vector instead of the requested shift. These algebraically equivalent forms produced slightly different FP32 edits; cached replay of the alternate edit reproduced the discrepancy. Independently recomputing the declared shift/norm operations from the native hook's recipient state restored exposed-check agreement within the original tolerance. A Pythia-only implementation amendment was frozen after the original Qwen execution completed and before fresh Pythia calibration or evaluation. It changes the ordinary checker, retains independent recipient-state projection, and leaves production edits, frozen directions, corpus, calibration, statistical rules and all numerical tolerances unchanged. The combined analysis retains Qwen's original frozen verifier and uses a separately bound Pythia verifier with unchanged statistical and raw-reconstruction code. The original failed attempt remains excluded from analysis and preserved by a complete file inventory. We also retain a preparation-harness failure caused by relocated corpus-test paths; an explicit workspace setting permits rerunning the unchanged tests. The amendment retains all prospective preflight, final-check and saved-output acceptance requirements.

O PROSPECTIVE RECOVERY-RESPONSE AND CROSSED-CONTROL RESULTS

The tables retain all declared methods, targets, pairwise gates, crossed cells and scalar selections. All intervals condition on fixed directions, loci, calibration groups and selected gains. Only Qwen's across-target preservation change H is primary. These are finite operating-point comparisons, not a continuous frontier or identification of a native causal variable.

Table 18: Qwen3-0.6B-Base: all six original-procedure pair contrasts. Positive preservation difference favors the subtracted procedure. Recovery differences have the same subtraction order. Each gate requires its own 95% recovery-difference interval strictly inside $[-0.02, 0.02]$ and positive 95% recovery lower bounds for both complete arms. These per-target comparisons are secondary; no failed gate removes the descriptive preservation contrast.

Target	Contrast	ΔP [95% interval], nats	ΔR [95% interval]	Gate
0.50	JS−CE	3.567×10^{-6} $[1.6428\times10^{-6}, 5.9064\times10^{-6}]$	0.0011751 $[-0.00079741, 0.0030208]$	Pass
0.50	U−CE	1.3103×10^{-5} $[6.3031\times10^{-6}, 2.1565\times10^{-5}]$	0.0038447 $[-0.0025347, 0.0099652]$	Pass
0.50	U−JS	9.5359×10^{-6} $[4.0538\times10^{-6}, 1.6059\times10^{-5}]$	0.0026695 $[-0.0035527, 0.0088396]$	Pass
0.90	JS−CE	8.0419×10^{-6} $[2.9784\times10^{-6}, 1.3775\times10^{-5}]$	0.0061257 $[0.003017, 0.0091676]$	Pass
0.90	U−CE	0.00012592 $[5.9242\times10^{-5}, 0.00021974]$	0.0038609 $[0.00025908, 0.0073342]$	Pass
0.90	U−JS	0.00011788 $[5.333\times10^{-5}, 0.00020996]$	-0.0022648 $[-0.0055866, 0.0011804]$	Pass

Table 19: Qwen3-0.6B-Base: all four actual factorial cells and their secondary symmetric decomposition, in nats with 95% intervals. The first index supplies requested amplitude and the second the signed axis. CC/JJ are original CE/JS operators; CJ/JC are diagnostic hybrid edits with no matched-subject-recovery claim. $A + S = F_{JJ} - F_{CC}$; I reports interaction separately and is not an additional additive term. Any missing required seed/control withholds the complete factorial aggregate.

Quantity	Target 0.50 [95% interval]	Target 0.90 [95% interval]
F_{CC}	4.6479×10^{-5} $[1.6457\times10^{-5}, 9.6989\times10^{-5}]$	0.00017834 $[6.3896\times10^{-5}, 0.00037021]$
F_{CJ}	4.3737×10^{-5} $[1.5899\times10^{-5}, 9.0419\times10^{-5}]$	0.0001682 $[6.1738\times10^{-5}, 0.00034651]$
F_{JC}	5.3133×10^{-5} $[1.9139\times10^{-5}, 0.00010973]$	0.00019735 $[7.2191\times10^{-5}, 0.00040467]$
F_{JJ}	5.0046×10^{-5} $[1.856\times10^{-5}, 0.00010242]$	0.00018638 $[6.9856\times10^{-5}, 0.00037916]$
A	6.4813×10^{-6} $[2.4715\times10^{-6}, 1.2442\times10^{-5}]$	1.8597×10^{-5} $[7.3092\times10^{-6}, 3.4197\times10^{-5}]$
S	-2.9143×10^{-6} $[-6.8409\times10^{-6}, -4.5566\times10^{-7}]$	-1.0555×10^{-5} $[-2.4508\times10^{-5}, -1.692\times10^{-6}]$
I	-3.4388×10^{-7} $[-8.0934\times10^{-7}, 1.5286\times10^{-8}]$	-8.2554×10^{-7} $[-1.9592\times10^{-6}, 1.8616\times10^{-7}]$
$F_{JJ} - F_{CC}$	3.567×10^{-6} $[1.6428\times10^{-6}, 5.9064\times10^{-6}]$	8.0419×10^{-6} $[2.9784\times10^{-6}, 1.3775\times10^{-5}]$

P CONSTRAINT STUDIES P10–P12 AND THEIR CORRECTIONS

These studies are reported as constraints on the fixed-vector result, not as additional headline successes. Each retains its own frozen endpoint. The 24 September corrections are additive: they preserve the original freezes, outputs and reports, correct interpretation or future analysis handling, and do not register new hypotheses.

P10: local-response prediction. This prospective Qwen study asked whether recipient-local output sensitivity predicts the P9 crossed amplitude/axis effects on 48 fresh lexical groups, with inherited fits and a fresh 16-group calibration to both targets. For each evaluation row and seed, forward-

Table 20: Qwen3-0.6B-Base: target-0.90 minus target-0.50 changes, in nats with 95% intervals. H is the difference of original JS-minus-CE preservation gaps; only Qwen H is primary. The remaining rows are recorded secondary factorial changes. The factorial total duplicates H only when all controls are available. Its absence does not hide an estimable original-method H. The interaction change is shown separately; no percentage explained or new cell-change inference is calculated.

