ONE BIT IN A HUNDRED: THE SPARSE SIGNED SUBSTRUCTURE OF BEHAVIORAL WEIGHT EDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Behavioral weight edits, such as task-vector negation, debiasing deltas, and persona or refusal edits, are usually treated as dense, full-precision objects that are trained and then merged. We ask how much of such an edit is actually doing the work. Using social-bias removal as a testbed, we train low-rank contrast vectors from a model's own likelihood-based self-diagnosis and then strip them to one bit per weight and 1% of coordinates. Under a deployment-honest collateral budget in which every metric is measured with the intervention active, binarization retains 87% to 105% of full-precision removal across models from 2.6B to 32B, and a seven-condition factorial shows where the information lives. The true sign field at magnitude-selected coordinates recovers $+0.353$ of bias while the same field at random coordinates recovers $+0.070$; sign permutation, layer shuffling, and bottom-magnitude support collapse to ≈ 0. On a fixed grid of edit scales, the selection effect is $+0.284$ [$+0.169$, $+0.407$] on the ten pre-registered cells, unanimous, and $+0.201$ [$+0.125$, $+0.284$] on a twenty-cell scale-up spanning four benchmark families, positive in 19 of 20 cells. The concentration is a property of the edit scale. An adaptive, bias-blind rule that sets each edit's scale from a held-out collateral budget finds that random-coordinate signs tolerate about four times the scale of magnitude-selected ones, and at those scales the gap between them shrinks to intervals covering zero at both tiers tested; magnitude selection therefore buys the effect at a fraction of the scale and collateral. The resulting edit merges into the checkpoint as a persistent intervention, with no serving-time hook and no additional inference-time compute. At matched collateral and selection space its removal is statistically tied with activation steering and SentenceDebias driven by the same signal, DPO trained on that signal removes no detectable bias, and prompting fails the budget in 16 of 18 conditions, so what distinguishes the edit is persistence rather than raw efficacy. Alongside the method we contribute an evaluation protocol built on integer-item budgets, always-on collateral measurement, equal selection spaces, and per-method nulls. We also characterize a regime of low pre-existing bias in which the edit backfires and show that an abstention gate built on it fails held-out validation.

1 INTRODUCTION

Behavioral weight edits—task-vector negation, debiasing deltas, persona and refusal edits—are usually treated as dense, full-precision objects: train a delta, merge it. This paper asks the prior question: **how much of that object is doing the work?** We use social-bias removal as the testbed because it supplies measurable target behavior across several benchmark families, an endogenous signal source (the model's own likelihood-based self-diagnosis (Schick et al., 2021)), and a deployment cost that must be accounted for. Three observations, all under a collateral budget measured with the intervention active, set the question up:

1. **Magnitude precision is nearly dispensable.** Binarizing the trained contrast vector to sign-plus-one-scale-per-tensor retains 86.8% of full-precision removal at $\leq$3.8B (deficit -0.040 [-0.093, $+0.007$], $n = 8$ cells, covers 0) 98.5% at 7–9B, a deficit that is small but measurable (-0.005 [-0.010, -0.002], $n = 4$ cells, excludes 0), and 105.4% at 27–32B (binary $-$ fp $+0.021$ [-0.006, $+0.074$], covers 0).

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.

2. **Density is largely dispensable.** On the registered sparsity panel, zeroing 99% of coordinates costs nothing measurable at $\leq$3.8B or 7–9B; on a broader panel that adds bbq_Age cells, 1% density keeps 98%, 75%, and 97% of the dense one-bit removal at $\leq$3.8B, 7–9B, and 27–32B, and every tier falls off a cliff by 99.9% (§5.5).
3. **The learned sign structure is not dispensable.** Randomizing signs at identical scale and sparsity removes most of the effect: at the cell unit the paired difference is $+0.135$ [$+0.049$, $+0.213$] over 44 target$\times$axis cells (sweep-level means $+0.221$ real vs. $+0.009$ random over $n = 502$ sweeps; App. G).

These facts do not yet say *where* the information lives—in the signs, in which coordinates were selected, or in where those coordinates sit. §5.3 answers with a factorial decomposition: destroy sign identity, coordinate–sign pairing, or layer placement and the effect vanishes; destroy tensor placement and it vanishes in every family but one (gemma retains about half of it, at 2.6B and at 27B); move the true sign field to random coordinates and a fifth survives; keep magnitude-selected coordinates with their true signs and it all survives. **Behavioral edit information is concentrated in a magnitude-identifiable sparse signed substructure—concentrated, not exclusive, and concentrated specifically at the edit scales a deployment uses.**

Contributions. (1) **Sparse signed structure**: a seven-condition decomposition establishing necessity (sign, pairing, and layer perturbations ≈ 0), sufficiency (1% of coordinates at 1 bit each), and concentration at the deployment scale ($\Delta_{\text{selection}} = +0.284$ [$+0.169$, $+0.407$] on the ten pre-registered cells, unanimous; $+0.201$ [$+0.125$, $+0.284$] on a 20-cell scale-up over four benchmark families and 2.6–32B, positive in 19 of 20), with the random-coordinate sign field carrying real signal ($+0.068$). (2) **Adaptive deployment**: a bias-blind rule that sets the edit scale from a held-out collateral budget, under which the concentration turns out to be a matter of scale rather than exclusivity: random-coordinate signs tolerate about four times the scale of magnitude-selected ones and then come close to them at both tiers tested (§5.3). (3) **Persistent intervention**: the edit merges into the checkpoint with no serving-time hook or per-token compute, and at matched collateral its removal ties inference-time steering and projection built from the same signal, while DPO trained on that signal removes none (§5.7). (4) **The evaluation protocol**: integer-item collateral budgets, always-on measurement, equal selection spaces, and per-method nulls, under which six method families sit on one frontier and every reported number traces to a measured artifact through a machine-checked manifest (App. G).

2 Related work

Delta compression and merging. BitDelta shows fine-tune deltas survive 1-bit sign compression (Liu et al., 2024); DARE shows task-performance deltas tolerate *random* 90–99% drop with rescaling (Yu et al., 2024); TIES traces merging interference to redundant small-magnitude updates and to sign disagreements, and resolves both by trimming and sign election (Yadav et al., 2023). Our decomposition speaks to all three and lands between them: the redundancy DARE exploits is real here (the true sign field at random coordinates carries $+0.068$ over random signs), but under a deployment collateral budget the effect is concentrated at magnitude-selected coordinates ($\Delta_{\text{selection}} = +0.284$), a dependence random-drop compression does not predict—**the behavioral-control regime rewards what DARE's regime forgives**—and the necessity side (sign randomization, pairing destruction, layer shuffles ≈ 0) is, to our knowledge, new at this granularity. The divergence is a property of the edit scale rather than of the sign field alone: once each arm may grow its scale to its own budget limit, the gap to random-coordinate signs is no longer distinguishable from zero, the DARE-consistent outcome, reached at about four times the edit scale (§5.3). **Task arithmetic.** Negating task vectors removes behaviors (Ilharco et al., 2023), including social bias (Dige et al., 2024); we inherit the mechanism and ask the information question of it. **Self-debiasing.** Prompt-level self-diagnosis (Schick et al., 2021) established that models can recognize their own biases; we move that endogenous signal into weights and show prompting itself fails an always-on collateral budget in 16 of 18 conditions (§5.7). **Inference-time debiasing.** Activation steering (Turner et al., 2023; Rimsky et al., 2024) and projection (SentenceDebias, INLP; Liang et al., 2020; Ravfogel et al., 2020, ported from Meade et al., 2022) are evaluated from the *same* elicited signal under the same budget; steering and SentenceDebias tie the edit, INLP fails its own null. **Preference optimization.** DPO (Rafailov et al., 2023) needs dense pairwise signal; in other settings it debiases about as well as task-vector

negation (Dige et al., 2024), but §5.7 shows self-elicited contrast is sparse in exactly that way—a signal–method mismatch, not a general inferiority. **Evaluation critiques.** Debiasing can hide bias rather than remove it (Gonen & Goldberg, 2019), "bias" is often left without a normative definition (Blodgett et al., 2020), intrinsic and extrinsic bias metrics disagree (Goldfarb-Tarrant et al., 2021; Delobelle et al., 2022), and benchmark gains often come with lost language-modeling ability (Meade et al., 2022). The protocol turns the last concern into requirements these works motivate but do not state: collateral measured with the intervention active, equal selection spaces, per-method nulls, and integer-item budget boundaries.

3 METHOD

Elicitation. For target T and axis a, a designer D (default $D = T$) produces paired corpora (C^+, C^-) by likelihood-based self-diagnosis: biased and debiased continuations are identified by D's own likelihoods under a diagnosis framing (free-generation and forced-choice elicitation fail at these scales; App. F). **Edit.** Train a rank-16 (LoRA; Hu et al., 2022) contrast task vector ΔW on the attention projections (per-family module resolution in §5.1), binarize per tensor ($E = s_t \operatorname{sign}(\Delta W_t)$), optionally sparsify to 1% by $|\Delta W|$, and merge the negation $W \leftarrow W - \alpha E$, choosing α among four values by the in-budget argmax. **`contrast_gap`.** The designer's likelihood separation between C^+ and C^- is a corpus property, computable before training; its mechanical reading (§5.7) is the fraction of items on which the model can express any contrast. Whether it tracks realized removal is population-dependent, and the abstention gate built on it does not validate (§5.8), so we make no predictor claim.

