What the Nonnegative Lens Misses: Energy-Matched Causal Tests of the J-Space in Open Language Models

Anonymous authors
Paper under double-blind review

Abstract

The Jacobian lens maps residual activations onto token-indexed output directions; "J-space" is a sparse nonnegative reconstruction of those atoms. We test how much causal work that reconstruction does on open-weight models with preregistered energy-matched interventions. When residual energy is matched, nonnegative J installs less of a contrastive concept shift than the leftover $R{=}D{-}J$. That $R \gg J$ ordering holds on the causal cells that report $J{-}R$—six families, rich and legacy reductions (future-summed as robustness), base models, and scales where capability permits (∼7B–32B)—and on a Gemma-4 invented-registry copy the remainder still transfers donor answers while reconstructed-J arms transfer none. Geometric capture stays in a 0.10–0.23 band from 0.5B to 32B. Nesting reconstructions on one dictionary is a selection-partition check: nonnegative OMP matches unconstrained least squares on the same support, while signed OMP raises reconstruction by 1.86–5.24 times at $k{=}100$ and flips J versus remainder on geography families. Increasing nonnegative budget fails to close the gap. Installation is not necessity: on Qwen2.5-32B-Instruct ($n{=}28$, descriptive), a post-hoc 2×2 of holding×answer-exposure shows an interaction, not an unconfounded holding effect; signed selection is unselective. Held-out known-attribute transport is strong for the full shift (0.82–0.93) and weak for the official nonnegative readout (0.15–0.23). The lens atoms track model knowledge; under the tested nonnegative-OMP reconstruction of contrastive shifts (rich-transport primary; future-summed as robustness), that code discards most of the equal-energy installable component. Preregistrations, frozen manifests, and code are released via an anonymous repository at https://anonymous.4open.science/r/jspace-iclr2027-anonymous-A1D0/.

1 Introduction

Gurnee et al. (2026) introduce the Jacobian lens: residual activations at layer ℓ map to token-indexed atoms $W_U[t]\, J_\ell$ by transporting the unembedding through an averaged linearization of the remaining network. "J-space" denotes sparse nonnegative combinations of those atoms. On Claude-family models that set is reported readable, steerable, used in flexible reasoning, and selectively necessary for holding intermediates—a profile that invites comparison with global workspace theory (Baars, 1988; Dehaene & Naccache, 2001; Mashour et al., 2020). Commentaries by Butlin et al. (2026) and Dehaene & Naccache (2026) already distinguish a privileged readable set from a privileged causal one. The original paper notes that the sparse decompositions are neither complete nor unique.

Two measurements already strain a simple workspace reading. A Qwen 3.6 27B reproduction found answer-token and intermediate swaps about equally effective, unlike Claude's intermediate-dominant pattern (Nanda, 2026). The original paper itself attributes only a median 6–7% of a concept vector's variance to the J-space component; verbal-report probes there already rescale directions to equal magnitude. Those do not settle the energy-matched installer question on open models: if J and $R{=}D{-}J$ are injected at the same residual energy, which moves the answer?

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.

1. Equal-energy install. Nonnegative J, remainder R, raw D, and a matched random control, scored by equal-energy margin AUC (integral of Δmargin over the dose grid) with item-clustered bootstrap. $R \gg J$ on the causal cells that report $J-R$ (six families; rich vs. legacy with future-summed robustness; base; ∼7B–32B where capability permits); on Gemma-4, remainder transfers donors while reconstructed J does not (§3, §4). Capacity-normalized ladders from 0.5B do not trend toward J-dominance.
2. Where the constraint sits. Nesting on one dictionary is a selection-partition sanity check: unconstrained LS on the nonnegative-OMP support equals nonnegative OMP. Signed OMP captures 1.86–5.24× more at k=100 and flips $J-R$ on geography. Increasing nonnegative budget fails to close the gap; capture near-saturates early in some models (§4).
3. Install versus need. On Qwen2.5-32B-Instruct (n=28), nonnegative J-ablation is selective against a primed control even though J is the weaker installer. A post-hoc 2×2 shows a holding×exposure interaction, not an unconfounded holding effect. Signed selection is unselective (§6). Held-out transport is strong for D and weak for the official nonnegative readout (§5).

The Jacobian lens finds real structure. The nonnegative reconstruction discards most of the installable, lens-aligned component.

2 Operational definitions and estimands

Glossary. Throughout, D is the contrastive residual shift, J its sparse nonnegative reconstruction (the operational J-space code), and $R{=}D{-}J$ the residual remainder. "Install" means moving answer margins under equal-energy injection; "need" means selective ablation. The primary causal statistic is equal-energy margin AUC, not recovery rate.

