ERASE IT WHERE YOU FOUND IT: MATCHED-POSITION ERASURE LOWERS DEMOGRAPHIC DEPENDENCE AT A MEASURED COST

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear concept erasure removes a direction from a model's hidden states. LEACE, the standard eraser, guarantees that no linear probe can recover the concept from the edited states. This paper shows that the token positions at which a direction is removed decide what the erasure removes and what it costs. The demographic direction comes from swapping the demographic term of a bias item and measuring how much the answer depends on it. That direction is then erased only at the term's tokens, in every layer. At those positions the erasure lowers the measured dependence by 11 to 28 percent on four open models. LEACE fitted to the same swap data, and random subspaces, remove nothing at those positions: a probe guarantee is not a guarantee about the answer. An additive steering vector at the same positions does as well as the erasure. At the last prompt token neither edit removes the dependence, and at every token both harm the model. On new BBQ items the erasure removes the demographic term's contribution to the answer, whether that contribution is a stereotype or the information the question needs. On items in which the term is the answer, accuracy falls from 0.80 to 0.16 on Gemma-2-2B. The evidence spans four models, three benchmarks, a pre-registered replication on new items, and controls for direction, rank, energy, readout and operator. The site of the edit decides what is removed and what is lost. An erasure result should state where the edit acts and what it costs.

1 INTRODUCTION

Linear concept erasure leaves open the choice of where a projection is applied, and this paper shows that the choice decides what the erasure removes and what it costs. In the picture behind erasure, a concept occupies a subspace of the residual stream and a projection removes it. Iterative nullspace projection (Ravfogel et al., 2020) and LEACE (Belrose et al., 2023) make the projection cheap, and their guarantee is precise: after the edit no linear predictor recovers the concept. The guarantee says nothing about where the projection should act. A generation wrapper projects the last prefill token and every generated token, a global hook projects every token, and a targeted edit projects only the positions that supplied the evidence for the direction. The direction is the same in all three cases; the choice among them is what this paper varies.

We measure the dependence by intervention rather than by a probe. A seed is a multiple-choice bias item containing a demographic term, and two variants of the seed differ only in that term. We patch the residual stream at the term's token positions from one variant into the other, at every layer, and record how far the answer moves; we call that movement the patching effect (Vig et al., 2020; Zhang & Nanda, 2024). The direction we erase is the leading direction of the same swap's activation differences at those positions, and we erase it where it was fitted: the matched-position edit. It touches the prefill only, because the term has already been processed when decoding begins, and a term the model writes while decoding is not edited.

The edit does what an erasure is meant to do, at the positions where the direction was fitted. Two evaluations follow: a *pooled audit* on four open models and three bias benchmarks, and a *fresh replication* pre-registered on new BBQ items. In the pooled audit the edit lowers the patching effect by 11 to 28 percent, and the reduction survives a Holm-corrected test on three of the four models

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.

(Table 1). It holds when the effect is read on the gold option's log-probability, so it is not a shift of the whole logit vector (Table 22). A pooled mean difference, a neutral-contrast basis and random subspaces of matched rank at the same positions remove nothing (Figure 3). LEACE fitted to the swap label removes nothing either, at fifty pairs or at every fit pair, with or without covariance shrinkage (Table 23). It keeps its guarantee that no probe recovers the swap side, and the patching effect stays, because the two are different quantities. The fresh replication repeats the comparison on new BBQ items with a readout that scores complete option strings. There the edit lowers the answer-level patching effect on both models and the answer's sensitivity to the swap on Llama-3.1-8B (Table 2). An additive steering vector at the same positions, matched in energy, does as well as the erasure. At the last prompt token it does nothing, and at every position it breaks the model (Table 3). In both evaluations the same direction applied at every position damages the model, on MMLU and WikiText-2 (Table 11) and on the very items of the fresh replication (Figure 1). Random subspaces applied everywhere, and an edit of equal energy at a non-demographic control span, change nothing.

The fresh replication also measures what the matched-position edit costs. At full strength, generation accuracy falls on both models, by 6.2 points on Gemma-2-2B. On a control set whose correct answer names the swapped entity it collapses, while the control-span edit leaves it unchanged (Table 15). The edit removes the demographic term's contribution to the answer, whether that contribution is a stereotype or the information the question asks for. A calibrated strength one tenth as large costs less and removes less.

This paper makes three contributions.

- This paper proposes the matched-position protocol for erasing a contrast-derived direction, with one implementation, one fitted direction and identity checks that pass exactly. It removes the measured dependence at the span where swap-label LEACE, a pooled mean difference and random subspaces remove nothing, and the reduction replicates on fresh items under a whole-answer readout.
- This work shows that the application site, more than the direction or the operator, decides what an edit removes and what it costs. The direction that is removable at its own positions damages the model everywhere else, an additive steering vector follows the same rule, and an edit of equal energy at a non-demographic span does nothing.
- This study contributes a pre-registered confirmatory protocol on fresh items, with calibrated energies, template-cluster intervals, Holm correction and non-inferiority tests, and uses it to measure the cost of the edit as a sensitivity–accuracy trade-off.

2 Related work

The projection this paper applies, the measurement it uses and the items it edits each come from a line of work that never asked where a projection should act. Removing a linear subspace to suppress an attribute begins, in its modern form, with iterative nullspace projection (Ravfogel et al., 2020); Haghighatkhah et al. (2022) showed that one projection along the class-mean difference achieves what the iterations do. Ravfogel et al. (2022) posed erasure as a minimax game over low-rank projections, and Belrose et al. (2023) gave a closed-form eraser that prevents any linear predictor from recovering the concept at the least least-squares cost. Basu Roy Chowdhury et al. (2025) derived limits on what survives perfect erasure, and Gur-Arieh et al. (2025) moved erasure into the parameters to localise it. These methods guarantee decodability at the positions where they are fitted; what the same projection does elsewhere is the question they leave open.