4 EVALUATION PROTOCOL

Budget. An intervention is in budget iff it costs at most 4 of 200 MMLU items (Hendrycks et al., 2021), compared in *integer items* so that configurations sitting exactly on the boundary are decided exactly (App. A), and perplexity ratio ≤ 1.10. **Always-on.** Collateral is measured with the intervention active—merged edit, added steering vector, applied projection, prepended prompt—during bias probes, MMLU, and perplexity alike. **Equal selection space.** Every method optimizes over exactly four configurations; cells with no in-budget configuration score 0 and are recorded. **Per-method nulls.** Each method must beat a structure null (sign-shuffle / matched-norm random direction / random subspace) and the partition null (identical pipeline on two halves of the same biased corpus). **Statistics.** Three seeds per cell; percentile bootstrap over target×axis cells; nan→0 with counts; every α persisted (`alpha_trace.jsonl`), so every budget decision can be recomputed without GPU time. Determinism is scoped in App. G.

5 RESULTS

5.1 SETUP

Six instruction-tuned checkpoints: gemma-2-2.6B (Gemma Team, 2024), qwen2.5-3B/7B (Qwen Team, 2024), llama-3.2-3B and llama-3.1-8B (Grattafiori et al., 2024), phi-3.5-3.8B (Abdin et al., 2024); the decomposition scale-up adds gemma-2-27B and qwen2.5-32B (one seed, one axis each) and the IFEval arm adds gemma-2-9B. All panels use instruction-tuned models (the single base-model measurement is in §6). Edited modules are `q_proj`, `k_proj`, `v_proj`, `o_proj` for gemma-2, qwen2.5, and llama-3.x, but the fused `qkv_proj` *alone* for phi-3.5 (`o_proj` not edited)—the "(q,k,v,o)" shorthand applies to the first group only. Axes: occupation→gender; CrowS-Pairs socioeconomic (Nangia et al., 2020); four BBQ group-vs-unknown axes (Parrish et al., 2022); StereoSet intrasentence (Nadeem et al., 2021); onboarding gates in App. F.

5.2 THE EDIT REMOVES BIAS AGAINST BOTH NULLS

Pooled over all 15 matched cells at the cell unit, real − sign-shuffle $= +0.243$ [$+0.139$, $+0.359$] and real − partition $= +0.228$ [$+0.138$, $+0.328$]; both exclude zero (per-cell examples in App. A, Table 4). Both nulls are reported unaveraged: the partition null is not vacuous (it reaches $+0.124$ on gemma|occ), and sign-shuffle is the tighter null.

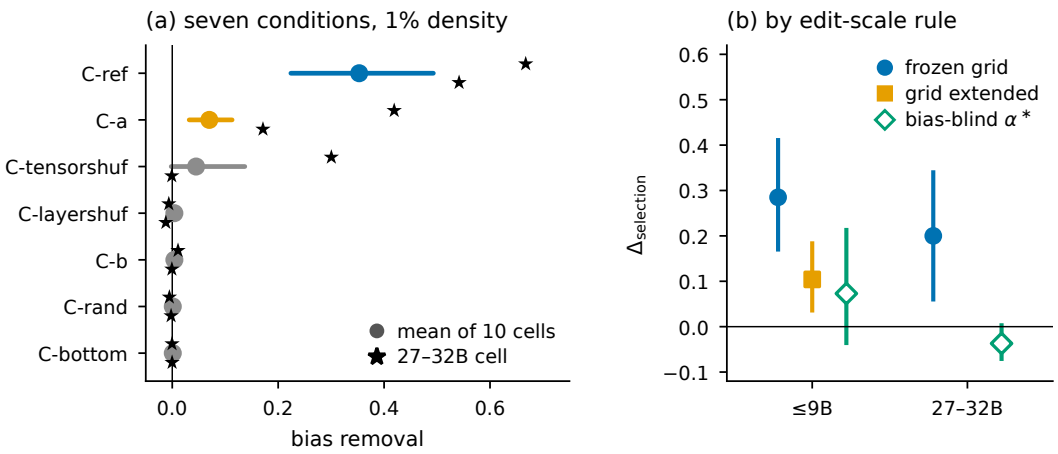


Figure 1: **Where the information lives.** (a) Seven conditions at 1% density with identical scale construction and 4-α selection: means and cell-bootstrap CIs over the 10 pre-registered cells (C-tensorshuf $n = 7$), with the two 27–32B cells as stars (one seed each). (b) The selection effect $\Delta_{\text{selection}} = C\text{-}ref - C\text{-}a$ under three edit-scale rules: the frozen deployment grid, the grid extended until neither arm clears the budget (≤9B only), and the bias-blind budget-limited rule α^*. ≤9B is the 10 registered cells; 27–32B is the three-seed four-cell panel.

Table 1: The decomposition at 1% density (cell means). [a] $n = 7$: phi's fused `qkv_proj` admits no projection shuffle.

condition	support	signs	removal
C-ref	top-$\|\Delta\|$	true, at those coords	**+0.353**
C-a	random	true, at those coords	**+0.070**
C-tensorshuf[a]	top-$\|\Delta\|$	patterns moved across projections	+0.045
C-layershuf	top-$\|\Delta\|$	patterns moved across layers	+0.004
C-b	top-$\|\Delta\|$	sign values permuted	+0.004
C-rand	random	random	+0.001
C-bottom	bottom-$\|\Delta\|$	true, at those coords	+0.001

5.3 The decomposition—where the information lives

At the dense tier, sign necessity holds at the cell unit (paired $\Delta = +0.135$ [+0.049, +0.213], 44 cells). The factorial in Table 1 and Figure 1 asks the finer question at 1% density, identical scale construction and 4-α selection, on 10 pre-registered in-envelope cells across three benchmark families, four axes, and two tiers (one 7B cell), and then on a 20-cell scale-up that adds eight in-envelope cells and two 27–32B cells (four benchmark families, three tiers); construction details in App. B, and a first-order account of why sign and magnitude selection should behave this way in App. H.

$\Delta_{\text{selection}} = +0.284$ **[+0.169, +0.407] on the ten pre-registered cells, excludes 0, unanimous 10/10;** $+0.201$ **[+0.125, +0.284] on the 20-cell scale-up, excludes 0, positive in 19 of 20.** The registered estimand reproduces bit-identically under a reinstalled stack; the scale-up magnitude is lower because weaker cells have less effect to decompose ($\text{corr}(C\text{-}ref, \Delta_{\text{selection}}) = +0.86$ over 20 cells, descriptive only), and a separate three-seed panel at 27–32B (gemma-2-27B and qwen2.5-32B on occupation→gender and bbq_Age) gives $+0.200$ [+0.060, +0.340] ($n = 4$ cells). Sign identity ($\Delta_{\text{sign}} = +0.350$ [+0.220, +0.490]), pairing (C-b ≈ 0), layer placement ($\Delta_{\text{location}} = +0.349$ [+0.221, +0.488]), and projection placement ($\Delta_{\text{tensor}} = +0.381$ [+0.249, +0.520], $n = 7$) are each necessary on average; bottom-magnitude support with true signs is dead (+0.001), so magnitude ranking is informative in both directions. Projection identity is necessary in every family but one: tensor shuffling collapses to at most 7% of C-ref in 12 non-gemma cells, while every gemma cell retains a large share (46% at 2.6B and 45% at 27B on occupation→gender)—a family property, not a size effect, and the reason the C-tensorshuf per-condition interval covers zero (App. A).

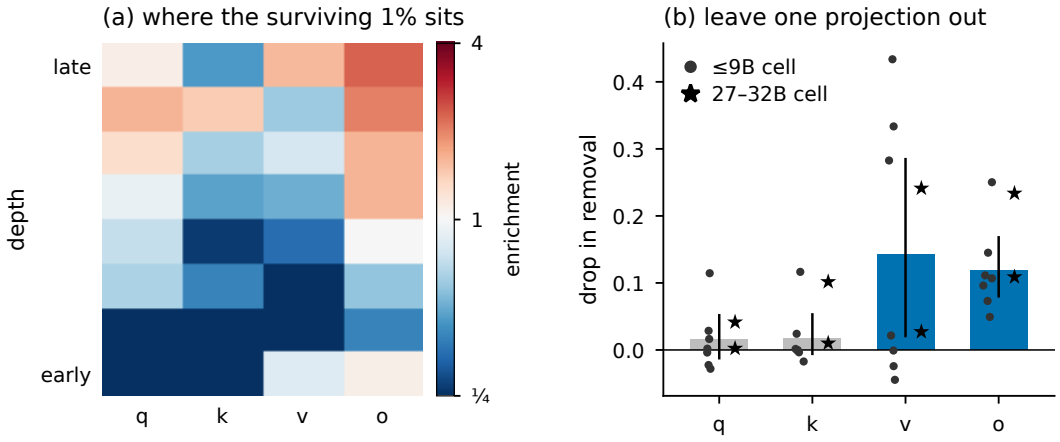


Figure 2: (a) Fold enrichment of the surviving 1% over a hypergeometric null, by depth and projection, pooled over the 27 non-fused ≤9B support dumps; pooled density multipliers are o 1.32×, q 0.83×, v 0.64×, k 0.49×. (b) Drop in removal when one projection is removed from C-ref: bars and CIs over the 7 four-projection ≤9B cells (blue: CI excludes 0), dots for those cells, stars for the two 27–32B cells (one seed each). Density is not importance: v is sparser than q and load-bearing.