Lens and equal-energy swap. Dictionaries are $D_\ell = W_U[S]\, J_\ell$ over a vocabulary sample S. Three sizes appear below: full word-vocabulary (∼40k atoms after the tokenizer-convention fix in Appendix F), a padded ∼4k-atom default in a 4096-dim residual, and a 52-atom answer/concept reference. The primary rich transport lens averages vector-Jacobian products over a frozen 24-prompt corpus and all source positions, last-position target, with the learned final-norm scale folded into W_U; we use it as primary because it matches our causal question, while Anthropic's future-summed reduction is the official operator and our robustness / 32B-prereg primary. A legacy last-position arm is kept for comparison. Reimplemented future-summed matches the official operator at cosine 0.843/0.918/0.980 (layers 7/13/19 on Qwen2.5-7B). Each battery item yields a contrastive concept shift D; nonnegative Orthogonal Matching Pursuit (OMP) (Pati et al., 1993; Yaghoobi et al., 2015) produces the operational reconstruction J (k=25 for swaps; k=100 for k-sweeps and signed headlines); $R{=}D{-}J$. Arms J, R, D, and a norm-matched isotropic random control are scaled to the same residual-relative RMS energy first, then swept on a frozen dose grid so magnitude cannot decide the winner. The primary statistic is equal-energy margin AUC: the trapezoid integral of Δmargin over the dose grid, with pairwise item-clustered bootstrap (n=10,000). DR1 / DR1fs use $\{0.25, 0.5, 1, 2, 4\}$; the post-hoc k-sweep uses $\{0.5, 1, 2, 4\}$. No best layer, dose, or band is a primary estimand. Contrastive D is a signed transition between endpoint states; reconstructing D with nonnegative coefficients is not the same estimand as reconstructing nonnegative endpoints and differencing them (§7). Protocol, batteries, eligibility, and model list: Appendix A.

Coefficient constraints. The post-hoc k-sweep varies reconstruction on the same atoms: nonnegative OMP (k=5–100), signed OMP, rank-k PCA, random-k-atom control; each also at native scale (α=1). Full-span reconstruction is rank-vacuous. An in-lens reference rebuilds from the 52 answer/concept atoms versus a matched-rank random subspace. Nesting nonnegative OMP, unconstrained LS on that support, and signed OMP is a selection-partition check: the first two coincide, so the nonnegative constraint binds at atom selection. Signed OMP raises $\|J\|/\|D\|$ and flips $J{-}R$. Cross-dictionary figures are bounds, not an additive split; reconstruction-norm ratio is not explained variance.

Table 1: Equal-energy margin AUC (logit·dose; grid $\{0.25, 0.5, 1, 2, 4\}$). Negative $J-R$: remainder out-installs J. Same-model eligible sets.

model (n)	$J-R$	$J-$rnd	$R-$rnd
Qwen2.5-7B-I (238)	$\mathbf{-32.65}$ [−33.7,−31.5]	+18.07 [17.0,19.1]	+50.72 [49.7,51.7]
OLMo-2-7B (542)	$\mathbf{-16.46}$ [−17.1,−15.9]	+24.86 [24.2,25.5]	+41.32 [40.8,41.8]
BLOOM-7b1 (542)	$\mathbf{-14.13}$ [−14.6,−13.7]	+27.96 [27.4,28.5]	+42.09 [41.5,42.7]

3 The remainder out-installs J at matched energy

At matched residual-relative energy, operational J is not the dominant installable swap component—$R{=}D{-}J$ is. That ordering matters for interpreting what the operational J-space is doing: if the leftover of the decomposition installs more of the concept than the reconstruction itself, a sparse nonnegative code of the lens cannot be treated as the causal substrate of the concept under this instrument. Table 1 reports equal-energy margin-AUC differences under rich transport, with item-clustered 95% CIs on same-model eligible sets.

Ordering: $R \gg J >$ random; every CI excludes zero. J beats random by +18 to +28 logit·dose units (so the reconstruction is not inert), but the remainder dominates on distilled Qwen, open OLMo, and non-distilled multilingual BLOOM. The preregistered DR1 encoding_contrast gate—which asked whether remainder dominance was confined to a subset of families—was false; remainder dominance on every reported causal cell is the observed outcome. Legacy last-position agrees and is more extreme ($J-R$ near −43). Eligibility follows the preregistered filters: Qwen 238/610 (instruct filter); OLMo/BLOOM 542/610.

Against a pure geometric tautology. Because $R{=}D{-}J$ and nonnegative capture of $\|D\|$ is typically 0.10–0.30, R is nearly collinear with D (mean $\cos(R, D)$ at nn-OMP $k{=}100$: 0.986 OLMo, 0.994 Qwen, 0.954 BLOOM). Equal-energy $R \gg J$ is therefore partly geometrically expected. We keep the headline because the comparison is still the right energy-matched question for the operational readout, and because two results are not foreordained by that geometry: signed selection flips $J-R$ on geography families, and on Gemma-4 the nonnegative remainder transfers donors while every reconstructed-J arm (nonnegative and signed) transfers 0% (below). The headline compares equal injection energy, not equal captured energy: equal-energy effects are not a native-scale efficiency claim.

Future-summed, paired within item, still leaves remainder ahead on OLMo-2-7B ($J-R{=}-23.20$ [−23.81, −22.60], $n{=}542$) and Qwen2.5-7B (−29.51 [−30.67, −28.35], $n{=}238$). The 32B extension (future-summed primary) matches the frozen prediction: $J-R{=}-27.16$ [−28.22, −26.13] on the shared eligible set with 7B ($n{=}224$). Base 32B same ordering (−22.8 [−23.5, −22.0], $n{=}396$).