A parallel line edits activations at inference without projecting them out. Activation addition and representation engineering add a contrast-derived vector (Turner et al., 2023; Zou et al., 2023), inference-time intervention shifts selected heads along probe directions (Li et al., 2023), and debiasing vectors steer along a bias direction (Li et al., 2025; Siddique et al., 2026). Other methods locate and edit the responsible components (Islam et al., 2026; Xu et al., 2025), ablate sparse-autoencoder features (Ahsan & Wallace, 2026), or change the prompt (Schick et al., 2021). Chand et al. (2026) report that reducing one targeted bias can worsen an untargeted one. We compare erasure with two such steering vectors at the same sites, so that the site question is answered for both operators.

The directions themselves can be found by intervention. Activation patching traces a prediction to the components that carry it (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023), Zhang & Nanda

(2024) examine its metrics and failure modes, and Geiger et al. (2021) frame such interventions as tests of causal abstraction (Pearl, 2009). We use patching to measure the dependence an eraser targets and re-read the same quantity after the edit, so the target and the evaluation share one instrument.

A few hidden dimensions of a trained transformer carry activations orders of magnitude larger than the rest (Sun et al., 2024; Yu et al., 2024) and have been tied to social bias (Oh et al., 2024). Section 5.6 tests, and does not support, the account that the selected direction is costly because it overlaps this structure.

The seeds come from bias benchmarks (Parrish et al., 2022; Nangia et al., 2020; Nadeem et al., 2021) whose scores have been criticised on construct-validity grounds (Blodgett et al., 2020; 2021; Bean et al., 2025; Cronbach & Meehl, 1955; Messick, 1995; Jacobs & Wallach, 2021). We use them as a source of items only, and the quantities we reduce are patching and sensitivity quantities read on generated answers and option scores.

3 METHOD

The method has three parts: the measurement of a dependence, the erasure of one direction at chosen positions, and the readout of the edited model (Figure 2, Appendix A); a set of bases supplies the controls.

3.1 PATCHING EFFECT AT THE DEMOGRAPHIC SPAN

The patching effect is the quantity every claim rests on: how far the answer moves when the demographic term's own activations are swapped. Let a *seed* be a multiple-choice item whose text contains a demographic term. Its gold option has a first token g, whose logit at the final position we write $\ell_g(\cdot)$. Two variants a and b of the seed differ only in that term. Let $S(x)$ be the token positions of the term in the prompt as fed to the model, chat template included, resolved through the tokenizer's offset mapping. Let $h_t^{(\ell)}(x)$ be the residual stream at layer ℓ and position t on variant x. We run the model on a and cache $h_t^{(\ell)}(a)$ for $t \in S(a)$ at every layer. We then run the model on b and, at every layer, overwrite the residual stream at the positions of $S(b)$ with the cached vectors, truncating to the shorter span. The quantity recorded is

$$C(a,b) \;=\; \ell_g(\mathrm{swap}(a \to b)) \;-\; \ell_g(b), \tag{1}$$

and we call it the patching effect. A model whose gold logit does not depend on the representation of the demographic term returns $C = 0$; $|C|$ is the demographic dependence of the seed.

One convention decides whether C is a valid instrument under an edit. The source run on a, the clean run on b and the patched run are all made with the edit hooks active, so the cached vector at layer ℓ already reflects the edits below. Under this *sequential* source state an identity pair ($a = b$) returns $C = 0$ exactly. The *frozen* alternative, which caches unedited activations and projects them before injection, returns non-zero C on identity pairs (Appendix B) and is kept only as a diagnostic.

The fresh replication reads the same patch at the level of the whole answer, because the first token of the gold option often also begins a competing option (Section 7). Each option k receives the score $s_k(x)$, the log-likelihood of the complete string {"answer": "⟨option k⟩"} as a continuation of prompt x, with the same prefill positions for every candidate. The gold margin $M(x) = s_g(x) - \log \sum_{k \neq g} \exp s_k(x)$ replaces the single logit. The answer-level patching effect is $C_{\mathrm{answer}}(a,b) = M(\mathrm{swap}(a \to b)) - M(\mathrm{id}(b))$, where the reference run injects b's own span states at b. Both passes then share one kernel path, and the same-input null is exactly zero in bfloat16. The *answer sensitivity* needs no patch: with $q(x) = \mathrm{softmax}\, s(x)$, $T(a,b) = \frac{1}{2}\sum_k |q_k(a) - q_k(b)|$ is the total variation between the two variants' option distributions, and a lower T means the answer depends less on the swap.

3.2 EDIT AND APPLICATION SITES

One map serves every edited condition; what the experiments vary is where it is applied. With U_ℓ an orthonormal $r \times d$ matrix at layer ℓ and $\alpha \in [0,1]$ a strength,

$$h_t'^{(\ell)} \;=\; h_t^{(\ell)} - \alpha\, U_\ell^\top U_\ell\, h_t^{(\ell)} \qquad \text{for } t \in \mathcal{S}, \text{ for every decoder layer } \ell. \tag{2}$$

At $\alpha = 1$ this is the orthogonal projection onto the complement of the row space of U_ℓ, applied through depth. The choice of $\mathcal{S}$ is the site policy. Under the *span* policy, $\mathcal{S}$ is $S(x)$, every token of the demographic term in the prompt, and nothing is edited at decode steps because the span has already been processed. Under the *last-token* policy, $\mathcal{S}$ is the final prefill position and every generated token, which is what an eraser wrapped around a generation loop does. Under the *every-position* policy, $\mathcal{S}$ is every prefill position (fresh replication) or every prefill and generated position (pooled audit). The fresh replication adds a *control-span* policy: a run of non-demographic context tokens with the same token count as the demographic span, nearest to it, overlapping neither it nor the second entity. One hook serves every policy; it logs every position it changes and the *activation energy* it removes, the squared norm of the removed component summed over layers and edited positions, relative to a fixed unedited denominator.

Identity checks run on every model before any result is recorded, and each passes exactly. Strength $\alpha = 0$ and rank 0 reproduce the unedited logits, positions outside the span are unchanged at the first edited layer, the span is changed, and no decode step is edited under the span policy. The every-position policy touches every prefill position and no decode step, and the same-input patch returns $C_{\text{answer}} = 0$.