Two registered qualifiers, stated as claims. *Concentrated, not exclusive:* C-a − C-rand = +0.068 [+0.030, +0.112], excludes 0; the true sign field at random coordinates carries about a fifth of the effect—redundant in the direction DARE predicts, concentrated in the direction it does not. *Concentration is a property of the edit scale, measured three ways.* Under the frozen grid $\{2, 4, 8, 16\}$, C-a is grid-limited (in-budget argmax at $\alpha = 16$ in 27 of 30 runs, spending almost no collateral) and C-ref is collateral-limited. A post-hoc extension of the same argmax rule lets the random field recover most of the effect (C-a +0.070 → +0.284; C-ref +0.353 → +0.388); the gap shrinks monotonically (+0.233 at $\alpha \leq 32$, +0.146 at ≤ 64) and saturates at +0.104 [+0.036, +0.183], since at $\alpha = 256$ neither arm clears the budget. A registered adaptive rule α^*, which grows each arm's edit scale to its first budget failure on validation items and never reads a bias probe (App. B), places the random field's budget limit at about four times the scale of the magnitude-selected field's (median 90.5 vs. 20.7 at ≤9B, 83.0 vs. 20.7 at 27–32B) and shrinks the gap to +0.073 [−0.036, +0.213] on the ten registered cells and −0.037 [−0.071, +0.003] on four 27–32B cells, both covering zero. The registered closure test at ≤9B is inconclusive, and the between-tier contrast licenses no scale claim. The three rules agree on one reading: **magnitude selection buys the effect where the signal is cheapest**, at about a quarter of the edit scale and a fraction of the collateral, the regime a deployed edit lives in. The frozen-grid estimand is the registered one; the grid extension is post-hoc, and the cell set was pre-selected in-envelope.

5.4 Which projections carry it

Leave-one-projection-out on the seven four-projection cells (Fig. 2b): dropping `q_proj` (+0.015 [−0.012, +0.052]) or `k_proj` (+0.017 [−0.006, +0.053]) costs nothing detectable, while `v_proj` (+0.143 [+0.021, +0.285]) and `o_proj` (+0.119 [+0.080, +0.168]) are load-bearing—the minimal edit is **v + o**, which compounds with sparsity and is a more robust route to a small patch than pushing density below 1%. `v_proj`'s load is carried by the templated occupation cells (positive in 4 of 7); `o_proj` is positive in all 7. At 27–32B the load-bearing projection is family-specific (`v_proj` for gemma-2-27B, `o_proj` for qwen2.5-32B; App. B). Two guard-rails: phi is the counterexample (its fused-qkv edit never touches `o_proj` and produces headline numbers, so o is load-bearing within the q,k,v,o parameterization, not universally), and **density is not importance**—`v_proj` holds fewer surviving coordinates (0.64×) than `q_proj` (0.83×) yet is load-bearing, while `o_proj`, the densest (1.32×), is also load-bearing; the support heatmap (Fig. 2a) must not be read as an importance map.

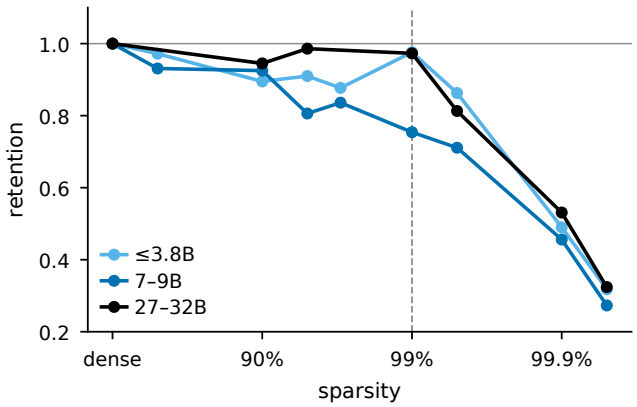


Figure 3: Retention of the dense one-bit removal as coordinates are removed (top-$|\Delta W|$ support, true signs, frozen grid), on a panel that adds bbq_Age to the occupation cells: $\leq$3.8B ($n = 7$ cells), 7–9B ($n = 4$), 27–32B ($n = 4$); ratio of cell means (Table 7). Dashed line: the deployed 1% density.

5.5 Binarization, sparsity, size—and sparsity as deployability

Binarization (paired on identical ΔW): retention 86.8% at $\leq$3.8B (deficit -0.040, $n = 8$ cells, covers 0) and 98.5% at 7–9B (deficit -0.005 [-0.010, -0.002], $n = 4$, excludes 0). At 7–9B the full-precision arm stands at $+0.328$ against the binary arm's $+0.322$: nearly free, not free. **Sparsity, $\leq$3.8B ($n = 4$):** 90% free ($+0.067$ [-0.031, $+0.182$]); 99% shows no measurable cost (-0.075 [-0.167, $+0.016$]); the cliff sits at 99.5% (-0.226 [-0.285, -0.155]) and 99.9% (-0.417 [-0.513, -0.331]). **Sparsity, 7–9B ($n = 2$):** under the registered nan$\rightarrow$0 scoring, 95–99.5% sparsity outscores dense (99%: $+0.112$ [$+0.013$, $+0.210$]; 99.5%: $+0.131$ [$+0.011$, $+0.251$]), but the margin is the one dense seed that fails the budget and scores 0—with it dropped the 99% difference is -0.010 [-0.033, $+0.013$], covers 0. The benefit at this tier is deployability, not removal quality; 99.9% is the cliff under either scoring (-0.074 [-0.121, -0.027]). **Sparsity removes the residual budget failures.** The only 7–9B configurations that fail the budget are one llama-8B seed at $\leq 90\%$ sparsity; the same edit clears at $\geq 95\%$, and **at $\geq 95\%$ sparsity no configuration in either tier fails**. Dense 7–9B edits deploy in five of six (target, seed) pairs; sparsification closes the sixth. **Across three tiers.** A broader sparsity panel that adds bbq_Age cells to the occupation cells gives, relative to the dense one-bit edit, retention at 1% density of 0.976 at $\leq$3.8B ($n = 7$), 0.754 at 7–9B ($n = 4$), and 0.973 at 27–32B ($n = 4$), with the same cliff at every tier (0.46–0.53 at 99.9%; Fig. 3); binarization at 27–32B retains 1.054 of full precision (binary $-$ fp $+0.021$ [-0.006, $+0.074$], covers 0). Per cell at 27–32B, 1% keeps 0.945–1.217 of the dense one-bit removal in three cells and 0.557 on gemma-2-27B|bbq_Age, which needs about 15% density to keep 90%. A pre-registered rule for predicting the needed density from the geometry of $|\Delta W|$ did no better than the constant in calibration and failed held-out validation (App. B), so 1% is used throughout. **Size.** The dense one-bit delta is 40.5–160.0 MiB; the 1% payload is 0.41–1.60 MiB; the support index costs $\sim$8 bits per retained coordinate at 1% density, so the shippable patch is 3.7–14.5 MiB entropy-coded (13–53 with 32-bit indices), versus 12–26 MiB for a rank-16 LoRA adapter (per-target table in App. A; all sizes MiB).

5.6 What the surviving support is

Overlap is reported as fold enrichment over a hypergeometric null (raw Jaccard at 1% density is uninformative). Across seeds the support is stable ($11.15\times$, signed agreement 0.953, $n = 36$); across designers on the one measured cell, more so ($14.64\times$, $n = 9$). Across axes the overlap is **bimodal and the split is semantic**: bbq_Age $\times$ bbq_Race share their sign field at designer-level strength (median signed agreement 0.765) while every other axis pair sits at chance (median 0.490): **the sign field tracks corpus content, not axis identity.** Support concentrates in late layers ($0.47\times$ early $\rightarrow$ $1.65\times$ late; Fig. 2a). The designer and axis arms use disjoint targets, so no within-target dissociation is claimed (method and caveats in App. C).

Table 2: **Table A—behavioral performance** on the eight frontier cells (always-on, integer-item budget, equal selection space; every configuration's collateral persisted). Removal and IFEval change carry cell-bootstrap CIs; ΔMMLU (items of 200) and perplexity ratio are medians over selected configurations; "in budget" is the share of candidate configurations within the budget.

method	removal	ΔMMLU	ppl	in budget	IFEval Δ
edit	+0.282 [+0.096, +0.487]	0	1.026	58.3%	−0.010 [−0.026, +0.007]
steering	+0.320 [+0.096, +0.542]	1	1.023	94.8%	+0.001 [−0.014, +0.018]
SentenceDebias	+0.326 [+0.113, +0.550]	0	1.006	78.1%	−0.002 [−0.012, +0.009]
DPO	+0.006 [−0.064, +0.095]	—	—	—	—

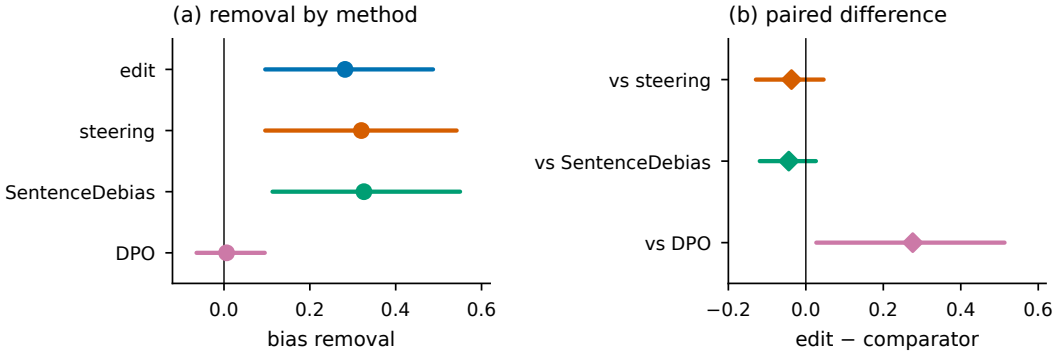


Figure 4: (a) Removal on the eight frontier cells for the edit, activation steering, SentenceDebias, and DPO, all driven by the same elicited signal at matched collateral and selection space (cell-bootstrap CIs; Table 2). (b) Edit minus each comparator, paired over cells.