Figure 1: post-hoc six-family $J-R$ runs −33.54 to −13.25. Qwen3-8B/14B ($-22.8/-25.6$) and hybrid-attention Qwen3.6-27B (−13.6 [−14.9, −12.4]; $J-$random $=+31.8$) sit in-band. On post-hoc Gemma-4-12B (held-out invented-registry, 128 TEST pairs; $\alpha{=}1$ heldout-relation): raw D donor rate 0.973, nonnegative remainder 0.969, signed remainder 0.691; every reconstructed-J arm (nn_nn, nn_ls, signed_nn, signed_ls) and both random/permuted controls transfer 0%. Margin Δ for nn_nn ($\approx$0.84) $\leq$ random-rank floor ($\approx$0.99), so $J \leq$ random here. Remainder wins beyond geography; the geography-signed flip does not transfer. An earlier best-strength positive for J-swap does not survive energy matching.

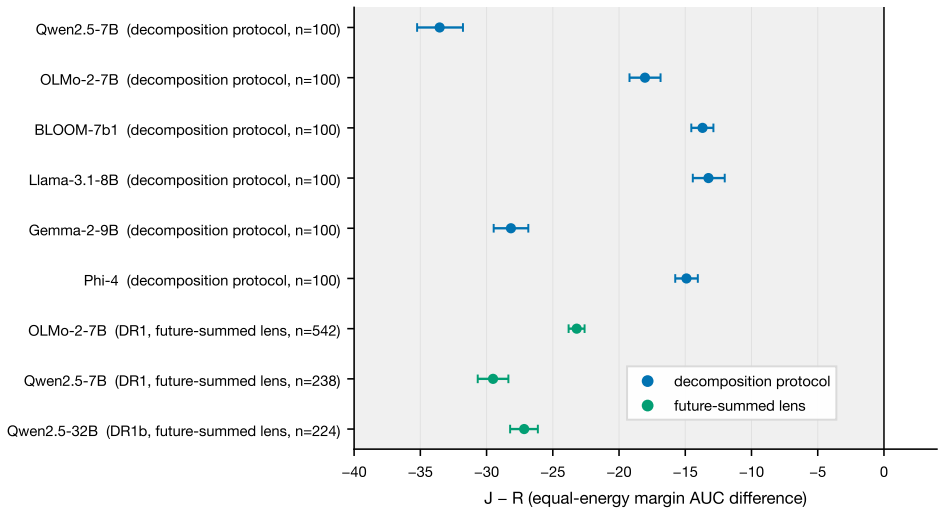

Figure 1: $J-R$ equal-energy margin AUC at matched residual energy (negative = remainder wins; item-clustered 95% CIs). Blue: post-hoc six-family k=100 (grid $\{0.5, 1, 2, 4\}$). Green: preregistered future-summed (DR1 grid). Axis matches the primary statistic in §2; no CI near zero.

4 Coefficient constraints, not a sparsity budget

Table 2 summarizes the post-hoc k-sweep on OLMo-2-7B (Qwen-7B and Gemma show the same qualitative pattern). Nonnegative capture near-saturates by k=5, so the primary swap budget k=25 was not too small on these models. Capture is not uniformly saturated everywhere—BLOOM still moves 0.243→0.296 from k=5 to 100—but the shared claim is that increasing the nonnegative budget fails to close the equal-energy $J-R$ gap.

Sign binds at selection. A signed sparse code on the same atoms captures ∼3× more on OLMo (0.51 vs. 0.165 at k=100) and moves $J-R$ from −18 to +6. At native scale (α=1), signed carries ∼10× the nonnegative causal effect (delta-margin +3.98 vs. +0.41). Nested reconstructions across the six k-sweep families (post-hoc) are a selection-partition check: unconstrained least squares on the nonnegative-OMP support yields nearly identical mean reconstruction ratios to nonnegative OMP, with small numerical exceptions (maximum ratio gaps ∼0.003 on BLOOM/Llama/Phi). Signed OMP raises the reconstruction-norm ratio to 0.43–0.60 at k=100 and flips $J-R$ on geography families: a large slice of what the operational decomposition calls "remainder" is signed lens-atom content the nonnegative code never selected. At signed k=100 on OLMo, $\|J\|/\|D\|\approx 0.51$ while $\|R\|/\|D\|\approx 0.86$; these are separate norms, not additive fractions of a partitioned shift.

Across six families, signed capture is 0.43–0.60 of the shift and 1.86–5.24× nonnegative at k=100; on Qwen3.6-27B the signed flip is +25.3 [+24.1, +26.4]. The 52-atom answer-token reconstruction captures 0.32–0.41 (∼3× matched-rank random). Cross-dictionary bounds and family-level capture detail: Appendix B. Capacity-normalized geometric ladders (Qwen2.5 0.5B–32B; Pythia 70m–12b; $k \propto$ width) keep nonnegative capture in a 0.10–0.23 band with no J-dominance (Appendix D). Causal $J-R$ is capability-gated (∼7B–32B). Net signed capture declines descriptively (slopes −0.074, −0.078 per $\log_{10}$ decade); not scaling laws. Preregistered 72B gate unmet; omitted.