3.3 Bases and controls

The erasure basis is fitted from the swap's own activation differences, so no part of it is chosen by its effect on the output; we call it a *contrast-derived projection*. It is fitted on 300 pairs of variants from the fit seeds, of which 54 substitute a neutral term such as *person* for the demographic term. For each pair we collect, at every layer, the difference $h_t^{(\ell)}(a) - h_t^{(\ell)}(b)$ at each span position, giving 327 to 330 difference vectors per layer. The rows of U_ℓ in Eq. (2) are the leading right singular vectors of the centred difference matrix, the directions along which the swap moves the residual stream. Rank one is the confirmatory setting; ranks two and four are exploratory. The fresh replication reuses the frozen rank-one basis of the pooled audit unchanged.

The pooled audit compares the basis with four controls that share the fitting data, the layers, the site and the hook. The normalised *mean difference* is the rank-one direction the centring discards, the direction Haghighatkhah et al. (2022) found sufficient for nullspace projection. *LEACE* is fitted with the authors' package on the same span activations, labelled by swap side, and applied as its exact affine map. It is fitted once on unedited activations (*frozen*) and once layer by layer on already-edited activations (*sequential*, the concept-scrubbing protocol of Belrose et al. (2023)). The swap side is not one coherent concept, since the fit pairs contain 991 distinct term contrasts. The fresh replication therefore adds a LEACE and a mean difference fitted on the one contrast with enough pairs, *man*/*woman* (45 pairs), with a held-out probe check; Section 5.5 refits both at larger sizes. Three *random* orthonormal subspaces of matched rank test whether any subspace of that rank at that site would lower the dependence. A *neutral-contrast* basis is fitted on the neutral pairs alone, and an *uncentred* SVD of the same differences tests the centring.

The fresh replication varies the site and the strength under one basis, with seven conditions on every seed. Three use the span basis: no edit, the span erasure at $\alpha = 1$, and the span erasure at a calibrated strength. Three are calibrated controls: the basis at the control span, and a random direction at the span and at the control span. The seventh applies the span basis at every prefill position at $\alpha = 1$. The calibration target is the smallest $\alpha = 1$ energy among the four calibrated cells on the development seeds, which is the random direction's. Each cell's α is found by at most eight evaluations to within five percent of it. The four calibrated cells therefore remove the same amount of activation, and only the direction and the position differ between them.

3.4 Readouts and statistics

Each claim is read out on the quantity that defines it and tested under a family fixed before the test phase. The pooled audit yields four readouts per seed and condition. The patching effect is re-read at the span with the edit active on both variants. *Option scoring* computes the log-likelihood of each option string and gives the argmax and its correctness. *Generation* decodes greedily for 48 tokens under a system prompt that requests a JSON object whose `answer` field is the exact option text. A deterministic parser maps the field to an option. For the 2.4 percent of test sides it cannot match, a judge (`deepseek-chat`, temperature zero, prompt in Appendix C) sees only the option list and the

answer string. A further 2.6 percent stay *invalid*, which counts as wrong. The *capability probes* score 200 MMLU test questions (Hendrycks et al., 2021) by option log-likelihood and 20,000 WikiText-2 test tokens (Merity et al., 2017) by perplexity, one pass per cell, under the every-position policy because these prompts have no demographic span. The fresh replication yields T, C_{answer}, the legacy C read from the same passes, and generation accuracy from the parser alone.

The pooled-audit statistic is the paired reduction $D = \mathrm{mean}(|C_{\mathrm{pre}}| - |C_{\mathrm{post}}|)$ over demographic pairs. The ratio of means $R = 1 - \mathrm{mean}|C_{\mathrm{post}}|/\mathrm{mean}|C_{\mathrm{pre}}|$ is secondary, and a win rate would hide that the reduction is concentrated on a minority of pairs (Table 9). Intervals are percentile bootstrap intervals over 2,000 draws that resample seeds as clusters (Efron & Tibshirani, 1993). Four comparisons were declared before the test phase: the erasure against the unedited model and against the mean of three random subspaces, on D and on generation accuracy. They are Holm-corrected within each model (Holm, 1979).

The fresh replication was pre-registered before any seed was scored, with eight confirmatory tests and six non-inferiority tests. Per model, H1 asks whether the span erasure lowers T, and H4 whether generation accuracy changes. H2 asks whether T under the calibrated span edit is lower than under the calibrated control-span edit, and H3 whether it is lower than under the calibrated random direction at the span. Each is a studentised bootstrap test over 10,000 resamples that clusters seeds by BBQ template, and the eight p-values are Holm-corrected as one family. Six one-sided tests ask whether accuracy under the span erasure, the calibrated span edit and the control-span edit stays within two percentage points of the unedited model, Holm-corrected as a second family. The outcome rules were fixed in advance. A reduction is claimed only when its test rejects in the hypothesised direction, and a rejection in the opposite direction is a measured harm. H2 and H3 are limited when the achieved test-set energies of the four calibrated cells differ by more than ten percent. Every other number is exploratory and labelled so.

4 Experimental setup

The models and items of each evaluation, and the split of the items, fix what it can claim. The pooled audit runs four open instruction-tuned models under every condition, Llama-3.1-8B (Grattafiori et al., 2024), Qwen2.5-7B (Qwen Team, 2024), Gemma-2-2B (Gemma Team, 2024) and Phi-4-mini (Abouelenin et al., 2025), in bfloat16 with flash attention through one hook implementation. Seeds derive from BBQ (Parrish et al., 2022), CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021), across ten demographic axes, each recast as a multiple-choice item with the gold option given as text (Appendix A). A manifest assigns the 596 seeds to fit, development and test splits with no seed in common (Table 4). Bases are fitted on the fit split only. The development phase screens the strength α (Table 6) on 100 development seeds and runs the pilot of Appendix B on 24 of them. The test phase evaluates a stratified sample of 160 test seeds once, the same list for every model and condition, selecting per seed the demographic pair with the largest unedited $|C|$.

The fresh replication runs Gemma-2-2B and Llama-3.1-8B on items built from BBQ alone. Each seed is an ambiguous BBQ item whose gold answer is the *unknown* option. It is paired with the sibling item of the same template in which one entity's descriptor differs and nothing else does, and the option list is kept fixed. The 160 test seeds come two per template from 80 templates that no earlier phase of this work touched, across 9 categories (Table 5). A separate 24 development seeds serve the calibration, and 36 disambiguated items whose correct answer names the swapped entity form the relevant-information control. The demographic span is every whole-word occurrence of the descriptor in the user prompt. The control span is resolved per seed, and the 7 Gemma-2-2B and 5 Llama-3.1-8B seeds without a qualifying run are excluded from the location tests only. Every number comes from one deterministic pass, so the intervals describe sampling over seeds.