5.7 The frontier—removal ties; the artifact differs

edit − steering = −0.037 **[**−0.129**,** +0.046**]; edit − SentenceDebias** = −0.044 **[**−0.119**,** +0.026**]; edit − DPO** = +0.276 **[**+0.027**,** +0.513**].** The tie was the pre-registered branch of this comparison and holds on this complete re-measurement (Table 2, Fig. 4); DPO removes no detectable bias on this signal.

Method notes. INLP fails its own random-subspace null (+0.025 [−0.070, +0.150]) and is the most collaterally expensive method tested; reported, not omitted. **DPO's failure is a signal–method mismatch:** the self-elicited corpus returns identical strings in both arms on 46–92% of pairs, the differing fraction correlates +0.82 with `contrast_gap` (Pearson, $n = 66$ corpus cells; App. E), and a direction contrast can use a sparse signal where a preference objective cannot. **Prompting** fails the always-on budget in 16 of 18 conditions; where it removes bias it pays in perplexity (ratio 1.28), and a mild fairness prompt *increases* occupational skew on all three targets (App. D). **Generation-side collateral:** on three 7–9B instruct targets and two axes (six cells, one seed), IFEval (Zhou et al., 2023) under the matched-norm contrast shows no instruction-following cost: sign − random ranges from −0.019 to +0.006 (mean −0.007), at most six instructions of 318, at removals of +0.128 to +0.810 (App. D).

5.8 Operating characterization—and a gate that does not validate

Removal tracks pre-existing bias (Pearson $r = +0.92$ [+0.82, +0.97] over the 31 measured cells across four benchmark families; +0.93 on the 8 BBQ cells), and below it the edit **backfires**: removal of −0.08 to −0.15 on three of the four lowest-pre-skew cells (Fig. 5a). An abstention gate on pre-skew and `contrast_gap` calibrates exactly as theorized in-sample (backfire 0.18 → 0.00, utility +0.135 → +0.145, rejecting 4 of 17 cells) and **does not survive held-out validation**: on the held-out split ($n = 14$) the pre-skew gate rejects no cell and the contrast-gap gate rejects one and costs utility (+0.188 → +0.179); across 200 alternative splits the gate's mean gain is −0.0051. The gate is not hidden by an unlucky split—it is not there, in a cell population where editing everything is already near-optimal. We report this as a pre-registered negative result. `contrast_gap` predicts

Table 3: **Table B—deployment properties**: what a tie makes decisive. Steering and projection win on bytes by two to three orders of magnitude; the edit's case is persistence.

property	edit (this paper)	steering	SentenceDebias
persists after merge; no serving-time hook; no per-token compute; no activation access	**yes**	no	no
bytes that must ship (MiB, 2.6–8B; dense one-bit delta 40.5–160.0)	3.7–14.5 entropy-coded; 13–53 with 32-bit indices	0.008–0.016	0.008–0.016
removes the residual budget failure at 8B	yes (§5.5)	n/a	n/a
reversible	subtract patch	remove hook	remove hook

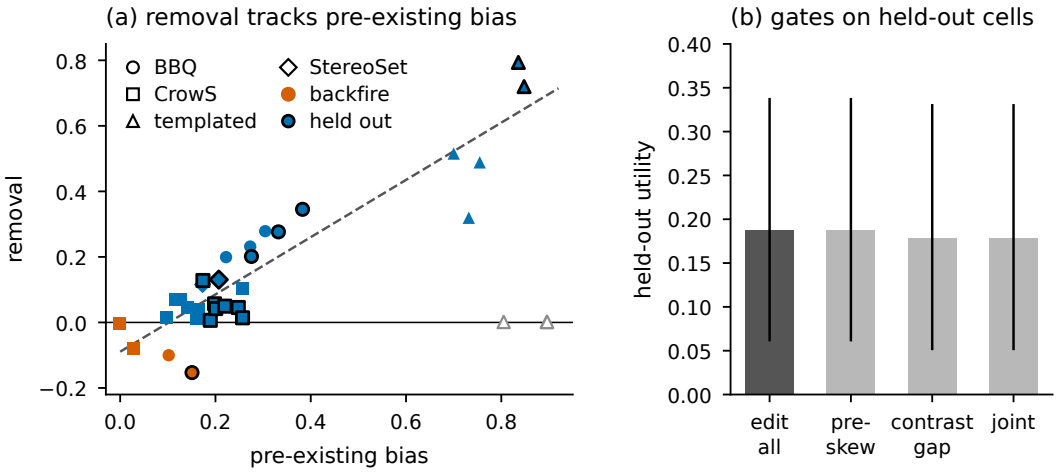


Figure 5: (a) Removal versus pre-existing bias over the 34 self-designer cells: 31 measured, and 3 whose every run fails the budget (hollow, excluded from the fit). Marker shape is the benchmark family, orange marks backfire (removal < 0), and black outlines mark the held-out split. Dashed: least-squares line over the 31 measured cells ($r = +0.92$ [$+0.82$, $+0.97$]). (b) Held-out utility (abstention scored 0, $n = 14$) for editing every cell and for the three gates; none beats editing everything, although the joint gate removes every backfire in calibration ($n = 17$).

removal only population-dependently (cell-level correlation from -0.018 to $+0.981$ across eleven defensible populations; $+0.305$ [-0.091, $+0.694$], covering zero, on the self-designer population the envelope uses; App. E), so we make no predictor claim.

6 Limitations

Granularity. MMLU is scored on 200 items against a 4-item budget; 1,205 probes (15.1% of 7,960) sit within one item of the boundary. The integer-item gate removes the knife-edge, but the budget remains $4\times$ the metric's granularity, and MMLU moves by one or two items across processes at identical weights (App. G), so decisions within two items of the boundary are not reproducible run to run; this bounds every gate-selected configuration, α^* included. Re-measuring selected configurations on 1,000 MMLU items rejects 1 of 60 decomposition points, none of 69 frontier points, and 7 of 48 at 27–32B.

Grid-conditional concentration. $\Delta_{\text{selection}}$ excludes zero at the deployment grid at every tier and after a saturating grid extension, but covers zero under the bias-blind rule α^* at both tiers tested; the registered closure test at $\leq$9B is inconclusive, and no difference between tiers is licensed. **Instruction-tuned throughout.** Every panel uses instruction-tuned checkpoints (a correction to earlier drafts of this project). One base-model measurement exists: with a corpus elicited from its instruction-tuned sibling and both checkpoints prompted in the same plain-text format, the base model is editable

($+0.605$ vs. a matched-norm null of $+0.000$), so the phenomenon is not an instruct artifact. No base-vs-instruct *relative* claim is licensed: in that format the base model's MMLU sits below chance, leaving it an effectively perplexity-only budget, and the format change alone moves the instruct model's MMLU by -0.175. **Envelope.** Described and calibrated, not validated out-of-sample; population-scoped. **Baselines and scope.** PCGU (Yu et al., 2023) and FairLoRA (Sukumaran et al., 2024) are cited, not reimplemented (legacy stack; vision-only formulation), and counterfactual data augmentation and dropout debiasing (Webster et al., 2020) are training-time interventions that cannot be placed on a frontier defined by one elicited signal and four configurations—the frontier spans six method families. Sample sizes: DEC $n = 10$ registered cells and $n = 20$ in the scale-up, whose two 27–32B cells are one seed and one axis each, plus a three-seed 27–32B panel of four cells; SPC 7–9B $n = 2$; frontier $n = 8$; IFEval 3 targets $\times$ 2 axes $\times$ 1 seed. English probes, social-bias axes; 2.6–32B for the decomposition and 2.6–8B elsewhere. Llama-3.1-70B was not run: it does not fit one device unquantized, and 8-bit loading would confound base-weight precision with the edit structure under study. Reproducibility and compute are detailed in App. G.

7 Conclusion

Strip a behavioral task vector of its magnitudes and 99% of its coordinates and the effect survives; perturb its learned signs, their pairing, or their layers and it dies; move the true sign field to random coordinates and, at the deployment scale, a fifth survives. Behavioral edit information is concentrated in a magnitude-identifiable sparse signed substructure, though not exclusively: an adaptive rule that grows each edit's scale to a held-out collateral budget lets a random 1% of the true signs come close at about four times the scale, so magnitude selection is what makes the edit cheap. The edit deploys as a persistent intervention, merged into the checkpoint with no serving-time hook; at matched collateral its removal ties steering and projection built from the same signal, while DPO on that signal removes none and prompting fails the budget. The decomposition travels beyond debiasing, since any behavioral delta can be asked the same seven questions, and integer-item budgets, always-on collateral, equal selection spaces, per-method nulls, and pre-registered forks should be the default conditions under which this literature compares interventions.

AI use statement

Generative AI tools were used in this work for tasks with required disclosure: research ideation and experiment planning (including the pre-registration texts and the adversarial audit plan), code drafting and review, and writing and editing of the manuscript. They were not used to generate, fabricate, or select experimental results: every number in the paper is produced by released scripts from measured artifacts and verified by a machine-checked manifest (App. G), and all LLM-drafted code was verified by unit tests, exhaustive gate checks, and bit-determinism reruns. Language models are also the *subjects* of the study and generate the elicitation signal used by every method (§3); that is the method under study, not a tool. We have reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

The intended use is bias reduction in publicly released language models, evaluated on public benchmarks; no human-subject data was collected. The paper's operating characterization is itself a safety finding: below the envelope the edit *increases* bias, and the abstention gate meant to prevent that does not validate, so deployment should be gated on measured pre-existing bias per axis rather than on a rule. The mechanism is dual-use—a sign-only negation can be reversed to *induce* a behavior, and a merged patch is harder to detect than a serving-time hook—which is why the release includes the per-method nulls and the always-on collateral protocol that make such edits measurable, and why the patch is reversible by subtraction.