Table 2: Post-hoc k-sweep on OLMo-2-7B (grid $\{0.5, 1, 2, 4\}$). recon_frac=$\|J\|/\|D\|$, not explained variance.

variant	k	recon_frac	$J-R$	$J-$rnd
nn_omp (operational)	5–100	0.165 (flat)	−18	+25
signed_omp	5	0.329	+0.5	+40
signed_omp	25	0.397	+2.8	+41
signed_omp	100	0.510	**+6.1**	+41
pca_rank_k	100	0.261	−14.8	+27
random_k	100	0.065	−37	+7

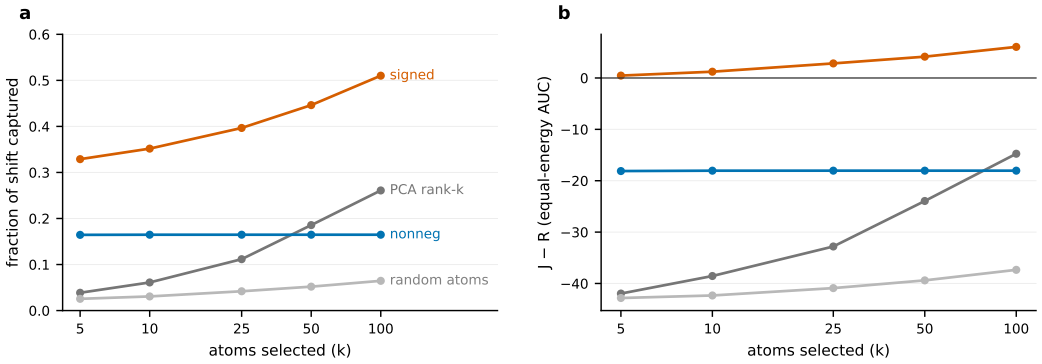

Figure 2: (a) Nonnegative capture near-saturates early on OLMo; signed keeps growing. (b) $J-R$ flips under signed selection.

5 Known-attribute transport as a calibration check

Entity-swap shifts constructed in a neutral frame are injected at the entity position of held-out single-attribute probes. Templates used to form the shift are unused at probe time; the intervention layer is chosen from the full-shift arm on a disjoint selection split; uncertainty is pair-clustered bootstrap. Matched-value pairs (same target and foil token) have margin identically zero by construction and are not negative-control evidence of specificity. The protocol tests template transport of known attributes—not open-ended unsupervised discovery of facts, and not native mediation along the model's ordinary computation.

On unequal-value pairs, full-shift recovery is 0.93 [0.89, 0.96] (Qwen3-8B), 0.82 [0.79, 0.85] (OLMo-2-7B), and 0.91 [0.87, 0.95] (Qwen2.5-7B), against random baselines of 0.016–0.033. Nonnegative J carries only 0.15–0.23 of the transport; signed codes carry 0.52–0.67; the remainder tracks the full shift. Nonnegative and signed intervals do not overlap. Selected layers land early (∼0.2 relative depth). On a Qwen2.5-7B capital cell, recovery stays near 1 through relative depth 0.71 and collapses by 0.86—a hopping-too-late signature (Biran et al., 2024).

Label-guided relation vectors (concept→answer) show within-family parallelism 0.86–0.97 vs. ∼0.3 cross-family; that gap is the expected signature of supervised, within-family relation geometry rather than unsupervised discovery of semantic organization.

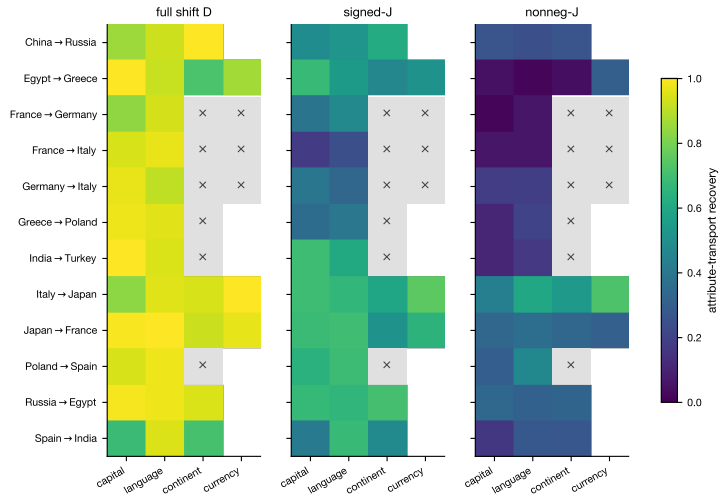

Figure 3: Counterfactual attribute transport per readout (Qwen3-8B; held-out pairs at the shared selection-split layer). Color = recovery; × = matched-value pairs (margin 0 by construction, not a negative control). Mean recovery: full shift 0.93, signed J 0.52, nonnegative J 0.23.