Quantity	Change [95% interval]
H (primary)	4.475×10^{-6} $[5.8447\times10^{-7}, 8.8115\times10^{-6}]$
A	1.2116×10^{-5} $[4.7622\times10^{-6}, 2.1812\times10^{-5}]$
S	-7.6409×10^{-6} $[-1.7681\times10^{-5}, -1.2429\times10^{-6}]$
I	-4.8166×10^{-7} $[-1.1635\times10^{-6}, 1.7578\times10^{-7}]$
$F_{JJ} - F_{CC}$	4.475×10^{-6} $[5.8447\times10^{-7}, 8.8115\times10^{-6}]$
Amplitude on CE axis	1.2357×10^{-5} $[4.8495\times10^{-6}, 2.2343\times10^{-5}]$
Amplitude on JS axis	1.1875×10^{-5} $[4.6679\times10^{-6}, 2.1364\times10^{-5}]$
Axis at CE amplitude	-7.4001×10^{-6} $[-1.7153\times10^{-5}, -1.2225\times10^{-6}]$
Axis at JS amplitude	-7.8817×10^{-6} $[-1.8204\times10^{-5}, -1.2502\times10^{-6}]$

Table 21: Qwen3-0.6B-Base: all fourteen prospectively declared original operators, including unavailable searches. Gains and achieved calibration recovery are locked before evaluation; they do not imply the same recovery on evaluation. A missing selection is printed as unavailable, never as zero. "Unattainable" is restricted to the declared finite search; "precision not reached" denotes its midpoint budget. Every seed remains listed.

Direction	Target	Gain	Calibration recovery	Search status
U	0.5	0.69336	0.49908	Attained
U	0.9	1.5625	0.9	Attained
CE_2111	0.5	0.56641	0.50026	Attained
CE_2111	0.9	1.1094	0.89983	Attained
CE_2112	0.5	0.56445	0.49971	Attained
CE_2112	0.9	1.125	0.9007	Attained
CE_2113	0.5	0.56543	0.50026	Attained
CE_2113	0.9	1.1094	0.89901	Attained
JS_2111	0.5	0.57324	0.50015	Attained
JS_2111	0.9	1.0938	0.89987	Attained
JS_2112	0.5	0.57227	0.49934	Attained
JS_2112	0.9	1.1172	0.90056	Attained
JS_2113	0.5	0.57227	0.49952	Attained
JS_2113	0.9	1.1172	0.90037	Attained

mode differentiation gave the directional logit derivative $J_z v$ at the clean post-block-19 residual, and the local quadratic prediction $F(a, v) = a^2 \operatorname{Var}_p(J_z v)/8$ was formed for all four crossed cells; a matched-information isotropic comparator replaced the two axis curvatures by their mean. All predictions were hashed before any finite outcome was executed. The primary conjunction required a positive group-MAE improvement over the comparator and all eight mean conditional-contrast residual intervals (simultaneous 95%) strictly within $\pm10^{-6}$ nats. Improvement was established,

Table 22: Pythia-410m: all original procedures at both nominal recovery targets. Donor error E and attractor disturbance P are full-vocabulary JS in nats; $R = 1 - E/B$ and $R - t$ use the shared clean-gap denominator. Recovery intervals are 95%; CE/JS average measurements over all three seeds. These are original operators, not crossed controls. Unavailable means a required complete arm is missing; a retained point with "CI withheld" has invalid normalized interval inference.

Target t	Method	E	P	R [95% interval]	$R - t$ [95% interval]
0.50	U	0.1654	2.5427×10^{-5}	0.52253 [0.50504, 0.53937]	0.022533 [0.0050359, 0.039371]
0.50	CE	0.1698	1.177×10^{-5}	0.50981 [0.49415, 0.52485]	0.0098115 [−0.0058489, 0.024845]
0.50	JS	0.16948	1.7205×10^{-5}	0.51073 [0.49572, 0.52527]	0.010732 [−0.0042848, 0.025268]
0.90	U	0.029126	0.00010081	0.91592 [0.90701, 0.92427]	0.015919 [0.0070118, 0.024271]
0.90	CE	0.030235	4.0706×10^{-5}	0.91272 [0.90363, 0.92109]	0.012716 [0.0036289, 0.021094]
0.90	JS	0.029445	6.3673×10^{-5}	0.915 [0.90646, 0.9231]	0.014998 [0.0064591, 0.023104]

Table 23: Pythia-410m: all six original-procedure pair contrasts. Positive preservation difference favors the subtracted procedure. Recovery differences have the same subtraction order. Each gate requires its own 95% recovery-difference interval strictly inside $[-0.02, 0.02]$ and positive 95% recovery lower bounds for both complete arms. These per-target comparisons are secondary; no failed gate removes the descriptive preservation contrast.

Target	Contrast	ΔP [95% interval], nats	ΔR [95% interval]	Gate
0.50	JS−CE	5.4348×10^{-6} $[3.4576 \times 10^{-6}, 7.7456 \times 10^{-6}]$	0.00092089 $[-0.0029636, 0.0046852]$	Pass
0.50	U−CE	1.3657×10^{-5} $[9.5141 \times 10^{-6}, 1.8536 \times 10^{-5}]$	0.012721 $[0.008091, 0.01719]$	Pass
0.50	U−JS	8.2225×10^{-6} $[4.6609 \times 10^{-6}, 1.1945 \times 10^{-5}]$	0.011801 $[0.0058844, 0.017706]$	Pass
0.90	JS−CE	2.2967×10^{-5} $[1.5002 \times 10^{-5}, 3.2518 \times 10^{-5}]$	0.0022816 $[-0.00063452, 0.0051098]$	Pass
0.90	U−CE	6.0108×10^{-5} $[4.2173 \times 10^{-5}, 8.1761 \times 10^{-5}]$	0.0032028 $[-0.00047166, 0.0067251]$	Pass
0.90	U−JS	3.7141×10^{-5} $[2.2223 \times 10^{-5}, 5.3084 \times 10^{-5}]$	0.00092115 $[-0.001576, 0.0032777]$	Pass

0.53 [0.24, 0.97]$\times 10^{-6}$ nats, but no residual interval lay inside the margin (estimates –10.0 to 4.1×10^{-6} nats), so the primary is inconclusive: prediction improvement is not adequacy. Independent reconstruction passed 172,800 full-JS and readout comparisons. The 24 September erratum corrects the post hoc exploration only: high-response bins are far from the JS ceiling (maximum finite response 9.4% of $\log 2$), a ratio direction was misstated, tied Spearman ranks now use average ranks (six exploratory correlation fields change), and anisotropy is associated with subject and nearest-noun number, which coincide in FRONTED. No registered quantity changes.