Reproducibility statement

The evaluation protocol and its rationale are in §4 and App. A; the construction of every condition and baseline is in Apps. B–D; axis onboarding, module resolution, and asset licenses are in App. F; determinism scope, compute, and the number-source manifest are in App. G. Code, the elicited

corpora, every edit with its per-α trace, the audit manifest, and the figure data CSVs are provided as anonymized supplementary material; every figure is generated from an audited CSV by the released script, every number in the paper maps to an artifact through the manifest, and the registered decomposition cells replay bit-identically under a reinstalled software stack.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. arXiv:2404.14219.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 5454–5476, 2020.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706, 2022.

Omkar Dige, Diljot Arneja, Tsz Fung Yau, Qixuan Zhang, Mohammad Bolandraftar, Xiaodan Zhu, and Faiza Khan Khattak. Can machine unlearning reduce social bias in language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 954–969, 2024.

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

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1926–1940, 2021.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 609–614, 2019.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR 2021)*, 2021. arXiv:2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR 2022)*, 2022. arXiv:2106.09685.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR 2023)*, 2023. arXiv:2212.04089.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 5502–5515, 2020.

James Liu, Guangxuan Xiao, Kai Li, Jason D. Lee, Song Han, Tri Dao, and Tianle Cai. BitDelta: Your fine-tune may only be worth one bit. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2402.10193.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1878–1898, 2022.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, 2021.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 1953–1967, 2020.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. arXiv:2305.18290.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 7237–7256, 2020.

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 (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

Rohan Sukumaran, Aarash Feizi, Adriana Romero-Soriano, and Golnoosh Farnadi. FairLoRA: Unpacking bias mitigation in vision models with fairness-driven low-rank adaptation, 2024. arXiv:2410.17358.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. arXiv:2308.10248.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models, 2020. arXiv:2010.06032.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. arXiv:2306.01708.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6032–6048, 2023.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024. arXiv:2311.03099.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. arXiv:2311.07911.

A PROTOCOL: RATIONALE AND DETAILS

Why the budget is measured always-on. A deployed edit, steering hook, projection, or system prompt is active on every request, so its capability cost is only meaningful if measured with the intervention active on the capability probes themselves. Evaluating bias with the intervention on and capability with it off—the common practice—credits an intervention with removal it can only achieve at a cost it never pays. Every method in this paper is therefore scored with the intervention active during the bias probes, the MMLU items, and the perplexity corpus alike. This is what exposes prompting: a fairness instruction that is prepended to MMLU items changes the answers.

Why the budget is an integer-item criterion. MMLU is scored on 200 items, so accuracy moves in steps of $1/200$ and a 4-item drop is exactly the 0.02 budget. A floating-point comparison of the form $(\text{pre} - \text{post}) \leq 0.02$ would reject exactly that boundary case ($0.545 - 0.525$ evaluates to 0.020000000000000018), and on a 200-item metric the boundary is not rare. We therefore compare in integer items, $(\text{pre}_{\text{items}} - \text{post}_{\text{items}}) \leq 4$, which decides the boundary exactly with no tolerance to tune; the perplexity ratio ≤ 1.10 keeps a float comparison with an explicit 10^{-9} tolerance. One shared implementation with 12 unit checks is used by every panel, and every persisted α trace is decided by it.

Why the selection space is equalized. A method tuned over more configurations gets more chances to land in budget. Every method in the frontier optimizes over exactly four configurations: the edit's four values of α; steering's two layers $\times$ two strengths; SentenceDebias's two layers $\times$ two ranks; DPO's four merge scales; prompting's three strengths plus none. A cell with no in-budget configuration scores 0 and is recorded as such, never imputed.

Why the resampling unit is the cell. Seeds within a target$\times$axis cell are not independent replicates of the phenomenon; treating them as such (pooling per-seed or per-sweep rows) shrinks intervals anti-conservatively. Every interval in this paper is a 10,000-resample percentile bootstrap over target$\times$axis cells, with n reported as the number of cells.

Nulls. Each method must beat a structure null that preserves everything but its purported signal—sign-shuffle for the edit (same scale, same support, random signs), a random direction at matched norm for steering, a random subspace of matched rank for projections—and the partition null, which runs the identical pipeline on two halves of the *same* biased corpus so that no biased-versus-debiased direction exists. A method that cannot beat its own nulls does not enter the frontier as a competitor (INLP, §5.7).

Per-condition intervals for Table 1. C-ref $+0.353$ $[+0.224, +0.493]$; C-a $+0.070$ $[+0.032, +0.113]$; C-tensorshuf $+0.045$ $[-0.002, +0.137]$ (covers 0; carried by gemma); C-layershuf $+0.004$ $[+0.002, +0.008]$; C-b $+0.004$ $[+0.002, +0.006]$; C-rand $+0.001$ $[+0.001, +0.002]$; C-bottom $+0.001$ $[+0.001, +0.002]$.



Figure 6: The information-removal pipeline. A full-precision contrast vector (bf16, 648–2,560 MiB over the edited tensors) is reduced to one bit per weight plus one scale per tensor (40.5–160.0 MiB), then to 1% of coordinates by magnitude (payload 0.41–1.60 MiB; 3.7–14.5 MiB once an entropy-coded support index is charged); the reduced edit still removes $+0.353$ $[+0.22, +0.49]$ of bias in 10/10 cells and merges into the checkpoint. Sizes are for 2.6–8B models.

Table 4: Removal against both pre-registered nulls: four designer×cell examples spanning three cells (pooled over those three alone: $+0.510$ and $+0.461$). The partition null is not vacuous—halves of a biased corpus can still yield a weakly usable direction on some cells—so both nulls are always shown, unaveraged.

cell	designer	real	sign-shuffle	partition
phi | occ_gender	self	$+0.316$	$+0.002$	$+0.014$
phi | occ_gender	cross-large	$+0.646$	$+0.002$	$+0.032$
qwen | crows_socio	cross-large	$+0.107$	$+0.002$	$+0.003$
gemma | occ_gender	same-large	$+0.842$	$+0.007$	$+0.124$

Table 5: Patch sizes in MiB (bytes/2^{20}), the unit under which this table reproduces. The payload is one bit per retained coordinate; the entropy-coded support index adds about 8 bits per retained coordinate at 1% density (App. H, Proposition 4). A bitmap costs exactly the dense payload and is never better; a plain 32-bit index costs 13–53 MiB.

target	dense one-bit	payload @ $s = 0.99$	payload + index (entropy floor)	LoRA r16 (fp16)
qwen-3B	40.50	0.41	**3.68**	14.06
gemma-2.6B	43.88	0.44	**3.98**	12.19
llama-3.2B	84.00	0.84	**7.63**	17.50
phi-3.8B	108.00	1.08	**9.81**	12.00
qwen-7B	98.00	0.98	**8.90**	19.25
llama-8B	160.00	1.60	**14.53**	26.00

B THE DECOMPOSITION: CONSTRUCTION AND DESIGN

What is being decomposed. A trained contrast vector ΔW (rank-16 LoRA product, materialized dense) is reduced to a k-sparse sign field E with $k = 1\%$ of the edited coordinates and one scale per tensor. Three kinds of information could carry the effect: *which* coordinates are in the support, *which sign* each carries, and *where* (which layer, which projection) the support sits. The seven conditions vary these one at a time while holding density, scale, and the four-value α selection fixed. All seven edit exactly the same number of coordinates, and every condition uses the per-tensor scale of C-ref. A condition that took its scale from the mean $|\Delta W|$ over its own support would give C-a a several-fold smaller edit norm, so it would lose on norm rather than on support information; matching the scale is the DARE-faithful convention, since DARE rescales survivors for the same reason. Under the per-support convention, run on two cells as a robustness arm, $\Delta_{\text{selection}}$ is larger.

Conditions. C-ref takes the top-k coordinates by $|\Delta W|$ with their true signs (the deployed edit). C-a draws a uniformly random k-subset of coordinates and evaluates the *true* sign field there, $\text{sign}(\Delta W_i)$ for i in the random support; it asks whether the sign field is portable, the condition DARE-style random dropping predicts should work. C-b keeps the top-k support but permutes the sign *values* within each tensor, destroying the coordinate–sign pairing while preserving the sign histogram. C-rand draws a random support with random signs (the complete null). C-bottom takes the k smallest-magnitude nonzero coordinates with their true signs, testing whether magnitude ranking is informative in both directions. C-layershuf keeps the top-k support globally but permutes each layer's sign pattern to another shape-compatible layer; C-tensorshuf permutes patterns across the shape-compatible projections within a layer. Permutation maps are logged per run. Phi's fused `qkv_proj` admits no projection shuffle, so its three cells carry no C-tensorshuf row ($n = 7$ for that condition). Random supports are drawn by exact multivariate-hypergeometric allocation over tensors followed by sampling without replacement within each tensor, so k is fixed exactly.

Two readings of "random support with learned signs." C-a evaluates the true sign field at random coordinates; C-b relocates sign values off their coordinates. They answer different questions and are both run. C-a is the scientifically loaded one: it is the condition under which the sign field would be redundant across coordinates. At the deployment grid the data place it at a fifth of C-ref (concentrated, not exclusive), while C-b sits at the null.