5 Results

The site of the edit decides each result below. At its own positions the erasure removes the dependence, everywhere else it damages the model, and on fresh items it costs accuracy where the demographic term is the answer.

Table 1: Pooled audit: the four pre-declared comparisons per model; * rejects after Holm correction.

		vs. unedited			vs. random subspaces	
Model	$\|C\|$ before	D	R (%)	Δacc.	D	Δacc.
Llama-3.1-8B	0.158	0.044 [0.019, 0.072]*	28.1	+0.025 [-0.003, +0.053]	0.039 [0.014, 0.064]*	+0.032 [+0.003, +0.061]
Qwen2.5-7B	0.663	0.076 [0.004, 0.146]	11.4	+0.022 [-0.013, +0.056]	0.064 [-0.007, 0.134]	+0.022 [-0.011, +0.056]
Gemma-2-2B	0.686	0.114 [0.044, 0.186]*	16.7	+0.072 [+0.028, +0.119]*	0.109 [0.038, 0.183]*	+0.069 [+0.025, +0.116]*
Phi-4-mini	0.408	0.113 [0.059, 0.163]*	27.7	+0.037 [-0.000, +0.078]	0.112 [0.061, 0.160]*	+0.035 [-0.003, +0.075]

Table 2: Fresh replication: the pre-registered tests per model; * rejects after Holm correction.

Model	Test (positive is the hypothesised benefit)	templates	estimate	Holm p
Gemma-2-2B	sensitivity T: no edit − span erasure	80	+0.021 [-0.014, +0.057]	0.760
Gemma-2-2B	position: T at control span − T at demographic span	79	-0.010 [-0.033, +0.011]	0.762
Gemma-2-2B	direction: T random direction − T span basis	80	-0.009 [-0.030, +0.010]	0.762
Gemma-2-2B	generation accuracy: span erasure − no edit	80	-0.062 [-0.100, -0.028]*	0.013
Llama-3.1-8B	sensitivity T: no edit − span erasure	80	+0.055 [+0.025, +0.084]*	0.006
Llama-3.1-8B	position: T at control span − T at demographic span	79	+0.034 [+0.015, +0.053]*	0.006
Llama-3.1-8B	direction: T random direction − T span basis	80	+0.031 [+0.012, +0.050]*	0.015
Llama-3.1-8B	generation accuracy: span erasure − no edit	80	-0.041 [-0.072, -0.013]	0.068

5.1 Pooled audit

At its own positions the rank-one span erasure lowers $|C|$ on every model, by 11 to 28 percent of its unedited mean (Table 1). The interval excludes zero on every model, and the Holm-corrected test rejects on Gemma-2-2B, Llama-3.1-8B and Phi-4-mini but not on Qwen2.5-7B. The comparison against the mean of three rank-matched random subspaces gives the same picture, so the reduction is not something any rank-one projection at the span would produce. Generation accuracy, with invalid outputs counted as wrong, rises on Gemma-2-2B under the corrected test and has an interval that includes zero on the other three; output validity stays high (Table 9). Read on its own, the pooled audit supports removal at the fitted site on three models and no accuracy-preservation claim, which is why the fresh replication measures accuracy again with a stricter readout.

Every condition appears on one axis per model in Figure 3 (Appendix C). The uncentred SVD tracks the centred one. The intervals of every other condition include zero: LEACE under either fitting protocol, the mean-difference direction that Haghighatkhah et al. (2022) found sufficient for probes, the random subspaces at ranks one and four, and the neutral contrast. LEACE's null is the consequential one: its guarantee concerns whether a probe can recover the swap side, and the gold logit still moves when the span is patched. Section 5.5 shows that neither the fit size nor the conditioning of the covariance changes this.

5.2 Every-position application

The direction that is removable at its own positions is destructive everywhere else. Table 11 (Appendix D) applies the same bases at every token position, the one policy that reaches an MMLU question or a WikiText-2 passage. The span-fitted erasure at rank one lowers MMLU accuracy on every model and raises perplexity by a factor that reaches 424 on Phi-4-mini. Random subspaces applied everywhere leave both probes close to their unedited values, and LEACE applied everywhere costs less than the erasure on every model. The last-token policy of a generation wrapper raises $|C|$ on three models and lowers output validity on two (Appendix B).

5.3 Fresh replication

On fresh items the reduction survives a pre-registered test on one model and the answer-level patching effect falls on both (Table 2; Figure 1 shows every condition). On Llama-3.1-8B the span erasure lowers the answer sensitivity T from 0.141 to 0.086, and the H1 test rejects after Holm correction. At calibrated energy the span edit also beats the same edit at the control span (H2) and a random direction at the span (H3), both after Holm correction. The achieved test-set energies of the four calibrated cells differ by more than the pre-declared ten percent on this model, so we report the two specificity

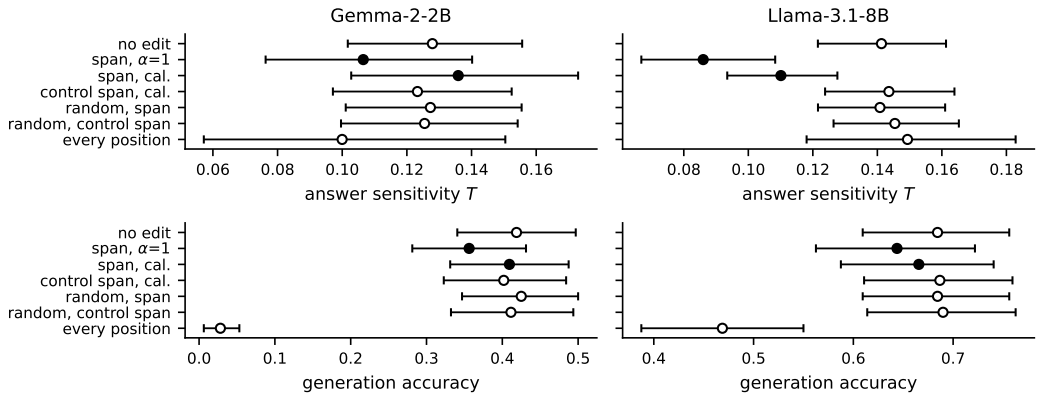


Figure 1: Fresh replication: T (top) and accuracy (bottom) per condition; filled markers, span erasure.

results as limited. The drift has a cause: the removed energy grows with the number of edited tokens, and the test spans are longer than the development spans (Appendix F). On Gemma-2-2B the H1 estimate is positive with an interval that includes zero, and H2 and H3 are null; the calibrated cells stay within the energy rule there.