P11: stratified signed-axis response. P11 reanalyzed the retained P9 evaluation outcomes of both checkpoints under a declaration made after those outcomes existed, with newly computed local curvatures; it is not fresh prospective confirmation. Each of eight cells (target $\times$ amplitude source $\times$ subject-number stratum) required agreement between a Bonferroni interval on the equal-weight group mean and an exact interval on the proportion of groups with the predicted sign. In Qwen, H1 (signed-axis reversal by subject number) and H2 (local curvature anisotropy by subject number) are established; for example, at target 0.90 with CE amplitude the singular-subject axis contrast is 5.18 [2.23, 9.36] and the plural-subject contrast –25.46 [–70.23, –5.18]$\times 10^{-6}$ nats. Plural-subject

Table 24: Pythia-410m: all four actual factorial cells and their secondary symmetric decomposition, in nats with 95% intervals. The first index supplies requested amplitude and the second the signed axis. CC/JJ are original CE/JS operators; CJ/JC are diagnostic hybrid edits with no matched-subject-recovery claim. $A+S = F_{JJ}-F_{CC}$; I reports interaction separately and is not an additional additive term. Any missing required seed/control withholds the complete factorial aggregate.

Quantity	Target 0.50 [95% interval]	Target 0.90 [95% interval]
F_{CC}	1.177×10^{-5} $[7.9589\times10^{-6}, 1.7051\times10^{-5}]$	4.0706×10^{-5} $[2.766\times10^{-5}, 5.8667\times10^{-5}]$
F_{CJ}	1.1767×10^{-5} $[7.9993\times10^{-6}, 1.6944\times10^{-5}]$	4.0716×10^{-5} $[2.7774\times10^{-5}, 5.8437\times10^{-5}]$
F_{JC}	1.7186×10^{-5} $[1.1948\times10^{-5}, 2.4483\times10^{-5}]$	6.3609×10^{-5} $[4.4346\times10^{-5}, 9.0247\times10^{-5}]$
F_{JJ}	1.7205×10^{-5} $[1.2026\times10^{-5}, 2.4397\times10^{-5}]$	6.3673×10^{-5} $[4.4603\times10^{-5}, 8.9955\times10^{-5}]$
A	5.427×10^{-6} $[3.4121\times10^{-6}, 7.8127\times10^{-6}]$	2.293×10^{-5} $[1.4837\times10^{-5}, 3.2756\times10^{-5}]$
S	7.8126×10^{-9} $[-1.4649\times10^{-7}, 1.5705\times10^{-7}]$	3.6487×10^{-8} $[-4.8725\times10^{-7}, 5.6025\times10^{-7}]$
I	2.1379×10^{-8} $[-3.7577\times10^{-8}, 7.7353\times10^{-8}]$	5.4723×10^{-8} $[-1.8204\times10^{-7}, 2.774\times10^{-7}]$
$F_{JJ}-F_{CC}$	5.4348×10^{-6} $[3.4576\times10^{-6}, 7.7456\times10^{-6}]$	2.2967×10^{-5} $[1.5002\times10^{-5}, 3.2518\times10^{-5}]$

Table 25: Pythia-410m: target-0.90 minus target-0.50 changes, in nats with 95% intervals. H is the difference of original JS-minus-CE preservation gaps; only Qwen H is primary. The remaining rows are recorded secondary factorial changes. The factorial total duplicates H only when all controls are available. Its absence does not hide an estimable original-method H. The interaction change is shown separately; no percentage explained or new cell-change inference is calculated.

Quantity	Change [95% interval]
H (secondary)	1.7532×10^{-5} $[1.1528\times10^{-5}, 2.48\times10^{-5}]$
A	1.7503×10^{-5} $[1.1409\times10^{-5}, 2.4939\times10^{-5}]$
S	2.8674×10^{-8} $[-3.4168\times10^{-7}, 4.0397\times10^{-7}]$
I	3.3344×10^{-8} $[-1.4466\times10^{-7}, 2.0164\times10^{-7}]$
$F_{JJ}-F_{CC}$	1.7532×10^{-5} $[1.1528\times10^{-5}, 2.48\times10^{-5}]$
Amplitude on CE axis	1.7487×10^{-5} $[1.1367\times10^{-5}, 2.4967\times10^{-5}]$
Amplitude on JS axis	1.752×10^{-5} $[1.1456\times10^{-5}, 2.4919\times10^{-5}]$
Axis at CE amplitude	1.2002×10^{-8} $[-3.062\times10^{-7}, 3.396\times10^{-7}]$
Axis at JS amplitude	4.5346×10^{-8} $[-3.9203\times10^{-7}, 4.7823\times10^{-7}]$

rows carry 92.1% of all finite divergence, and every aggregate Qwen axis contrast is the sum of opposing established strata. In Pythia, H1 and H2 are unresolved. Subject number and most-recent-noun number coincide in FRONTED, so the strata do not separate grammatical role from recency. Independent reconstruction passed 18,432 full-JS comparisons from the retained P9 logits. The

Table 26: Pythia-410m: all fourteen prospectively declared original operators, including unavailable searches. Gains and achieved calibration recovery are locked before evaluation; they do not imply the same recovery on evaluation. A missing selection is printed as unavailable, never as zero. "Unattainable" is restricted to the declared finite search; "precision not reached" denotes its midpoint budget. Every seed remains listed.