6 Necessity is not installation

Selective ablation tests need, not install: a component can be necessary for a prompt class without being the dominant installer under equal-energy swaps. Preregistered P5f compares a faithful J-ablation on a two-hop holding prompt against a primed control that holds nothing and states the answer twice. We report Qwen-32B at n=28 (descriptive; no CI), k=10, layers 24–34 of 64:

Table 3: P5f ablation, Qwen2.5-32B-Instruct (argmax acc.). Registered: reason2−primed =0.25.

arm	base	abl.	drop	n
auto (0 held, 1×)	0.929	0.714	+**0.214**	26
primed (0 held, 2×)	0.786	0.821	−0.036	22
reason1 (1 held)	0.821	0.786	+0.036	23
reason2 (2 held)	0.929	0.714	+**0.214**	26
reason3 (3 held)	0.964	0.750	+**0.214**	27
cot / cot3	1.000 / 1.000	1.000 / 0.893	0 / +0.107	28

Registered H1 passes (reason2−primed =0.25>0.20; primed retention preserved). Auto matches reason2. Selectivity vs. primed is not an unconfounded holding test: primed restates the answer and reason2 does not. The numeric H1 gate still passes; the pre-committed "holding beyond priming" reading is superseded by the 2×2, because primed differs from reason2 on answer exposure as well as holding.

Table 4: Post-hoc holding×answer-exposure 2×2 (same 28 items; nonnegative abl.; drop). Descriptive (n=28, no CI). Holding effect given absent answer =0; given restated ≈+0.393.

	answer absent	answer restated
hold (2-hop)	+0.214	+0.357
reaccess (0-hop)	+0.214	−0.036

The 2×2 (incl. hold×restated +0.357) shows an interaction: ablation sensitivity is null for holding when the answer is absent (both hops +0.214) and large when restated (hold +0.357

vs. reaccess −0.036). Descriptively at n=28, this is a holding×answer-exposure interaction, not an unconfounded holding effect. Matched-random drops $\leq$ 0.07. Signed selection is unselective (reaccess-redundant drop +0.39). Same model's 32B swap has $J-R=-27.2$: a small J-component can break some prompts without being the dominant installer or shown necessary for holding.

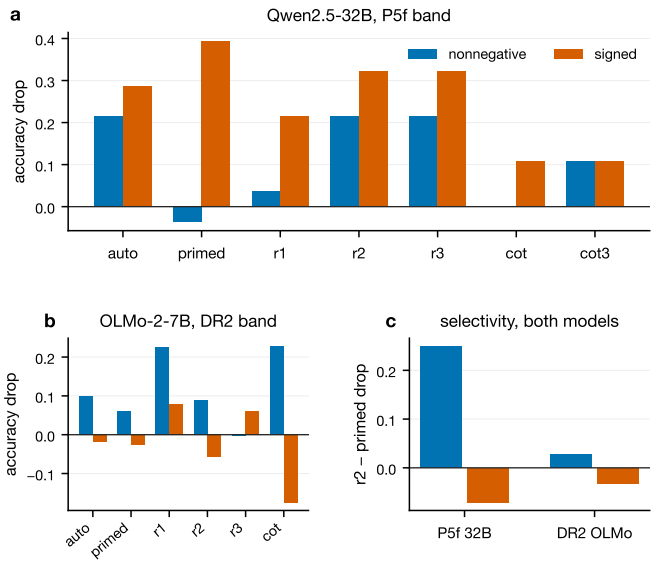

Figure 4: (a) Qwen-32B at the P5f band: nonnegative ablation hits reason2/auto and spares primed; signed ablation hits the priming control hardest. (b) OLMo-2-7B at the matched DR2 band. (c) reason2−primed selectivity on both models.

W3a (remove nonnegative from ablation) failed. On OLMo, signed ablation is inert (reason2 drop −0.057). On Qwen-32B, signed is destructive without selectivity (reason2 +0.32, primed +0.39). One signed atom captures as much as saturated 25-atom nonnegative (recon_frac 0.170 vs. 0.172) yet ablating it does not reproduce the primed-control pattern.

7 Discussion

Interpreting the operational readout. Operational J-space under-captures causally relevant, lens-aligned content: signed codes recover 1.86–5.24× more by reconstruction-norm ratio. Two incorrect readings should be resisted. First, nonnegative J is not the dominant installer (recon 0.10–0.30 padded / 0.10–0.22 unpadded; it loses at matched energy on the reported cells). Second, the concept does not live entirely outside the lens either: roughly one-third to two-fifths of the shift sits in the answer-token subspace, about 3× a matched-rank random baseline. Caveat: if endpoints are nonnegative combinations, their difference D need not be; poor nonnegative reconstruction of D does not alone prove incompleteness of nonnegative endpoint codes.

Reporting. Anyone using J-space as a measurement of where concepts live should report signed reconstruction alongside nonnegative. The gap varies ∼3× across families—large enough to look like family differences in "having" a J-space when it is mostly the sign constraint. Necessity: on Qwen-32B (n=28), nonnegative ablation is selective vs. primed and the post-hoc 2×2 shows holding×exposure interaction, not unconfounded holding; signed is unselective; OLMo has no clean necessity result.

Related work. The Jacobian lens extends logit and tuned lens tools (nostalgebraist, 2020; Belrose et al., 2023). Contrastive activation addition and difference-in-means share the equal-energy install question without sparse nonnegative coding (Turner et al., 2023; Rimsky

et al., 2024). Gurnee et al. already magnitude-match verbal-report probes; our contribution is energy-matched J versus R on open models with contrastive shifts, dose curves, and coding comparisons—not the first use of magnitude matching. Signed versus nonnegative coding is a sparse-coding control, not a second workspace hypothesis; compare AbsTopK dropping the SAE orthant (Zhu et al., 2026). Installation is not unique necessity (Makelov et al., 2024; Vaidyanathan et al., 2026), and linear representation is not linear accessibility (Garg et al., 2026). Extended citations live in Appendix C.