Cells. Ten in-envelope cells were pre-registered before any condition ran, all with the self designer: gemma, llama, qwen, and phi on occupation→gender; phi and qwen on bbq_Age; qwen on bbq_Race_ethnicity; qwen and phi on StereoSet intrasentence; and qwen-7B on occupation→gender (three benchmark families, four axes, two tiers). Backfire cells were excluded because a decomposition of an effect that is negative in sign is uninterpretable. Three seeds per cell.

Scale-up. After the registered panel was reported, three reportings were fixed after the first of the panel's 54 runs had finished and before the remaining 53: the registered ten cells; the ten plus every new cell that passed a C-ref-only screen (the screen reads C-ref, never $\Delta_{\text{selection}}$); and all cells. Eight new cells at 2.6–8B (a fourth benchmark family, CrowS, enters) and two cells at 27–32B (gemma-2-27B and qwen2.5-32B, self-designer, one seed, occupation→gender, frozen grid) were then run; every new cell passed the screen, so the second and third reportings coincide. Every published cell was replayed alongside the new ones and reproduced bit-identically. Table 6 lists the 20 cells.

Table 6: The 20-cell scale-up, sorted by C-ref (cell means; $\Delta_{\text{selection}} = C\text{-}ref - C\text{-}a$). Pooled $\Delta_{\text{selection}} = +0.201$ [$+0.125$, $+0.284$], positive in 19 of 20. The two 27–32B cells are one seed each.

cell	C-ref	C-a	$\Delta_{\text{selection}}$
llama-8B | occ_gender	$+0.698$	$+0.465$	$+0.233$
gemma-2.6B | occ_gender	$+0.693$	$+0.095$	$+0.598$
gemma-27B | occ_gender	$+0.667$	$+0.419$	$+0.248$
qwen-7B | occ_gender	$+0.663$	$+0.214$	$+0.450$
llama-3.2B | occ_gender	$+0.615$	$+0.043$	$+0.572$
qwen-32B | occ_gender	$+0.542$	$+0.171$	$+0.370$
qwen-3B | occ_gender	$+0.391$	-0.011	$+0.402$
phi-3.8B | bbq_Age	$+0.331$	$+0.078$	$+0.253$
qwen-7B | bbq_Age	$+0.324$	$+0.192$	$+0.132$
qwen-3B | bbq_Age	$+0.253$	$+0.133$	$+0.120$
qwen-3B | bbq_Race_ethnicity	$+0.253$	$+0.107$	$+0.146$
gemma-2.6B | bbq_Age	$+0.171$	$+0.095$	$+0.076$
qwen-3B | ss_intra	$+0.117$	$+0.014$	$+0.103$
phi-3.8B | occ_gender	$+0.116$	$+0.014$	$+0.102$
phi-3.8B | ss_intra	$+0.100$	$+0.011$	$+0.090$
gemma-2.6B | crows_socio	$+0.084$	$+0.045$	$+0.039$
llama-3.2B | ss_intra	$+0.079$	$+0.015$	$+0.064$
phi-3.8B | crows_socio	$+0.029$	$+0.019$	$+0.011$
llama-8B | bbq_Age	$+0.023$	$+0.026$	-0.003
llama-3.2B | crows_socio	$+0.011$	$+0.006$	$+0.005$

At 27–32B the full condition set behaves as at small scale: for gemma-2-27B the seven conditions read $+0.667$ (C-ref), $+0.419$ (C-a), $+0.300$ (C-tensorshuf), and ≈ 0 for C-b, C-rand, C-layershuf, and C-bottom; for qwen2.5-32B they read $+0.542$, $+0.171$, and ≈ 0 for all five perturbations including C-tensorshuf. Both large models were grid-limited rather than budget-limited at the top of the frozen grid (gemma-2-27B's C-ref fails the budget at $\alpha = 16$ with perplexity ratio 1.55), so their $\Delta_{\text{selection}}$ values are deployment-grid values like the rest of the panel. Leave-one-projection-out at this tier: gemma-2-27B $+0.626$/$+0.566$/$+0.426$/$+0.559$ and qwen2.5-32B $+0.539$/$+0.532$/$+0.515$/$+0.308$ without `q_proj`/`k_proj`/`v_proj`/`o_proj` respectively.

An adaptive, bias-blind edit scale. A registered rule α^* takes the first budget failure on the ladder $\{1, 2, 4, \ldots, 512\}$, refined by three log-bisection steps, with the integer-item gate applied to 200 MMLU validation items and 20 WikiText validation chunks and no bias probe read; the removal reported is the as-deployed removal at α^*. It was run on the ten registered $\leq$9B cells and on a separate three-seed 27–32B panel (gemma-2-27B and qwen2.5-32B on occupation→gender and bbq_Age, self designer), where the deployment grid gives $\Delta_{\text{selection}} = +0.200$ [$+0.060$, $+0.340$] ($n = 4$). The random field's budget binds at about four times the scale of the magnitude-selected field's (median α^* 90.5 vs. 20.7 at $\leq$9B, 83.0 vs. 20.7 at 27–32B), and $\Delta_{\text{selection}}$ shrinks from $+0.285$ [$+0.170$, $+0.411$] to $+0.073$ [-0.036, $+0.213$] at $\leq$9B and from $+0.200$ to -0.037 [-0.071, $+0.003$] at 27–32B. The registered closure criterion (α^* gap at most a quarter of the frozen-grid gap) is missed at $\leq$9B by

0.0014 (ratio 0.255), a margin that MMLU's run-to-run variation (App. G) could move to either side, so the registered verdict is inconclusive. Scoring runs that fail the evaluation budget as zero ($+0.072$ [-0.072, $+0.240$]) or restricting $\leq$9B to the axes the large panel uses ($+0.104$ [-0.043, $+0.301$]) does not change it, and the $\leq$9B minus 27–32B contrast, $+0.109$ [-0.008, $+0.260$], licenses no scale claim. The validation-chosen α^* passes the budget on evaluation items in 57% (C-ref) and 60% (C-a) of $\leq$9B runs and 83% and 92% of 27–32B runs; the extension above instead selects only among configurations that pass on the evaluation items themselves, which is why the two scale-extending rules can differ. At 27–32B, α^* lowers C-ref's removal relative to the frozen-grid argmax (-0.123 [-0.248, $+0.001$]) and raises C-a's ($+0.113$ [-0.082, $+0.326$]).

Is the 1% density predictable from the weights? Before the same panel ran, we registered a rule that predicts the density needed to keep 90% of the dense one-bit removal from the geometry of $|\Delta W|$: choose the density whose top coordinates carry the median share of $|\Delta W|$ mass that the needed density carries on eleven calibration cells at $\leq$9B, and pass if the predicted density keeps at least 85% on three of the four 27–32B cells. In leave-one-target-out calibration the fixed 1% constant beat this rule and three alternative geometry rules (median $|\log_2|$ error of the predicted density 1.59 against 1.67–1.81; target met in 64% of cells against 45–55%). The rule predicted 0.72–0.90% for every held-out cell and failed held-out validation (2 of 4). We therefore keep 1% as a constant and make no claim that the needed density can be read off the weights. On the same four cells, 1% keeps 0.557–1.217 of the dense one-bit removal (§5.5).

Table 7: Retention of the dense one-bit removal by sparsity, on a panel that adds bbq_Age cells to the occupation cells (ratio of cell-mean removals; the dense one-bit edit is 1).

sparsity	$\leq$3.8B ($n = 7$)	7–9B ($n = 4$)	27–32B ($n = 4$)
0.5	0.972	0.931	—
0.9	0.895	0.925	0.945
0.95	0.910	0.806	0.986
0.97	0.877	0.836	—
0.99	0.976	0.754	0.973
0.995	0.863	0.711	0.813
0.999	0.490	0.456	0.531
0.9995	0.318	0.273	0.324

Pre-registered adjudication. Two outcomes were written before the run: *DARE-consistent* ($\Delta_{\text{selection}}$ covers 0; the sign field is distributed and magnitude merely certifies coordinates) and *DARE-divergent* ($\Delta_{\text{selection}}$ excludes 0 with a substantial fraction of C-ref). The second fired, unanimously on the registered panel, and survives the scale-up (19 of 20) at the deployment grid.

The α-extension. The deployment grid $\{2, 4, 8, 16\}$ was frozen in pre-registration before the decomposition ran. Because C-a selected $\alpha = 16$ in 27 of 30 runs while spending almost no collateral, we ran a post-hoc extension to $\alpha \in \{32, 64, 128\}$ on the same cells, seeds, corpora, and gate; the frozen arm reproduces exactly under the extended runner. The frozen-grid estimand remains the registered one; the extension is reported as a measured qualifier. A second extension probed $\alpha \in \{256, 512\}$ under a criterion registered beforehand (stop when neither arm clears the budget): at $\alpha = 256$ no run of either arm is admissible (median MMLU drop 34 items, median perplexity ratio 1.74; at 512, 88 items and a ratio above 1,000), so the depth series is flat from 128 onward and the gap at the maximum admissible depth, $+0.104$ [$+0.036$, $+0.183$], is the saturated value.