Direction	Target	Gain	Calibration recovery	Search status
U	0.5	0.61523	0.49953	Attained
U	0.9	1.2188	0.89943	Attained
CE_2111	0.5	0.52246	0.50012	Attained
CE_2111	0.9	0.96875	0.89913	Attained
CE_2112	0.5	0.52344	0.50075	Attained
CE_2112	0.9	0.97656	0.90074	Attained
CE_2113	0.5	0.52246	0.50026	Attained
CE_2113	0.9	0.97656	0.90097	Attained
JS_2111	0.5	0.54688	0.49949	Attained
JS_2111	0.9	1.0469	0.90036	Attained
JS_2112	0.5	0.54297	0.49938	Attained
JS_2112	0.9	1.0469	0.90085	Attained
JS_2113	0.5	0.54688	0.49926	Attained
JS_2113	0.9	1.0547	0.89993	Attained

Table 27: Original JS-minus-CE preservation gaps with the two 97.5% marginal sign intervals used for reversal interpretation, in nats. Bonferroni gives nominal 95% joint sign coverage per checkpoint; bootstrap coverage is approximate. The P8-motivated pattern is positive at target 0.50 and negative at 0.90. Opposite signs alone do not establish comparable recovery; both gates must also pass. Pythia remains secondary.

Checkpoint	Target	ΔP [97.5% interval]	Recovery gate
Qwen3-0.6B-Base	0.50	3.567×10^{-6} $[1.4283\times10^{-6}, 6.283\times10^{-6}]$	Pass
Qwen3-0.6B-Base	0.90	8.0419×10^{-6} $[2.292\times10^{-6}, 1.4694\times10^{-5}]$	Pass
Pythia-410m	0.50	5.4348×10^{-6} $[3.2077\times10^{-6}, 8.1306\times10^{-6}]$	Pass
Pythia-410m	0.90	2.2967×10^{-5} $[1.4093\times10^{-5}, 3.4065\times10^{-5}]$	Pass

correction adds a fail-closed reporter that refuses to present classifications when a common gate fails; the accepted report regenerates byte for byte.

P12: independent refit. P12 refitted CE and JS on FRONTED with a disjoint lexicon, recalibrated to both targets and evaluated 48 new groups. This changed fitting construction, fitting cohort, wording and evaluation population together. Family A (the two target-specific JS-minus-CE gaps and their across-target change H) uses Bonferroni-6 intervals; all three cells are unresolved in each checkpoint, and the family label is partial. Qwen H is 11.61 [–2.30, 25.79]$\times10^{-6}$ nats and Pythia H –0.11 [–4.50, 3.02]$\times10^{-6}$ nats. Read at the P9 standard of one 95% interval, which is descriptive and not a P12 registered quantity, Qwen's H was again positive, $+11.6$ $[+1.3, +21.9]\times10^{-6}$ nats, but no target-specific gap excluded zero: Qwen's higher-target gap was $+11.2$ $[-1.8, +23.5]\times10^{-6}$ nats, and its lower-target gap and all Pythia gaps were below 10^{-6} nats in magnitude.

The P12 correction adopts the stricter reading that its local-prediction family D requires numerical gates on both arms. Pythia's POST_OBJECT arm exceeded its 10^{-7} nat budget (1.086×10^{-7}), so all six Pythia D versions are unavailable confirmatorily; their numbers remain descriptive. In Qwen, three D versions are supported under their own criteria and three are inconclusive; a supported version does not substitute for another or yield a study-wide claim. The frozen metadata and executed

Table 28: Recorded reversal interpretation flags, without new tests. Expected replication requires the positive-low/negative-high sign pattern and both recovery gates. An opposite-pattern matched reversal is displayed separately. A missing primary arm remains unavailable rather than being interpreted as evidence of no reversal.

Checkpoint	Sign pattern	Both recovery gates pass	Expected matched reversal	Any matched reversal
Qwen3-0.6B-Base	Not established	Yes	No	No
Pythia-410m	Not established	Yes	No	No

Table 29: Qwen3-0.6B-Base: conditional crossed-control effects in units of 10^{-6} nats, with paired 95% intervals. Amplitude changes replace CE-sourced by JS-sourced magnitude while holding the signed axis fixed; axis changes replace CE by JS while holding the amplitude source fixed. Negative values reduce disturbance. These are secondary, unadjusted finite-response comparisons. The hybrid controls have no matched-subject-recovery interpretation.

Controlled change	Target 0.50	Target 0.90
Amplitude on CE axis	+6.6532 [+2.5046, +12.8143]	+19.0099 [+7.3966, +35.1205]
Amplitude on JS axis	+6.3093 [+2.4286, +12.0817]	+18.1844 [+7.1337, +33.2967]
Signed axis at CE amplitude	-2.7424 [-6.4427, -0.4376]	-10.1424 [-23.5866, -1.6530]
Signed axis at JS amplitude	-3.0862 [-7.2391, -0.4669]	-10.9680 [-25.4427, -1.7175]

code disagreed on two multiplicity divisors and a per-stratum equivalence bound; the executed constants are primary and a metadata sensitivity leaves every top-level A/B/C outcome unchanged. A diagnostic labeled as training CE used recipient labels; the fitting objective itself used donor labels correctly, and no fit is changed. Family C now requires derivative acceptance for its compound decision; both accepted studies passed. Family A numbers are unchanged by every correction.

Q CONTROLLED REFITTING (P14) AND CALIBRATION-POLICY SENSITIVITY (P14-S)

Design. Two fixed 32-group fitting cohorts, each with prior study exposure, are rendered in both FRONTED and POST_OBJECT with eight context/number cells per group. In each of the four cohort-by-construction cells, CE and donor-JS directions are fitted with three paired seeds (14111–14113), four epochs, 44 updates and identical optimizer settings; the final parameter is retained without checkpoint selection. Each cell's subject-displacement mean initializes its directions, so initialization follows one rule but is not one numerical vector across constructions. All 56 operators per checkpoint (CE, JS and U, per cell and target) are calibrated on one exposed, disjoint 16-group FRONTED population with the P9 selector (coarse gains $\{0, .25, .5, .75, 1, 1.25, 1.5, 2, 3, 4\}$; the smallest accepted coarse gain, else at most 12 midpoint steps in the first ascending bracket), then evaluated on 48 FRONTED groups whose forms are disjoint from the recorded P1–P13 lexical union. Both checkpoints use post-block 19. Inference resamples evaluation groups jointly across all cells (100,000 draws); seeds and fitting cohorts are not treated as independent sampled observations. The primary $K(t)$ intervals are adjusted over the two targets within each checkpoint (97.5% marginal); cohort-specific and paired policy-difference intervals are descriptive 95% intervals. The held-out recovery family comprises 32 simultaneous checks across both targets: within-cohort objective and construction recovery differences inside ± 0.02 and positive recovery lower bounds. It does not compare recovery across fitting cohorts or across calibration policies.