The patching instrument agrees on both models (Table 17, Appendix E). The answer-level patching effect $|C_{\text{answer}}|$ falls under the span erasure on both models, by roughly two fifths, with intervals that exclude zero, and the legacy first-token effect falls with it. The matched-position edit therefore removes the demographic term's influence on the answer that a patch can measure, on both models, and lowers the behavioural sensitivity of the answer distribution on one.

The edit position, not the direction, decides the outcome on the same items. The span basis applied at every prefill position collapses generation accuracy and output validity on both models while leaving T close to its unedited value (Table 14). The calibrated edit at the control span removes as much activation energy as the calibrated span edit and changes neither T nor accuracy (Table 16). On Llama-3.1-8B it is the one edited condition that passes the non-inferiority test. A random direction at the span, at the same energy, changes nothing either.

5.4 Cost of the edit

The full-strength erasure lowers generation accuracy on both models (H4 in Table 2): by 6.2 points on Gemma-2-2B, which rejects after Holm correction, and by 4.1 points on Llama-3.1-8B, which does not. Neither passes the two-point non-inferiority test, and the share of pairs with both sides correct falls on both (Table 16). The accuracy gain the pooled audit measured on Gemma-2-2B does not replicate. Table 15 shows where the loss comes from. On the control seeds, whose correct answer names the swapped entity, accuracy falls from 0.80 to 0.16 on Gemma-2-2B and to under half of its unedited value on Llama-3.1-8B. The control-span edit leaves it unchanged (Appendix E; the last column of Table 3 repeats the figures). The edit removes the demographic term's contribution to the answer, and when that contribution is the information the question asks for, the answer is lost with it. The per-category estimates agree (Table 20): the accuracy loss concentrates on the disability and physical-appearance items, where the descriptor also describes the person, and the sensitivity reduction on gender, nationality, race and religion.

The calibrated strength makes the trade-off adjustable (Table 16). At about one tenth of the full energy, the accuracy loss shrinks to within two points on both models. The reduction in T shrinks with it, to about half of the full-strength reduction on Llama-3.1-8B and to nothing on Gemma-2-2B. The strength is therefore a choice the practitioner makes, not one the method fixes. The earlier pool agrees in behavioural terms (Table 10). There the full-strength edit raised the share of pairs with both sides correct and lowered answer flips on all four models, on items where the demographic term was rarely the answer.

Table 3: Steering baselines at three sites against the span erasure on the fresh items; control accuracy on the relevant-information set.

Model	Edit	T [95% CI]	R_T (%) (p)	Acc.	Valid	R_C (%) (p)	Control acc.
Gemma-2-2B	no edit	0.128 [0.102, 0.156]	–	0.42	1.00	–	0.80
Gemma-2-2B	span erasure, $\alpha = 1$	0.107 [0.076, 0.140]	17 (0.248)	0.36	0.95	44 (<0.001)	0.16
Gemma-2-2B	debiasing vector at the span	0.127 [0.100, 0.156]	1 (0.951)	0.44	0.99	28 (<0.001)	0.55
Gemma-2-2B	contrast vector at the span	0.082 [0.059, 0.108]	36 (<0.001)	0.43	1.00	44 (<0.001)	0.36
Gemma-2-2B	debiasing vector at the last prompt token	0.121 [0.095, 0.149]	5 (0.009)	0.66	0.92	–	–
Gemma-2-2B	contrast vector at the last prompt token	0.124 [0.099, 0.152]	3 (0.197)	0.32	0.69	–	–
Gemma-2-2B	debiasing vector at every prefill position	0.240 [0.184, 0.298]	-88 (0.002)	0.00	0.00	–	–
Gemma-2-2B	contrast vector at every prefill position	0.133 [0.095, 0.175]	-4 (0.821)	0.29	1.00	–	–
Llama-3.1-8B	no edit	0.141 [0.122, 0.162]	–	0.68	1.00	–	0.84
Llama-3.1-8B	span erasure, $\alpha = 1$	0.086 [0.067, 0.108]	39 (<0.001)	0.64	1.00	37 (<0.001)	0.40
Llama-3.1-8B	debiasing vector at the span	0.082 [0.065, 0.102]	42 (<0.001)	0.70	0.99	39 (<0.001)	0.41
Llama-3.1-8B	contrast vector at the span	0.067 [0.051, 0.086]	52 (<0.001)	0.72	0.98	44 (<0.001)	0.16
Llama-3.1-8B	debiasing vector at the last prompt token	0.139 [0.120, 0.160]	1 (0.425)	0.52	0.89	–	–
Llama-3.1-8B	contrast vector at the last prompt token	0.142 [0.123, 0.162]	0 (0.810)	0.89	0.91	–	–
Llama-3.1-8B	debiasing vector at every prefill position	0.222 [0.171, 0.275]	-57 (0.006)	0.00	0.00	–	–
Llama-3.1-8B	contrast vector at every prefill position	0.193 [0.142, 0.246]	-37 (0.079)	0.00	0.00	–	–

5.5 Robustness of the reduction and steering baselines

The reduction is not an artefact of a shifted logit vector. A patch can move every logit at once, and the raw effect C would count that shift. Re-read on the gold option's log-probability, the reduction matches the raw one on all four models, and the corrected test rejects on the same three (Table 22). On the option margin it agrees on three models and vanishes on Llama-3.1-8B.