C THE SURVIVING SUPPORT: METHOD

Fold enrichment. Two independent 1%-dense supports overlap on about 1% of either set by chance, so raw Jaccard (0.015–0.079 in our data) is uninformative. Every overlap statistic is reported as the ratio of observed overlap to its hypergeometric expectation under independent selection at the observed densities; a value of 1 means chance. Signed agreement is the fraction of shared coordinates whose signs agree. The expectation is computed both globally and within module, conditioning on each module's support size so that shared layer placement alone cannot produce enrichment; within

module the enrichment is lower but well above chance (5.10× across seeds, 8.44× across designers, 2.56× across axes). The estimator is validated by a committed self-test harness (45 checks).

Arms. *Seeds:* the three seeds of each cell (36 support dumps; 11.15× global enrichment, signed agreement 0.953). *Designers:* self, same-family-large, and cross-family designers on one cell, qwen-7B|occupation→gender (9 pairs; 14.64×; per-pair signed agreement 0.319–0.965, including one reversal—a single-cell arm). *Axes:* pairs of axes on qwen-3B and phi-3B (27 pairs; 2.96× overall, bimodal: bbq_Age×bbq_Race at median signed agreement 0.765, all other pairs at 0.490). No target appears in both the designer and the axis arm, so the two are not compared within a target.

Density and importance. Projection densities are fold-enrichment marginals over the 27 non-fused support dumps (`o_proj` 1.32×, `q_proj` 0.83×, `v_proj` 0.64×, `k_proj` 0.49×); the depth gradient is $0.47\times$ in the earliest eighth of layers to $1.65\times$ in the latest. The leave-one-projection-out panel (§5.4) rests on the 21 dumps of the seven four-projection cells and is the causal statement; the heatmap is descriptive.

D BASELINES: HOW EACH WAS BUILT FROM THE SAME SIGNAL

Steering. For each cell, the steering vector is the mean residual-stream activation on the biased half of the elicited corpus minus the mean on the debiased half, at a chosen layer; it is added to the residual stream at inference with a strength coefficient (two layers × two strengths). Its structure null is a random direction at matched norm; its partition null uses the two halves of the same biased corpus. Steering beats both nulls ($+0.297$ [$+0.180$, $+0.415$]; $+0.298$ [$+0.183$, $+0.413$]) and keeps 94.8% of its configurations in budget (median perplexity ratio 1.023), so the tie with the edit is a real contest.

SentenceDebias and INLP. Both are ported from the bias-bench implementation (Meade et al., 2022) and fitted on the same elicited corpora (two layers × two ranks); the structure null is a random subspace of matched rank. SentenceDebias beats its null ($+0.308$ [$+0.087$, $+0.531$]) and has the lowest perplexity cost of the frontier methods (median ratio 1.006; 78.1% of configurations in budget). INLP fails its null ($+0.025$ [-0.070, $+0.150$]), is worse than random on phi, and is the most collaterally expensive (median perplexity ratio 1.083; 49 of 96 configurations in budget); it is reported but does not enter the frontier as a competitor.

DPO and the identical-string phenomenon. Preference pairs are built from the elicited contrast: for each probe item the biased and debiased continuations become the rejected and chosen responses. The elicitation, however, chooses between the same two candidate continuations under the biased and the debiased framing, so when a model cannot articulate a contrast it returns the identical string in both arms. Across designers only 8–54% of pairs differ (for gemma-3B on occupation→gender, 92% of the corpus is the same text twice), and the differing fraction correlates $+0.82$ with `contrast_gap` (Pearson, $n = 66$ corpus cells: the large designers plus all 3B siblings, two DPO axes). A direction contrast extracts signal from a minority of differing items; a preference objective needs the pairs to differ. Cells with fewer than 20 informative pairs are skipped and scored 0.

Prompting. Three prompt strengths (mild, moderate, strong) on three targets (phi, qwen, gemma) and two axes; 16 of 18 conditions fail the collateral budget because the instruction is also prepended to the MMLU items and the perplexity corpus. Where a prompt removes bias it does so at perplexity ratio 1.28 (qwen|crows, $+0.176$) and 1.11 (gemma|crows, $+0.122$); the mild prompt raises occupational skew on all three targets (qwen: $0.326 \rightarrow 0.407$, removal -0.081).

IFEval. Six cells, one seed each, 318 verifiable instructions: sign-only edit minus the matched-norm random-sign null on IFEval accuracy is $+0.006$ (gemma-2-9B|occ, removal $+0.810$), $+0.003$ (qwen-7B|occ, $+0.698$), -0.003 (qwen-7B|bbq_Age, $+0.382$), -0.013 (gemma-2-9B|bbq_Age, $+0.337$), -0.019 (llama-8B|bbq_Age, $+0.128$), and -0.019 (llama-8B|occ, $+0.735$); unedited accuracies are 0.59–0.73. The random-sign null moves IFEval by a similar amount, which is why the matched-norm contrast, not the raw delta, is the reported quantity.

Baselines cited but not run. PCGU's reference implementation pins a legacy stack (`torch` 1.4, `transformers` 4.10), and FairLoRA's published formulation targets vision models with no canonical language-model implementation; a badly tuned baseline is worse than an absent one. Counterfactual data augmentation and dropout debiasing (Webster et al., 2020), both available in bias-bench, are training-time interventions with no elicited-signal analogue and no four-configuration selection space, so they cannot be placed on this frontier without changing what the frontier measures.

E OPERATING ENVELOPE IN FULL

Inventory and split. 36 cells across four benchmark families; 34 self-designer cells are plotted in Fig. 5a (31 measured; 3 with every run failing the budget). A calibration/held-out split was dealt by a hashed rule fixed before any threshold was chosen; after restricting to cells decided under the integer-item gate the held-out half is 14 cells.

Calibration objective. Thresholds τ (pre-skew) and γ (`contrast_gap`) were chosen on the calibration cells to maximize utility with abstention scored zero, subject to a cap on the backfire rate among edited cells. On calibration ($n = 17$) the joint gate ($\tau = 0.116$, $\gamma = 0.012$) raises utility from $+0.135$ to $+0.145$ and drives the backfire rate from 0.176 to 0.000, rejecting 4 cells.

Held-out result. Edit-all $+0.188$ $[+0.062, +0.337]$ ($n = 14$); the pre-skew gate rejects no held-out cell and is identical to edit-all; the contrast-gap gate and the joint gate reject one cell and score $+0.179$ $[+0.052, +0.330]$. Across 200 alternative splits the gate's mean utility gain is -0.0051. The gate is not hidden by the particular split; it is not there, in a population where editing everything is already near-optimal.

`contrast_gap` as a predictor. Across eleven defensible populations the cell-level correlation between `contrast_gap` and removal spans -0.018 (sibling designers) to $+0.981$ (cross-family small designers); on the self-designer population that the envelope analysis uses it is $+0.305$ $[-0.091, +0.694]$, covering zero. No predictor claim is made.

F AXIS ONBOARDING, MODULE RESOLUTION, AND LICENSES

Onboarding gates. Each axis passes, before any edit is trained, a polarity unit test (the probe scores a deliberately biased continuation as biased), an item floor of 40 per disjoint probe/edit half, and a per-target pre-skew measurement. BBQ axes are rebuilt as group-versus-unknown likelihood probes on ambiguous negative-polarity contexts; StereoSet intrasentence uses 1,053 probe and 1,053 edit items; all twelve BBQ target×axis combinations passed with pre-skew between $+0.10$ and $+0.38$.

Elicitation method. Free-generation elicitation fails on sub-4B models (they drop the pronoun rather than choose one) and forced-choice judging sits at chance (it measures recognition rather than exhibition); likelihood-based self-diagnosis is used throughout.

Module resolution. `resolve_targets` returns {`q_proj`, `k_proj`, `v_proj`, `o_proj`} for gemma-2, qwen2.5, and llama-3.x and {`qkv_proj`} alone for phi-3.5, whose `o_proj` is therefore never edited. Phi's headline numbers are produced through the fused projection, which is why §5.4 scopes the load-bearing role of `o_proj` to the q,k,v,o parameterization.

Assets and licenses. Models: Gemma 2 (Gemma Terms of Use), Qwen2.5 (Apache-2.0; the 3B checkpoint under the Qwen Research License), Llama 3.1/3.2 (Llama Community License), Phi-3.5 (MIT). Datasets: BBQ (CC-BY-4.0), CrowS-Pairs (CC-BY-SA-4.0), StereoSet (CC-BY-SA-4.0), MMLU (MIT), IFEval (Apache-2.0); SentenceDebias/INLP via bias-bench (MIT). All are used under their stated terms; licenses to be re-verified at camera-ready. New assets released with this paper: code, the elicited corpora, every edit and its α trace, the audit manifest, and the figure data CSVs, under a permissive license (to be specified).

G ENGINEERING, DETERMINISM, COMPUTE, AND REPRODUCIBILITY

Determinism, scoped. Sweep-pipeline bit-determinism is verified: an independent rerun reproduces 108/108 sweeps with zero differing values, including an exact match between a re-scorable twin panel pair. MMLU scoring on 200 items is stable within a process (repeated reads agree exactly; TF32 on/off agree) and sensitive to batch composition by up to 1% (batch sizes 4/6/8/16 score 131/129/131/129 on llama-3.2-3B-Instruct). Across processes at identical weights it can move by one or two items: a controlled three-process check agreed exactly, but across later panels the same unedited model read 0.660 and 0.670, while bias skew and perplexity were bit-identical across the same processes. The code path draws no random numbers and the cause is not identified; because the budget boundary is 4 items, this bounds the reproducibility of every gate-selected configuration. The one cross-panel divergence occurred at identical batch size and is attributed, stated not proven, to a `transformers` 5.14.1→4.57.1 rebuild; the environment is pinned in the release, and after a later reinstallation of the GPU stack every published decomposition cell replayed bit-identically before any new cell was added; a later re-run of the ten registered cells replayed 59 of 60 configurations bit-identically, the miss being the MMLU variation above.