Figure 12 in the main text is generated by `make_revision_figures.py` from the authenticated P14-S analyses, which contain both the original and the alternative replays.

Table 30: Pythia-410m: conditional crossed-control effects in units of 10^{-6} nats, with paired 95% intervals. Amplitude changes replace CE-sourced by JS-sourced magnitude while holding the signed axis fixed; axis changes replace CE by JS while holding the amplitude source fixed. Negative values reduce disturbance. These are secondary, unadjusted finite-response comparisons. The hybrid controls have no matched-subject-recovery interpretation.

Controlled change	Target 0.50	Target 0.90
Amplitude on CE axis	+5.4163 [+3.3947, +7.8101]	+22.9030 [+14.7700, +32.7931]
Amplitude on JS axis	+5.4376 [+3.4291, +7.8162]	+22.9577 [+14.8832, +32.7297]
Signed axis at CE amplitude	-0.0029 [-0.1425, +0.1339]	+0.0091 [-0.4452, +0.4745]
Signed axis at JS amplitude	+0.0185 [-0.1541, +0.1836]	+0.0638 [-0.5471, +0.6612]

Table 31: Cohort-specific fitting-construction contrasts, 10^{-6} nats, with descriptive unadjusted 95% intervals. K averages the two columns. No inference about a population of fitting cohorts follows.

Checkpoint	Policy	*t*	Cohort A	Cohort B
Pythia-410m	Original	.50	–4.96 [–11.71, 1.32]	–3.01 [–7.43, 0.59]
	Original	.90	–6.78 [–24.55, 12.13]	–1.59 [–9.84, 6.63]
	Alternative	.50	–4.96 [–11.71, 1.32]	–3.01 [–7.43, 0.59]
	Alternative	.90	–6.78 [–24.55, 12.13]	–1.59 [–9.84, 6.63]
Qwen3-0.6B	Original	.50	–2718.18 [–4859.33, –1081.31]	1.67 [–0.81, 4.35]
	Original	.90	–485.67 [–927.09, –168.61]	193.89 [81.31, 338.22]
	Alternative	.50	5.65 [2.22, 9.49]	1.67 [–0.81, 4.35]
	Alternative	.90	–92.34 [–193.62, –21.00]	–179.81 [–357.83, –58.65]

Why a sensitivity analysis, and what it changes. Five of the 112 original selections accept a later coarse gain although an earlier sampled ascending crossing exists. The original rule specifies exactly this order, so it is not a coding error. After the original outcomes were known, P14-S froze an alternative that prioritizes the first sampled ascending crossing, replayed all 112 selections, and executed only the new measurements; every fit, group and seed was retained, and no direction was refitted. The changed selections are Pythia cohort-A FRONTED U at 90% (gain 2 to 1.0625), Qwen cohort-A FRONTED JS seeds 14111 and 14113 at 50% (3 to 0.5459), Qwen cohort-A FRONTED JS seed 14112 at 90% (2 to 1.1875), and Qwen cohort-B FRONTED CE seed 14111 at 90% (2 to 1.21875). Pythia's learned-operator K and intervals are therefore unchanged; its changed U baseline moves the descriptive U attractor disturbance from 2512.32 to 685.56$\times 10^{-6}$ nats at held-out subject recovery 0.899 and 0.893. Qwen's cohort-A POST_OBJECT U at 90% is unavailable under both selectors and does not withhold the CE/JS primary. Paired alternative-minus-original K is 1361.91 [542.33, 2434.12] at 50% and 9.82 [–4.73, 27.77]$\times 10^{-6}$ nats at 90% for Qwen, and exactly zero for Pythia. The 90% interval crossing zero is not a policy-equivalence finding.

Interpretation limits. K is a difference of JS-minus-CE disturbances across fitting constructions, not an overall ranking of objectives or models. The construction treatment jointly changes syntax, word order, punctuation, anchor position and initialization; it does not identify grammatical role or construction dominance. Inference conditions on two particular fitting cohorts and one calibration sample, so it does not estimate variation over a population of fitting cohorts. The alternative selector is outcome informed; its intervals are descriptive. Full original and alternative reports retain absolute disturbances, all 56 source calibrations per checkpoint, U baselines, signed behavioral descriptives and recovery checks.

Verification. Independent reconstruction of the original runs passed 112,768 raw JS comparisons in Pythia and 102,144 in Qwen, with 56 calibration-selector replays per checkpoint and exact statistical replay. P14-S verification reconstructed its new raw endpoints, rehashed the complete parent

payload, replayed all alternate selectors and sparse replacement tensors, and replayed the statistics exactly; it relied on the parent's earlier raw reconstruction for retained parent endpoints. An initial private P14-S launcher stopped before any model execution on a JSON-serialization identity check; an additive reviewed launcher completed both runs without changing frozen scientific files. Local compact replays of all four analyses reproduce the statistics exactly without model weights.

R CONTROLLER MECHANISM (P13) AND LARGER CHECKPOINTS (P13-X)

Design. For group k and construction X, the subject flip S_k^X is the mean anchor difference between plural and singular subjects, and the attractor flip A_k^X the corresponding difference for attractors. The controller $\hat{w} = \mathrm{unit}(\bar{S}^F + \bar{S}^P)$ is computed from the 48 exposed P12 study groups at each checkpoint and locus and locked before any evaluation row enters a model stage. Subject displacement $s_k^X = S_k^X \cdot \hat{w}$ and leakage displacement $\lambda_k^X = A_k^X \cdot \hat{w}$ define leakage $L^X = \bar{\lambda}^X / \bar{s}^X$. If the readout depended on the anchor only through $\hat{w}$ with one slope, attraction divided by the subject swing would equal leakage divided by subject displacement; ρ_{B2} is the ratio of these two sides and has no fitted parameter. Causal edits replace the anchor of a mismatched POST_OBJECT row by its matched counterpart in full (FULL), along $\hat{w}$ only (CTRL) or in the orthogonal complement (ORTH); restoration fractions are ratios of group means. C3 adds the POST_OBJECT leakage along $\hat{w}$ to FRONTED rows and compares the induced attraction with the linear prediction. Every construction has eight number/context rows per group: FRONTED ("next to the *attractor*, the *subject* probably"), POST_OBJECT ("the *subject* next to the *attractor* probably") and POST_RELATIVE ("the *subject* that sat next to the *attractor* probably"). Inference uses 100,000 paired whole-group draws with Bonferroni control within R (3 cells), B (4) and C (4); B2 and C require established positive attraction; all nested point and bootstrap denominators must exceed 10^{-8} in magnitude. Table 32 lists every cell.