The reduction is larger where the two spans have equal length, and the gap is a property of the alignment rule. Truncating the patch to the shorter span shrinks the reduction on unequal-length pairs (Table 26), and padding the shorter span to the longer one restores it (Table 27).

Rank deficiency does not explain the LEACE null; the label does. Refitting the swap-side LEACE at three fit sizes, with or without shrinkage, brings the held-out probe to chance every time and lowers the patching effect on no model (Tables 23 and 24). The fitted direction is close to orthogonal to the contrast direction, so the subspace that best hides the swap side is not the one the answer reads. A LEACE fitted to a coherent gender label on thousands of descriptor rows behaves differently. On a gender-flip set it lowers the answer-level patching effect on both models at a small fraction of the erasure's energy, where the paper's 61-row fit removes little (Table 25). The erasure still removes about twice as much.

The site matters more than the operator. Two additive steering baselines were calibrated to the erasure's energy and applied at three sites (Table 3): a debiasing vector in the manner of Li et al. (2025), and a contrast vector in the manner of representation engineering (Zou et al., 2023; Turner et al., 2023). At the span both lower T and $|C_{\text{answer}}|$ by about as much as the erasure. At the last prompt token, the site a generation wrapper uses, neither lowers the dependence, and the debiasing vector raises the rate of *unknown* answers instead. At every prefill position both collapse accuracy and validity, as the projection did. On the control set the span-site vectors cost accuracy as the erasure does, although the debiasing vector was fitted on ambiguous items alone. The reduction, and its cost, belong to the position.

5.6 Explanatory accounts and forecast

Three pre-registered accounts of the cost, each of which would replace the site finding with a simpler mechanism, all fail (Table 12, Appendix D). A random subspace at the development strength changes neither $|C|$ nor option accuracy. It removed only 1.6 to 3.6 percent of the erasure's energy on the test seeds, so the energy-matched comparison rests on the fresh replication's calibrated cells. A blockwise refit on already-edited activations removes the same dependence as the frozen basis, and restoring the massive-activation coordinates after erasure changes neither quantity. The unedited $|C|$ of a prompt, its audit score, adds no held-out forecast of post-edit errors over a confidence-and-length baseline (Table 13).

6 DISCUSSION

The results turn on a distinction that erasure methods leave implicit: where a basis is fitted and where it is applied. The LEACE guarantee holds on the fitted distribution, where perfect erasure must already discard part of the representation (Basu Roy Chowdhury et al., 2025). Our measurements add three things. At the positions where the direction was fitted, a LEACE map fitted to the swap label removes no dependence at any fit size, and its direction is nearly orthogonal to the contrast direction. The subspace that best hides the swap side is not the one the answer reads. Fitted to one coherent concept on enough rows, LEACE removes part of the dependence, and the contrast-derived projection removes more, on both instruments. An additive vector at the same positions removes as much as the projection, so the reduction belongs to the site rather than to the operator. Away from those positions the same edits discard what other tokens need, while a random direction of matched rank, or the control span at equal energy, discards nothing.

An erasure fitted at a site is applied at that site, and its cost is measured where it is applied, on items where the erased term may be the answer. The matched-position edit removes the demographic term's contribution to the answer. On ambiguous items that contribution is a dependence the benchmark calls bias, and T and $|C_{\text{answer}}|$ fall. On items where the term carries the answer, it is information, and accuracy falls. On Gemma-2-2B the two instruments part: the answer-level patching effect falls while T does not, and their per-seed changes are nearly uncorrelated (Appendix F). The patched pathway is cut, but the unedited answer distribution depends on it less than the patch suggests.

7 LIMITATIONS

The reductions are measured on the instruments that define the dependence, so they are removals on those instruments and not a general fairness claim. The pooled audit's patching effect scores one token, which identifies a unique option on a minority of seeds (Table 7), which is why the fresh replication scores whole answers. The sensitivity reduction on Llama-3.1-8B holds under the frozen scoring and a fixed option permutation, but not under label-only candidates or length-normalised scores (Table 21). The gender-flip set of Table 25 is built by descriptor substitution, as the pooled audit's variants were, because BBQ pairs one female with one male entity in every gender item. The method needs an explicit span, so an attribute that no single token carries is out of its reach. The capability probes are one deterministic pass each, and the test seeds were used once but not held by a third party.

8 CONCLUSION AND FUTURE WORK

The positions at which a demographic direction is erased decide what the erasure removes and what it costs. This paper measured a demographic dependence by patching at the demographic term and erased the swap's leading direction at the same positions, in four language models and on fresh BBQ items. At its own positions the erasure lowers the patching effect on three of the four models, where swap-label LEACE and random subspaces remove nothing. An additive steering vector at the same site removes as much. The same edits applied everywhere damage capability, and an equal-energy edit at a non-demographic span does nothing. The edit costs accuracy at full strength, most of all where the demographic term is the answer, and a calibrated strength trades the two.

One question remains open: why the span-fitted direction is costly to remove at other positions; a layer-wise decomposition of the every-position damage would answer it. A direction fitted on ambiguous items alone does not spare the relevant-information control, since the debiasing vector is fitted that way and still pays the cost at the span. Four open models of at most eight billion parameters and one attribute family bound the evidence.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the edit, the site policies, the source-state convention, the bases, the readouts, the calibration and the statistics. Appendix A shows a seed as the model sees it. The code and every result file are in an anonymised repository at `https://anonymous.4open.science/r/Cure_Audit_Benchmark/`. The repository holds the split manifest, the frozen seed lists of

both evaluations, the pre-registered protocol file with its hash, the fitted bases, the per-item result files for every stage and the summary tables. Provenance files there map every number in this paper to the result file that produced it. The code runs each stage from one entry point, and two scripts generate every table and figure from the result files. All benchmarks and models are public.

Ethics statement

This work studies demographic bias in language models using three public benchmarks. No human subjects were involved, no personal data were collected, and the demographic terms in the variants come from those benchmarks. The finding that a matched-position erasure lowers a demographic dependence is not a certificate of fairness. The dependence is measured on two instruments at one site, the same edit lowers accuracy on fresh items, and it removes needed information when the demographic term is the answer. We report those costs and the negative results on forecasting and on the massive-activation account so that none is cited as established.