Resampling unit. All statistics use the target×axis cell as the resampling unit. For sign necessity the sweep-level contrast is $+0.212$ $[+0.188, +0.236]$ over $n = 502$ sweeps against $+0.135$ $[+0.049$, $+0.213]$ at the cell unit ($n = 44$): pooling sweeps narrows the interval considerably, which is why the cell unit is the reported one.

Compute. All method panels ran on a single NVIDIA L40S (46 GB); the decomposition, support, envelope, sparsity, and instruct arms (523 measured rows, 36 support dumps) took two days; the post-hoc α-extension ran on the same GPU. The dense contrast materialization was refactored from a 34 GB to a 17.6 GB peak, verified bit-exact.

Reproducibility. Every α of every collateral search is persisted (`alpha_trace.jsonl`), so every budget decision can be recomputed without GPU time; every number in this paper maps to an artifact through a released manifest and audit script (`audit/manifest_v10.yaml`, `src/v10_audit.py`), run green at freeze; every figure is generated from an audited data CSV by `figures/make_figs.py`. Code and artifacts are provided as anonymized supplementary material and will be released publicly at camera-ready.

H A FIRST-ORDER ACCOUNT OF SIGN-ONLY, MAGNITUDE-SELECTED EDITS

This appendix gives the simplest quantitative reading of the decomposition. It is a first-order (linearized) account and makes no claim beyond what the propositions state; its role is to say why the ordering of the seven conditions, the near-free binarization, the grid-conditional concentration, and the patch-size floor are the expected shape of the phenomenon rather than accidents.

Setup. Let $B(W)$ be the bias score of the model with weights W and $g = \nabla_W B(W)$. An edit E applied as $W \leftarrow W - \alpha E$ changes the score, to first order, by $-\alpha\langle g, E\rangle$; removal is the decrease of B. The trained contrast vector ΔW is, by construction, a direction along which B decreases, so we take $\langle g, \Delta W\rangle > 0$ and, as the working approximation, $g \propto \Delta W$. Collateral is a second capability score $C(W)$ whose first-order change along E is small (the elicited direction is fitted to B, not C) and whose second-order change grows as $\alpha^2\, E^\top H_C E$ for a curvature H_C; the budget of §4 caps the admissible α through C.

Proposition 1 (sign is first-order optimal at a fixed entrywise scale). Among all E with $|E_i| \le 1$, the inner product $\langle \Delta W, E\rangle$ is maximized by $E^\star = \operatorname{sign}(\Delta W)$, with value $\|\Delta W\|_1$. *Proof.* $\langle \Delta W, E\rangle = \sum_i \Delta W_i E_i \le \sum_i |\Delta W_i|$, with equality iff $E_i = \operatorname{sign}(\Delta W_i)$ wherever $\Delta W_i \neq 0$. □ Binarization therefore preserves the entire first-order alignment with the trained direction per unit entrywise scale; what it loses is directional, $\cos(\Delta W, \operatorname{sign}\Delta W) = \|\Delta W\|_1/(\sqrt{d}\,\|\Delta W\|_2) \le 1$, which is why the measured retention is high but not exactly one (86.8% and 98.5%, §5.5).

Proposition 2 (magnitude-selected support maximizes alignment among k-sparse true-sign fields). For a support S with $|S| = k$ and $E_S = \operatorname{sign}(\Delta W)$ on S, zero elsewhere, $\langle \Delta W, E_S \rangle = \sum_{i \in S} |\Delta W_i|$. This is maximized by the top-k magnitudes (the C-ref support) and minimized by the bottom-k (C-bottom); a uniformly random S has expectation $(k/d)\,\|\Delta W\|_1$ (C-a); and independent fair signs on any support have expectation 0 (C-rand). *Proof.* Immediate from the sum, and from linearity of expectation over a uniform k-subset or over independent signs of mean zero. □

The remaining perturbations need not have zero expectation. *Sign permutation (C-b).* Let tensor t carry scale s_t on its top-k support, with $n_t^{\pm}$ positive and negative coordinates of total magnitude $A_t^{\pm}$, and let $m_t = (n_t^+ - n_t^-)/(n_t^+ + n_t^-)$ be its sign imbalance. If the signs within each tensor's support are permuted uniformly at random, each permuted sign has mean m_t, so

$$\mathbb{E}\,\langle \Delta W, E_{\mathrm{b}} \rangle = \sum_t s_t\, m_t\, (A_t^+ - A_t^-), \qquad \left|\mathbb{E}\,\langle \Delta W, E_{\mathrm{b}} \rangle\right| \le \max_t |m_t|\ \langle \Delta W, E_{\mathrm{ref}} \rangle,$$

where $\langle \Delta W, E_{\mathrm{ref}} \rangle = \sum_t s_t (A_t^+ + A_t^-)$ is the reference alignment. The expectation is zero only when every support is balanced ($m_t = 0$) or its two signs carry equal mass; otherwise the permuted edit keeps a fraction of the reference alignment bounded by the largest imbalance. *Layer and tensor shuffles.* Moving tensor t's signed pattern onto a shape-compatible tensor t' gives alignment $s_t \sum_{i \in S_t} \Delta W_i^{(t')} \operatorname{sign}(\Delta W_i^{(t)})$: the agreement of the two sign fields at matched coordinates, weighted by $|\Delta W^{(t')}|$. It vanishes only when the two fields are uncorrelated there.

Define the *concentration ratio* $\rho_k = \sum_{\text{top-}k} |\Delta W_i| \big/ \big((k/d)\|\Delta W\|_1\big) \ge 1$, the tail heaviness of $|\Delta W|$. The account predicts C-ref > C-a > C-rand ≈ C-bottom ≈ 0, which is what Table 1 shows. The measured C-b and C-layershuf (≈ 0) are consistent with near-balanced supports and with sign fields that are uncorrelated across layers; the retention of about half of the effect under the tensor shuffle in gemma, at 2.6B and at 27B, indicates sign agreement between gemma's shape-compatible projections. C-a recovers a fifth of C-ref rather than $1/\rho_k$ because α is selected per condition (C-a selects the largest admissible α).

Proposition 3 (why concentration is a deployment-grid property). Let E_{top} and E_{rand} be the top-k and random-k true-sign fields ($\|E_{\text{top}}\|_2 = \|E_{\text{rand}}\|_2 = \sqrt{k}$). To match the first-order removal of E_{top} at scale α, the random field needs scale $\alpha_{\text{rand}} = \rho_k\,\alpha$ in expectation; at that scale its second-order collateral is ρ_k^2 times larger, $\alpha_{\text{rand}}^2 E_{\text{rand}}^\top H_C E_{\text{rand}} \approx \rho_k^2\,\alpha^2 E_{\text{top}}^\top H_C E_{\text{top}}$ when the two supports see comparable curvature. *Proof.* Equate $\alpha_{\text{rand}}\,(k/d)\|\Delta W\|_1$ with $\alpha \sum_{\text{top-}k} |\Delta W_i|$ and substitute into the quadratic collateral term. □ Two consequences match §5.3. First, on a bounded scale grid the random field is *grid*-limited whenever $\rho_k \alpha$ exceeds the grid maximum (observed: C-a at $\alpha = 16$ in 27 of 30 runs, spending almost no collateral), while the top-k field is *collateral*-limited (observed: C-ref exceeds the frozen grid in only 5 of 30 runs under the extension because its budget binds first). Second, extending the grid lets the random field catch up in removal, but at a collateral cost that grows as ρ_k^2; the gap shrinks with extension depth and, under a binding budget, need not close (observed up to 9B: $+0.233 \to +0.146 \to +0.104$ and then flat, because at $\alpha = 256$ neither arm clears the budget; the saturated value excludes zero). When the random field reaches the required scale before its own budget binds, the gap should close instead, and the bias-blind rule α^* shows why it can: the random field's budget binds at about four times the scale of the magnitude-selected field's (App. B), so it sees markedly less collateral curvature per unit scale, and the gap covers zero at both tiers tested. Magnitude selection buys the effect where it is cheapest: at the smallest scale and the least collateral. This is also where the account departs from DARE: random dropping with rescaling by $1/(1-p)$ is an unbiased estimator of ΔW and works when the effect is linear in ΔW and the variance averages out over many coordinates; a sign field has no magnitudes to rescale, and a budget on α prevents the rescaling that would substitute for them.

Proposition 4 (the support index has an entropy floor). A k-subset of d coordinates at density $p = k/d$ costs at least $\log_2 \binom{d}{k} \approx d\,H_b(p)$ bits, i.e. $H_b(p)/p$ bits per retained coordinate, where H_b is the binary entropy. At $p = 0.01$, $H_b(p) = 0.0808$ bits, so the index costs 8.08 bits per retained coordinate and the shippable patch is at least 9.08 bits per retained coordinate—about nine times the one-bit sign payload—which is the "∼8 bits per coordinate" and "≈ 9× payload" of §5.5 and Table 5. A plain 32-bit index costs 33 bits per coordinate; a bitmap costs d bits, exactly the dense

payload, and is never better at $p < 1/2$. The sizes reported in this paper sit at this floor; nothing below it is available without modeling the support itself.

What the account does not claim. It is first-order in α and ignores saturation of removal, interaction between coordinates, and the discreteness of the MMLU budget; it treats curvature as comparable across supports, which the adaptive-scale data contradict for random supports (their budget binds at about four times the scale); and it does not explain why the sign field of one axis transfers to another (App. C), which is an empirical statement about corpora, not about ΔW.