Table 32: All P13 and P13-X R/B/C cells: estimate, family-adjusted interval (R: two-sided $\alpha = .05/3$; B and C: $.05/4$) and frozen status. R1 cosine difference; R2a POST_OBJECT leakage; R2b leakage ratio (equivalence ± 0.1); B1 attraction difference (log-odds); B2 parameter-free ratio (accurate [0.8,1.25], same order [0.5,2]); B3/B4 Spearman correlations; C1 ϕ_{CTRL} (threshold 0.5); C2 $\phi_{\mathrm{CTRL}} - \phi_{\mathrm{ORTH}}$; C2b ϕ_{ORTH} (equivalence ± 0.2); C3 transplant ratio (bands as B2). Two-rule sign cells (R2a, B1, C2) also require the exact group-sign interval; the mean interval is shown. The original Pythia-410m cohort failed its numerical gate and has only R partial. est. = established; same = same order; acc. = accurate; n.s. = not supported; unres. = unresolved; contr. = contradicted.

Cell	Qwen3-0.6B, P13 cohort	Pythia-410m, confirmation	Qwen3-0.6B, matched
R1	0.55 [0.52, 0.60] est.	0.25 [0.23, 0.33] est.	0.53 [0.50, 0.58] est.
R2a	0.07 [0.07, 0.08] est.	0.19 [0.17, 0.22] est.	0.09 [0.07, 0.10] est.
R2b	0.21 [0.06, 0.35] unres.	0.08 [0.01, 0.15] unres.	0.13 [–0.00, 0.25] unres.
B1	0.46 [0.36, 0.55] est.	1.10 [0.97, 1.22] est.	0.61 [0.50, 0.72] est.
B2	0.80 [0.74, 0.88] same	0.80 [0.73, 0.88] same	0.75 [0.69, 0.81] same
B3	0.53 [0.16, 0.79] est.	0.47 [0.12, 0.73] est.	0.85 [0.65, 0.94] est.
B4	0.78 [0.55, 0.89] est.	0.74 [0.51, 0.88] est.	0.82 [0.62, 0.92] est.
C1	0.85 [0.78, 0.92] est.	1.07 [0.98, 1.17] est.	0.89 [0.83, 0.96] est.
C2	1.04 [0.91, 1.17] est.	1.20 [1.02, 1.38] est.	1.10 [0.99, 1.21] est.
C2b	–0.19 [–0.26, –0.12] unres.	–0.13 [–0.21, –0.04] unres.	–0.21 [–0.26, –0.16] unres.
C3	0.58 [0.56, 0.61] same	0.84 [0.82, 0.86] acc.	0.59 [0.57, 0.62] same
Cell	Qwen3-1.7B, primary	Pythia-1.4B b14, primary	Pythia-1.4B b19, sensitivity
R1	0.38 [0.36, 0.45] est.	0.17 [0.14, 0.24] est.	0.41 [0.39, 0.45] est.
R2a	0.15 [0.14, 0.17] est.	0.35 [0.30, 0.40] est.	0.15 [0.13, 0.17] est.
R2b	0.39 [0.30, 0.48] contr.	0.07 [0.04, 0.10] unres.	0.20 [0.14, 0.28] contr.
B1	0.82 [0.70, 0.94] est.	0.96 [0.84, 1.09] est.	0.96 [0.84, 1.09] est.
B2	0.79 [0.70, 0.88] same	0.34 [0.32, 0.37] n.s.	0.81 [0.75, 0.86] same
B3	0.32 [–0.01, 0.61] unres.	0.68 [0.40, 0.86] est.	0.80 [0.61, 0.90] est.
B4	0.50 [0.19, 0.72] est.	0.65 [0.35, 0.84] est.	0.83 [0.60, 0.94] est.
C1	0.64 [0.59, 0.70] est.	0.76 [0.71, 0.82] est.	1.17 [1.10, 1.24] est.
C2	0.82 [0.70, 0.94] est.	1.06 [0.96, 1.17] est.	1.53 [1.40, 1.67] est.
C2b	–0.18 [–0.25, –0.11] unres.	–0.30 [–0.37, –0.24] contr.	–0.36 [–0.44, –0.30] contr.
C3	0.42 [0.39, 0.44] n.s.	0.34 [0.32, 0.37] n.s.	0.96 [0.93, 0.99] acc.

What the side-effect endpoint measures. The original P13 family S applied the fitted operators' attractor-donor edits, at their locked 90% gains, to fresh rows. In the Qwen cohort and in the Pythia

confirmation, all 12 POST_OBJECT cells were established in the predicted direction: the edits reduced correct-verb log-odds in number-matched rows (introducing attraction) and increased them in mismatched rows (removing it), for U, CE and JS from both the inherited and the P12 fits. In FRONTED, 6 and 4 of 12 cells were established, respectively. This measures directly that the unsigned preservation endpoint pools introduced and removed attraction. P13-X omits family S by design.

Populations, conditions and gates. The original P13 study used 64 fresh groups for Qwen3-0.6B and Pythia-410m at post-block 19. The original Pythia run failed an independent ordinary-forward numerical check, so its B, C and S families are withheld. A separate Pythia-410m confirmation used the next 128 nouns of the registered shuffle with the numerical comparison corrected before any fresh group ran. P13-X reuses this confirmation lexicon (16 development and 48 evaluation groups) for four prospectively frozen conditions on CPU in FP32: Qwen3-1.7B-Base at post-block 19 (primary), Pythia-1.4B-deduped at post-block 14 (selected primary), Pythia-1.4B-deduped at post-block 19 (descriptive locus sensitivity) and Qwen3-0.6B-Base at post-block 19 (matched comparator). Each larger checkpoint's locus was chosen by an independently recorded selector before development. Each condition ran a 16-group development attempt, independent verification, a per-condition study freeze that must bind a development passing the identity, ordinary-forward numerical and edit-realization gates, the 48-group study and independent study verification. Ordinary-forward budgets are $\max(10^{-4}\text{ nats}, 10^{-5}|G|)$ for the readout and $\max(10^{-9}\text{ nats}, 10^{-3}\,\text{JS})$ for full-vocabulary JS, evaluated inside the original 24-row prefix partitions. JVP acceptance passed for every condition and withholds only derivative-dependent descriptions when it fails.