AI use statement

The writing process included a final language-editing step. ChatGPT was used to correct grammar and spelling and to improve the clarity of the writing. All content was reviewed and verified by the authors, who take full responsibility for it.

References

Abdelrahman Abouelenin et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. arXiv:2503.01743, 2025.

Hiba Ahsan and Byron C. Wallace. Can SAEs reveal and mitigate racial biases of LLMs in healthcare? In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.00177.

Somnath Basu Roy Chowdhury, Kumar Avinava Dubey, Ahmad Beirami, Rahul Kidambi, Nicholas Monath, Amr Ahmed, and Snigdha Chaturvedi. Fundamental limits of perfect concept erasure. In *Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 901–909, 2025.

Andrew M. Bean et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. arXiv:2511.04703.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2306.03819.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proc. 58th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pp. 5454–5476, 2020. doi: 10.18653/v1/2020.acl-main.485.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proc. 59th Annual Meeting of the Assoc. for Computational Linguistics (ACL-IJCNLP)*, pp. 1004–1015, 2021. doi: 10.18653/v1/2021.acl-long.81.

Shireen Chand, Faith Baca, and Emilio Ferrara. No free lunch in language model bias mitigation? targeted bias reduction can exacerbate unmitigated LLM biases. *AI*, 7(1):24, 2026. doi: 10.3390/ai7010024.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.02997.

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

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

Yoav Gur-Arieh, Clara Haya Suslik, Yihuai Hong, Fazl Barez, and Mor Geva. Precise in-parameter concept erasure in large language models. In *Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 18986–19006, 2025. doi: 10.18653/v1/2025.emnlp-main.960.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In *Proc. 2022 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8395–8416, 2022. doi: 10.18653/v1/2022.emnlp-main.575.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proc. 9th Int. Conf. Learning Representations (ICLR)*, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Sekh Mainul Islam, Nadav Borenstein, Siddhesh Milind Pawar, Haeun Yu, Arnav Arora, and Isabelle Augenstein. BiasGym: A simple and generalizable framework for analyzing and removing biases through injection. In *Findings of the Assoc. for Computational Linguistics: EMNLP 2026*, 2026. arXiv:2508.08855.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. FairSteer: Inference time debiasing for LLMs with dynamic activation steering. In *Findings of the Assoc. for Computational Linguistics: ACL 2025*, pp. 11293–11312, 2025. doi: 10.18653/v1/2025.findings-acl.589.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. arXiv:2202.05262.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *Proc. 5th Int. Conf. Learning Representations (ICLR)*, 2017.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (9):741–749, 1995. doi: 10.1037/0003-066X.50.9.741.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proc. 59th Annual Meeting of the Assoc. for Computational Linguistics (ACL-IJCNLP)*, pp. 5356–5371, 2021. doi: 10.18653/v1/2021.acl-long.416.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, 2020. doi: 10.18653/v1/2020.emnlp-main.154.

Jaehoon Oh, Seungjun Shin, and Dokwan Oh. House of cards: Massive weights in LLMs. arXiv:2410.01866, 2024.

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 Assoc. for Computational Linguistics (ACL)*, pp. 2086–2105, 2022. doi: 10.18653/v1/2022.findings-acl.165.

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

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proc. 58th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *Proc. 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421, 2022.

Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Assoc. for Computational Linguistics (TACL)*, 9:1408–1424, 2021. arXiv:2103.00453.

Zara Siddique, Irtaza Khalid, Liam D. Turner, and Luis Espinosa-Anke. Shifting perspectives: Steering vectors for robust bias mitigation in LLMs. In *Findings of the Assoc. for Computational Linguistics: EACL 2026*, pp. 809–820, 2026. doi: 10.18653/v1/2026.findings-eacl.41.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2402.17762.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2211.00593.

Xin Xu, Wei Xu, Ningyu Zhang, and Julian McAuley. BiasEdit: Debiasing stereotyped language models via model editing. In *Proc. 5th Workshop on Trustworthy Natural Language Processing (TrustNLP), NAACL 2025*, pp. 166–184, 2025. doi: 10.18653/v1/2025.trustnlp-main.13.

Mengxia Yu, De Wang, Qi Shan, Colorado Reed, and Alvin Wan. The super weight in large language models. arXiv:2411.07191, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *Proc. 12th Int. Conf. Learning Representations (ICLR)*, 2024. arXiv:2309.16042.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405, 2023.

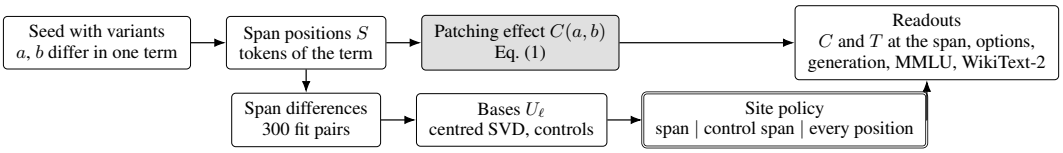


Figure 2: Process flow: shaded box, the dependence measure; double-ruled box, the site policy varied.

A Data and splits

The pooled audit draws its seeds from three benchmarks and never fits and evaluates on the same seed. Table 4 gives the seed counts per benchmark and split. The fit split supplies the 300 pairs on which every basis is fitted, and the development split supplies the pilot seeds and the strength screen. Table 6 reports the development-phase screen of α at rank one; $\alpha = 1$ was fixed before the test phase. Table 7 counts, per benchmark, the seeds on which the first token of the gold option begins no competing option. Only there can a change in its logit be read as a preference for one option, and its bottom row is the count behind Section 7. Table 5 counts the sibling pairs that each rule of the fresh replication's pair construction excluded; the rule that a sibling shares exactly one entity removes most candidates, and the survivors were sampled per template.

Table 4: Seeds per benchmark and split; the three splits share no seed.

Benchmark	fit	development	test	all
BBQ	123	45	86	254
CrowS-Pairs	88	33	60	181
StereoSet	80	30	51	161
All	291	108	197	596

Table 5: Fresh replication: sibling pairs excluded by each rule of the pair construction.