Limitations. Untested above 32B. Ladder slopes are checkpoint-conditional on the released Qwen2.5 and Pythia-deduped suites (Appendix D / Figure 5); they are not a claim that every training run will show the same model-size trend. Necessity is one model and 28 items (descriptive), with selectivity versus primed and a post-hoc holding×exposure interaction 2×2, not an unconfounded holding test. Transport covers one geography domain plus one Gemma-4 invented-registry cell; matched-value pairs are tautological zero margins. Nested reconstructions are a selection-partition sanity check; cross-dictionary figures are bounds rather than an additive variance split. The k-sweep, nested partition, Gemma-4 remainder curves, and P5f 2×2 are post-hoc relative to the original manifests. Some preregistered DR1 damage-related secondaries (including KL and unrelated-control accuracy) were never implemented or retained; they are disclosed as absent rather than unsummarized. DR1/DR2 ablation and routing aggregates that were run are in the supplement and committed results for inspection. Substrate without the lens (Appendix E) is distributed and only partly direction-selective. Nothing here refutes a workspace in Claude. Claims concern the operational decomposition at or below 32B.

Taken together, the Jacobian lens identifies real, knowledge-aligned structure. When intervention energy is matched and dictionary coding is examined directly, the tested nonnegative-OMP reconstruction of contrastive shifts discards most of the equal-energy installable component under our readout (with future-summed as a robustness operator); reporting both signed and nonnegative decompositions is necessary for claims about that operational readout.

AI use statement

Generative models were used as coding and drafting assistants: protocol design, intervention and analysis code, reading results against frozen files, literature search, bibliography, structure, and manuscript drafting and editing. They were not used to invent training data, theorems, or proofs. Batteries are frozen JSON. Some prompt and protocol text was drafted with assistance and then frozen before any inference. Code, numbers, and wording were checked against preregistrations, manifests, and result files. The authors remain solely responsible for the scientific claims, experimental design, analysis, final wording, and all submitted artifacts.

Reproducibility statement

Preregistrations, frozen manifests, hash-embedded batteries, and committed aggregates are in the anonymous code/results archive at https://anonymous.4open.science/r/jspace-iclr2027-anonymous-A1D0/ deposited with this submission; Appendix G maps artifacts to results. Primary statistics use item- or pair-clustered bootstrap as in §2. Lens *.pt* weight exports and an export_lens_assets.py exporter are not included: E-series and transport cells reported here are analysis from saved aggregates (lens rebuild deferred). Manifests record revision: main at freeze; immutable content hashes for primary aggregates under results/nested/ and results/recovered/ are listed in Appendix G. The calibrated swap, k-sweep, scale, and necessity runs are tied to those committed aggregates. The appendix and anonymous code archive are the intended reproduction entry points.

Ethics statement

The work evaluates an interpretability instrument on public open-weight models and adds no new capability. The intended use is more careful auditing: report signed and nonnegative

readouts, match intervention energy, and do not treat installation as necessity. Misuse potential is limited; the corrections make over-claiming easier to catch.

References

Marah Abdin et al. Phi-4 technical report. arXiv preprint arXiv:2412.08905, 2024.

Anthropic. jacobian-lens (official implementation). GitHub repository, 2026. URL https://github.com/anthropics/jacobian-lens.

Bernard J. Baars. A Cognitive Theory of Consciousness. Cambridge University Press, 1988.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023.

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

BigScience Workshop. BLOOM: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100, 2022.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

Patrick Butlin, Derek Shiller, Dillon Plunkett, and Robert Long. Consciousness and cognitive access in LLMs: A commentary on "verbalizable representations form a global workspace in language models". Eleos AI, 2026. URL https://eleosai.org/papers/eleos_gwt_commentary_20260706.pdf.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition, 79:1–37, 2001.

Stanislas Dehaene and Lionel Naccache. Does claude possess a conscious global workspace? a commentary on "verbalizable representations form a global workspace in language models" (Gurnee et al., Anthropic). Unicog, 2026. URL https://unicog.org/wp_2025/wp-content/uploads/2026/07/Dehaene-and-Naccache-Workspace-commentary-on-Gurnee-Lindsey-June-2026.pdf. Invited/external commentary on Gurnee et al. 2026.

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? In Proceedings of Thirty Ninth Conference on Learning Theory, PMLR 336, 2026. URL https://arxiv.org/abs/2602.11246.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR), PMLR 236, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118, 2024.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

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

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In International Conference on Learning Representations (ICLR), 2024.

George A. Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. Neuron, 105:776–798, 2020.

Neel Nanda. A review of anthropic's global workspace paper. LessWrong, 2026. URL https://www.lesswrong.com/posts/zFJ3ZdQwrTWE9jT5S/a-review-of-anthropic-s-global-workspace-paper. Qwen 3.6 27B replication with Camila Blank and Agam Bhatia.

Neel Nanda and Joseph Bloom. Transformerlens. GitHub repository, 2022. URL https://github.com/TransformerLensOrg/TransformerLens.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

OLMo Team. 2 OLMo 2 Furious. arXiv preprint arXiv:2501.00656, 2025.