The selected Pythia primary. The original block-14 development attempt failed the unchanged numerical budget for POST_RELATIVE FULL (readout) and CTRL (readout and JS); maximum readout discrepancies were 1.14×10^{-4} and 1.74×10^{-4} nats. It remains a failure and supplies no result. A diagnostic on the exposed development data reproduced the failure and showed that aligning suffix batches with the original prefix partitions, with the unchanged anchor-only head, restores agreement at unchanged budgets. This amendment was independently reviewed and frozen as a separate 97-file design after other conditions' outcomes were known but before any primary Pythia evaluation. Its complete new development passed every gate, with zero readout discrepancy and at most 5.3×10^{-12} nats JS discrepancy, and only then was its study frozen and executed. The block-19 sensitivity cannot replace this primary.

Verification and replay. Independent verification of every P13-X study reconstructed all 12,096 retained readout and full-vocabulary JS comparisons (maximum discrepancy below 5×10^{-16} nats), all requested causal edits for the 17 causal conditions and the key estimates, and replayed the frozen analysis exactly. A single report from the unchanged reporter authenticates all four conditions together. Local compact replay without model weights authenticates every archive and admission chain, restores six original layer-anchor arrays per condition, and reproduces every family and cell classification; a few R1 bootstrap interval endpoints differ from the saved server values by 1–3 units in the last place (at most 1.7×10^{-16}), and every difference is recorded. Reported values are the authenticated saved server values.

Numerical-criterion history. The historical P13 protocol prose stated a norm-relative edit-realization tolerance of 10^{-5}, while the frozen specifications and executed code enforced an elementwise FP32 rounding bound. All retained historical attempts pass the implemented elementwise criterion; none passes the prose criterion over all rows, with maximum norm-relative errors of 0.0096, 0.0248 and 0.0090 in the original Pythia, Qwen and confirmation studies, mostly for very small FRONTED edits. Historical results are therefore reported under the implemented specification with this qualification. P13-X states the elementwise criterion as its single operational gate and records norm-relative exceedances as diagnostics (12, 0, 9 and 8 condition rows for the four conditions in the order above). The corrected v2 analysis adds denominator and eligibility guards; it reproduces every retained numeric value and classification exactly.

Interpretation limits. C1 tests restoration above 50%, not complete removal or a uniquely necessary native mechanism. FULL replaces one final anchor while earlier cached positions stay recipient-specific, so residual attraction does not locate where attraction first arises. B2 relates quantities

measured on the same evaluation cohort and is not a source-frozen prediction. The Qwen size comparison shares cohort, protocol and absolute block index but not relative depth, width or training, and separate intervals are not a paired test. A shared-primary claim requires the cell supported in both larger primaries. P13-X and P15 share lexical groups and are not independent lexical replications.

S SOURCE-FROZEN RESPONSE TRANSFER (P15)

Design and chronology. The P15 specification was frozen before any new source or destination forward. In each checkpoint, FRONTED and POST_OBJECT anchor states of the 48 exposed P12 source groups were re-extracted on the execution host; the plural-oriented controller $\hat{w}$ is the normalized sum of their mean subject-number differences, and five random unit directions orthogonal to it were drawn and locked with it before source interventions. Source FRONTED subject-counterpart edits use dose $d = \hat{w}^\top(h_{\text{donor}} - h_{\text{recipient}})$ and displacement $d\hat{w}$ at the final anchor's zero-based post-block-19 residual. With $G = \log p(\texttt{are}) - \log p(\texttt{is})$, the zero-intercept least-squares slope is $\beta = \langle d\,\Delta G\rangle / \langle d^2\rangle$, with means over the 384 source rows. Controller and slope use the same exposed source groups, so they are not split-sample estimates. The slope and the source response scale (mean absolute natural subject-counterpart change in G) were locked before destination prefixes were evaluated; all destination predictions βd were locked before the first edited destination forward. Pythia's slope is 1.187 with scale 12.07 nats; Qwen's are 0.248 and 11.35 nats.

The destination is the 48-group confirmation lexicon in OBJECT_RELATIVE ("{context}the *subject* that the *attractor* could obscure probably") and FRONTED, each with two contexts and four number cells (384 rows). In OBJECT_RELATIVE the matrix subject is the relative clause's object gap and the attractor is the embedded subject; the modal prevents a second number-inflected verb. Plausibility, frequency, length and processing difficulty are not matched, and the construction changes syntax, role, order and position together. A separate template review and tokenizer-only preparation checks preceded execution. The normalized MAE is the equal-group mean of within-group mean absolute prediction error divided by the fixed source scale; the zero-response comparator predicts no change, and improvement is its paired MAE minus the predictor's on the same rows. Support requires the 98.75% upper bound of the OBJECT_RELATIVE normalized MAE to be at most 0.10, the 98.75% lower bound of improvement to exceed zero, and eight gates: authentication, clean competence (group accuracy lower 95% bound above 0.5), positive FULL control, exact NOOP identity, ordinary-versus-cached numerical agreement, prediction-lock ordering, FP32 edit realization and source-parameter floors. All gates pass in both arms of both checkpoints.