Yagyensh C. Pati, Ramin Rezaiifar, and Perinkulam S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, 1993.

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. arXiv preprint arXiv:2606.27510, 2026.

Mehrdad Yaghoobi, Di Wu, and Mike E. Davies. Fast non-negative orthogonal matching pursuit. IEEE Signal Processing Letters, 22(9):1229–1233, 2015.

Xudong Zhu, Mohammad Mahdi Khalili, and Zhihui Zhu. Abstopk: Rethinking sparse autoencoders for bidirectional features. In International Conference on Learning Representations (ICLR), 2026. URL https://arxiv.org/abs/2510.00404.

A Protocol, batteries, and models

Frozen, hash-embedded batteries: 610-item factual counterfactual swap set (100-item subset for the k-sweep); 280-item held-intermediate set varying holding, priming, and externalized computation; held-out known-attribute transport set (12 single-token countries, 24 pairs; capital, language, continent, currency). Eligibility is preregistered: the unmodified model must argmax the expected answers. Cross-model contrasts use shared eligible sets. Causal runs use frozen manifests with tested verdict functions; deviations logged (Appendix G). Models run in bf16 via TransformerLens (Nanda & Bloom, 2022) at frozen revisions, with an architecture-agnostic autograd backend for hybrid-attention

models the standard path cannot load. Families: Qwen2.5 (Qwen Team, 2024), OLMo-2 (OLMo Team, 2025), BLOOM (BigScience Workshop, 2022), Llama-3.1 (Grattafiori et al., 2024), Gemma-2 (Gemma Team, 2024), Gemma-4-12B (Gemma Team, 2026), Phi-4 (Abdin et al., 2024), Qwen3, Qwen3.6-27B, Pythia-deduped (Biderman et al., 2023). Dictionaries follow the official jacobian-lens construction (Anthropic, 2026). Future-summed operator: $J_\ell = \text{mean}_p[\sum_{p' \geq p} \partial h_{\text{final}}[p']/\partial h_\ell[p]]$. Dose grids and bootstrap as in §2. Default dictionary pads to ∼4000 atoms in a 4096-dim residual.

B Coefficient extras and cross-dictionary bounds

On the original three families, signed capture is fairly flat (0.51–0.55) while nonnegative capture varies 3× (0.10–0.30) and rank-orders the screening-era causal J-effect. Qwen's apparently inert J is where the nonnegative constraint most strongly limits capture (∼10% of $\|D\|$). BLOOM, where nonnegative capture is highest among those families, is the model whose J carried the most causal effect. Causal k-sweep on Llama-3.1-8B, Gemma-2-9B, and Phi-4 repeats the pattern: nonnegative under-capture with remainder dominance, a signed flip to parity or better, and an in-lens reference near 3× a matched-rank random baseline. Cross-dictionary bounds: 10–22% of the shift in nonnegative J on the unpadded answer-token dictionary (10–30% on the padded vocabulary dictionary); 32–41% in the answer-token reconstruction; 43–60% in the 100-atom signed vocabulary code. Neither "all sign" nor "all geometry" alone accounts for the full pattern. Native-scale OLMo path: results/recovered/ksweep_olmo_native_scale.json.

C Extended related work

Distributed alignment search and related subspace methods raise illusion risks beyond those already flagged for activation patching (Geiger et al., 2024; Makelov et al., 2024). On overlapping measurements we agree with the original paper and the Qwen 3.6 27B reproduction (Gurnee et al., 2026; Nanda, 2026). Those prior numbers are native variance share and swap recovery at the official code's own scale, which is not installer potency at matched residual energy $E(D)$. Matching energy upsamples J to the full-shift dose; the official reconstruction still loses to R. We retain the Jacobian atoms—which provide causal transport into a vocabulary-aligned map—and do not force that map through the nonnegative cone. Occupancy and capacity claims are readout-relative in the sense of Garg et al. (2026).

D Scale ladders

Geometric metrics use structural eligibility only, so they are valid at every size; the item set is tokenizer-determined and identical across rungs within a family. Causal metrics are capability-gated. On Qwen2.5 (0.5B–32B) nonnegative capture stays in a 0.10–0.23 band with no consistent drift toward J-dominance. Signed capture at fixed k=100 declines with width, as expected when a fixed atom budget is a shrinking slice of a widening residual. Capacity normalization (k=$d/64$, $d/32$, $d/16$) removes most of that decline. Item-resampling slopes on net signed capture at k=$d/64$ are -0.074 [-0.077, -0.072] per $\log_{10}$ decade of parameter count on Qwen and -0.078 [-0.081, -0.076] on Pythia, on these checkpoints and the 150 shared tokenizer-identical items. Neither matched-width depth contrast met a frozen trend criterion. Pythia causal cells exist only at 6.9b and 12b; smaller rungs fail the capability gate and are geometric-only by design.

E Substrate without the lens

Protocol-matched ablation on OLMo (DR2) is J-specific (a dimension-matched random ablation is about null) but not reasoning-selective: the faithful localized ablation drops reason2 by only +0.089 while damaging the primed retention control. The routing question is inconclusive because the ablation is too destructive.