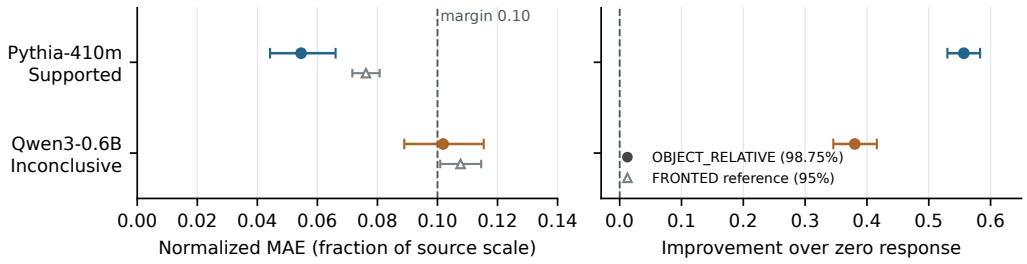


Figure 15: Source-frozen response prediction. Left: normalized MAE on OBJECT_RELATIVE (98.75% four-component adjusted intervals) with the FRONTED same-lexicon reference (descriptive 95%); the dashed line is the prespecified 0.10 margin. Right: paired improvement over the zero-response predictor (98.75%). Intervals condition on the fixed source fit. Generated by `make_revision_figures.py` from the authenticated P15 analyses.

Reference, controls and scope. Source-training normalized MAE is 0.083 (Pythia) and 0.110 (Qwen); these in-sample values are not held-out performance. Qwen's FRONTED reference error 0.108 [0.101, 0.115] has its descriptive lower bound above the margin, a descriptive status confined to that reference arm; it does not change the OBJECT_RELATIVE primary. Pythia's paired OBJECT_RELATIVE-minus-FRONTED error is –0.022 [–0.030, –0.012], which does not isolate syntax as its cause; Qwen's spans zero, which establishes neither equivalence nor a construction-

specific failure. Clean conditional accuracy is 1.0 in both arms of both checkpoints. OBJECT_RELATIVE CTRL donor-directed effects are 7.38 [7.07, 7.70] (Pythia) and 5.47 [5.25, 5.70] nats (Qwen), FULL controls are positive, and NOOP readouts equal clean readouts exactly. The five random directions use exactly the controller's signed target-conditioned dose, so they match its edit norm; their OBJECT_RELATIVE effects lie between -0.17 and 0.19 nats in Pythia and between -0.10 and 0.30 nats in Qwen. They are descriptive controls, not an additional confirmatory family. No destination dose fell outside the source dose range. Destination doses use each row's clean subject counterpart, so the test is target-conditioned interchange without destination parameter fitting, not a fixed-dose deployment. Each checkpoint has its own controller and slope; no cross-model representation transport, isolated syntax effect or population claim follows.

Verification. For each checkpoint, independent server verification reconstructed complete full-vocabulary readouts for all 20 conditions, the source geometry and random directions, the source OLS slope by an independent formula, every requested edit and numerical gate, the lock chronology and the statistical analysis, which replayed exactly; model forwards were not independently rerun. Local compact replay authenticated both archives and replayed source geometry, doses, edit requests, predictions, realization diagnostics and the saved analyses. The frozen reporter checked a receipt's PASS status without binding the current analysis bytes; an additive reporter that authenticates the complete chain produced the reported tables, and the frozen experiment was not re-frozen.

T VERIFICATION SCOPE, PRESERVED FAILURES AND AMENDMENTS

Numerical acceptance is separate from scientific outcome. Original failure statuses and validation-only amendments remain in the record; a successful audit neither promotes a secondary endpoint nor resolves an unavailable primary. Saved-output reconstruction recomputes readouts and full-vocabulary divergences from retained arrays; it does not rerun model forwards. Compact replay re-runs frozen statistical analyses from included small records and selected arrays; it cannot reconstruct omitted full-vocabulary logits. Table 33 summarizes P9–P15; earlier studies are described in Appendix M.

Table 33: Verification scope and preserved histories for the later studies. Counts are independent full-vocabulary JS/readout comparisons unless stated. "Exact" means byte- or value-identical replay of the frozen analysis.

Study	Reconstruction and replay	Preserved failures, amendments and corrections
P9	Qwen 860,800 and Pythia 848,000 comparisons; exact unchanged replay	Failed Pythia exposed preflight; separately frozen FP32 checker-ordering amendment before fresh Pythia calibration
P10	172,800 comparisons; exact replay	Exploratory erratum (ceiling, ratio direction, tied ranks); registered primary unchanged
P11	18,432 comparisons from retained P9 logits; exact replay	Fail-closed reporter; accepted report byte-identical
P12	Retained full-array receipts; six corrected analysis views replayed; eight adverse fixtures	Strict both-arm D gate (Pythia D unavailable); conflicting frozen constants; CE diagnostic label; Family A unchanged
P13	Retained receipts; corrected v2 reanalysis reproduces every value; 43 inherited and 26 adverse checks	Original Pythia numerical failure (B/C/S withheld); prose versus implemented realization criterion
P13-X	12,096 comparisons per study; exact server replay; local compact replay within 3 ULP	Failed Pythia-1.4B block-14 development; separately frozen batch-alignment amendment; independent verifier v2
P14	Pythia 112,768 and Qwen 102,144 comparisons; 56 selector replays each; exact compact replay	None in execution
P14-S	New endpoints reconstructed; parent payload rehashed; exact compact replay	Post hoc policy; first launcher stopped before models
P15	All 20 conditions per checkpoint; source OLS and geometry; exact compact replay	Reporter authentication correction; frozen design unchanged

In P9, a preserved failed Pythia exposed preflight motivated a separately frozen correction to ordinary-checker floating-point arithmetic after Qwen execution but before fresh Pythia calibration or evaluation; declared interventions, scientific procedures and numerical tolerances were unchanged (Appendix N). In P13-X, the analogous amendment for the selected Pythia-1.4B primary was chosen from exposed development numerical comparisons after other conditions' outcomes were known; it repeated development at unchanged gates before any primary evaluation (Appendix R). The P13-X larger checkpoints were run on CPU because their GPU routes failed an unchanged CPU-envelope certification criterion even though separate endpoint budgets passed; GPU availability did not waive that rule. In P12, the executed analysis checked Family D's numerical gate on one arm only, while its family-specific protocol required both; the correction adopts the stricter reading, which makes all Pythia D versions unavailable confirmatorily.

"Prespecified" and "frozen" refer to local, timestamped and hashed designs, not external preregistration. Several analyses were declared after related outcomes existed and are labeled as such: P11 reanalyzes retained P9 outcomes; P14-S chose its alternative selector after original P14 outcomes; the P13-X Pythia amendment followed other conditions' outcomes. Corrections to P10–P13 were made after observing results; they correct interpretation, eligibility and reporting, and preserve the original records.