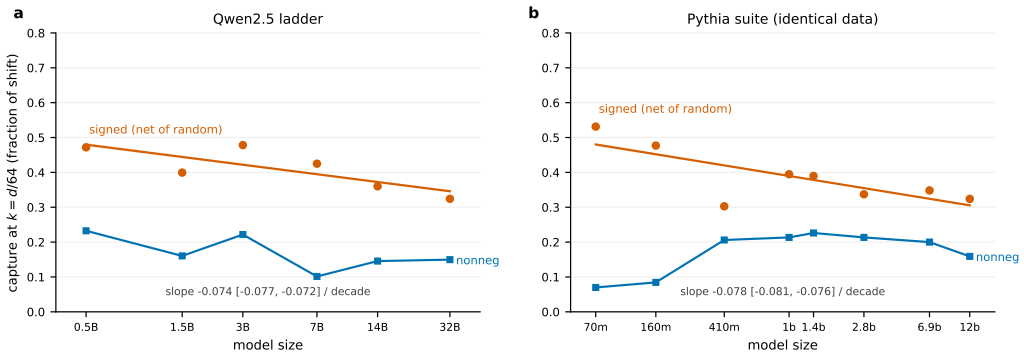

Figure 5: Capacity-normalized signed capture at $k{=}d/64$, net of matched-k random (orange), and flat nonnegative capture (blue), across (a) Qwen2.5 and (b) Pythia-deduped under an identical protocol (same eligibility rule and capacity-normalized k). Slopes are conditional on these checkpoints and items, and run away from J-dominance.

A lens-free activation-patching trace (DR3; 90/90 discovery/confirmation split) selects 11 four-layer windows with mean recovery above 0.20: eight at the changed-state slot, three at prediction positions, none at the final query. On confirmation, all 11 have $\mathrm{CI}_{\mathrm{low}} > 0$ (73 eligible items) and the windows span all 32 layers. The assay does not pin a compact workspace band and does not, by itself, establish redundant storage at every layer.

Interchange (DR4) on a disjoint 64-item battery is position-selective (neighbor mean +0.02; same-state identity 0; different-state fully moves the answer) but only partly direction-selective: a shuffled-coordinate donor keeps about half the effect (+0.48). We do not take the protocol's reusable-workspace-candidate label. About half of the different-state effect is the magnitude of a large 32-layer perturbation, the same magnitude confound the calibrated swap was built to remove.

F Instrument audits

Two silent, family-level failure modes were caught by diagnostics, not by p-values on results. First, the vocabulary sampler originally admitted GPT-style leading-space markers but not the SentencePiece leading-space marker. SentencePiece models got nine-atom dictionaries and near-zero swap success that looked like a family-level causal split. After the fix, Gemma-2-9B swap rises from 0.038 to 0.423 with a proper ∼40k-atom dictionary. All pre-fix cross-family numbers were discarded. Second, the learned final-norm scale has to be folded into W_U; omitting it disagreed with a naive builder by more than the signal.

G Preregistration, deviations, and artifacts

DR1–DR4 run under frozen manifests with tested verdict functions. DR1b, W3a, and the 32B interpretation table were written before inference. Deviations, none of which changed frozen inputs: DR2's matched-random control was omitted on the first pass and re-run; DR3/DR4 ran past a frozen gate because the halt reason (remainder routing) made lens-free localization the next question; DR1b supersedes an earlier 32B trigger with its own namespace; DR3 confirmation and four DR4 arms were unimplemented stubs at freeze and were implemented before those runs. The k-sweep, nested partition, portability replications, capacity-normalized reanalysis, Gemma-4 remainder-curve audit, and P5f holding×exposure 2×2 are post-hoc robustness analyses.

Primary numbers in the main text are tied to committed JSON aggregates. 2026-09-16 addenda (sha256): xval_jbar_result.json abab1af8···b18381; Gemma-4 remainder curves 520b687a···9f2afc; nested summary f9ba6f34···206c5774; nested jsonl Qwen/OLMo/BLOOM/Llama/Gemma/Phi 04c32404···007f9e65

/ 97d0f3f4· · · 788853a7 / f4dc3f5f· · · de6728b7 / 328017ae· · · c5291e7d / 9c7754ac· · · 0c398304 / 0519f3a3· · · 4e14fa9f; P5f DEV/TEST 3091acbe· · · ee5ad01a / 5924072f· · · 79df1cc6. Full digests in results/nested/SHA256SUMS. Native-scale OLMo +3.98 vs +0.41: results/recovered/ksweep_olmo_native_scale.json (source jsonl sha256 52f18570· · · 662854ee). Base-32B $J-R= -22.8$ (n=396): results/recovered/dr1b_qwen32b_base_aggregate.json (items sha256 cc24419e· · · acf3c466c). Table 1 BLOOM same-model $R-$rnd +42.09 [41.5,42.7] (n=542): results/recovered/dr1_bloom7b1_same_model_r_minus_random.json (aggregate sha256 58883c5e· · · aa5fca23; point estimate from DR1 pairwise means; CI from Astra same-model recomputation—raw DR1 JSONL not shipped). A short non-archival workshop version of the instrument-calibration argument exists; this paper is the archival causal-validity account.