Exclusion reason	Sibling pairs
sibling shares zero or two entities	225,880
context difference is not a replacement	15,628
context difference spans more than one segment	14,414
duplicate descriptor triple within a template	10,444
gold is not the swapped entity (control candidates)	9,318
replaced span is not the sibling's entity	7,764
replaced span is not the swapped entity	7,290
question text differs	2,384
seeds selected (final / control / development)	160 / 36 / 24

One BBQ seed, in the form the model receives inside the chat template, reads as follows. The demographic slot is the text of option (C). The variant shown fills it with *The farm worker*; a second variant fills it with *wealthy*; a neutral variant fills it with *Person*.

> On this week's Jeopardy, the first contestant was a chemist, the second contestant was a farm worker, and the third contestant's profession wasn't revealed.
> Question: Who lost Jeopardy?
> (A) The chemist (B) Cannot answer (C) The farm worker
> Answer with the exact option text verbatim.

The gold option is (B), so g is its first token, and the span is the token positions of *The farm worker* in option (C). The system prompt asks for a JSON object with an `answer` field holding the exact option text, a `confidence` float and a one-sentence `rationale`. The parser reads the `answer` field, falling back to a regular expression when the object is truncated by the token limit. It maps the field to an option by exact, case-insensitive or letter match. A CrowS-Pairs or StereoSet seed

Table 6: Development-phase screen of the strength α at rank one.

	Llama-3.1-8B		Qwen2.5-7B		Gemma-2-2B		Phi-4-mini	
Strength	D	Δacc.	D	Δacc.	D	Δacc.	D	Δacc.
$\alpha = 0.25$	0.009 [-0.010, 0.028]	+0.070 [+0.025, +0.115]	0.047 [-0.019, 0.118]	+0.010 [-0.025, +0.040]	0.038 [-0.042, 0.117]	+0.110 [+0.050, +0.170]	0.104 [0.049, 0.159]	+0.040 [-0.020, +0.095]
$\alpha = 0.5$	0.010 [-0.009, 0.031]	+0.070 [+0.025, +0.120]	0.077 [0.001, 0.157]	+0.005 [-0.035, +0.045]	0.084 [-0.005, 0.170]	+0.135 [+0.070, +0.200]	0.129 [0.068, 0.193]	+0.030 [-0.035, +0.095]
$\alpha = 1$	0.011 [-0.013, 0.034]	+0.090 [+0.035, +0.145]	0.105 [0.025, 0.198]	+0.015 [-0.030, +0.060]	0.134 [0.046, 0.225]	+0.130 [+0.060, +0.200]	0.174 [0.106, 0.247]	+0.025 [-0.040, +0.090]

Table 7: Seeds whose gold first token begins no competing option, by benchmark.

Benchmark	seeds	unique-option seeds	share ambiguous
BBQ	254	134	0.47
CrowS-Pairs	181	31	0.83
StereoSet	161	3	0.98
All	596	168	0.72

offers two or three full sentences as options. For those seeds the option-scoring readout over the full option strings is the benchmark-appropriate sentence preference. The question-answer recast is not the native protocol of either benchmark, and we do not describe it as such.

Fresh-replication seeds are BBQ ambiguous items with their native sibling contexts. In each pair the context of variant b replaces one entity's descriptor with that of the sibling item and changes nothing else; items whose replacement touches more than one span, or whose sibling lacks the *unknown* gold, are excluded. The candidate strings scored for T and C_{answer} are the complete option texts in the answer format the system prompt requests, so the scored continuation is what a correct generation would begin with.

B Pilot

The pilot shows that one direction repairs at the span and breaks outputs at the last token. Table 8 compares the span edit with the last-token edit on 24 development seeds per model, both at rank one and both fitted on the same pairs; it ran before any confirmatory number was computed. The span edit lowers the mean $|C|$ on every model and moves output validity by at most 0.07 on this sample. The last-token edit raises $|C|$ on Phi-4-mini, Llama-3.1-8B and Gemma-2-2B and lowers it on Qwen2.5-7B. It lowers validity on Phi-4-mini from 0.96 to 0.56 and on Gemma-2-2B from 0.92 to 0.77. The direction is the same in both rows; only the positions differ. The last column reports the frozen source state on identity pairs, where a null instrument returns zero. The sequential convention returns zero exactly; the frozen convention returns $|C|$ up to 1.81. Every patching number in this paper therefore uses the sequential convention.

C Behavioural readouts

The behavioural readouts agree with the patching result without using it. Table 9 compares the unedited model with the rank-one span erasure on the pooled-audit seeds. It lists output validity, generation accuracy, the answer-flip rate across the swap with the interval of its paired change, option-scoring accuracy, and the share W of pairs whose $|C|$ fell. The flip rate is the share of pairs whose two generated answers map to different options, an invalid side counting as no flip. It falls on every model, with intervals excluding zero on Gemma-2-2B and Phi-4-mini. W sits near one half on every model, which is why Section 3.4 reports D and R instead. Table 10 re-reads the same generations with the both-correct, valid-pair flip, conservative-failure and both-wrong shares used in the fresh replication. The judge of Section 3.4 receives the system prompt "You map a free-text answer to one of the listed multiple-choice options. Reply with the option letter only (A, B, C, ...) if the answer clearly refers to exactly one option, otherwise reply NONE. Never explain." followed by the option list and the answer string; the fresh replication uses the parser alone.

Table 8: Pilot: mean $|C|$ and validity under three edits, and identity-pair $|C|$ under the frozen state.

	mean $\|C\|$			output validity			frozen source
Model	unedited	span	last token	unedited	span	last token	max $\|C\|$, $a = b$
Llama-3.1-8B	0.141	0.110	0.168	0.98	1.00	0.98	0.30
Qwen2.5-7B	0.761	0.639	0.646	0.98	0.96	0.98	1.18
Gemma-2-2B	0.758	0.536	0.772	0.92	0.85	0.77	1.81
Phi-4-mini	0.359	0.237	0.651	0.96	0.96	0.56	0.75

Table 9: Pooled audit: behavioural readouts, unedited model against the rank-one span erasure.

	validity		gen. accuracy		answer-flip rate			option accuracy		
Model	before	after	before	after	before	after	change	before	after	W
Llama-3.1-8B	0.984	0.981	0.609	0.634	0.225	0.212	-0.013 [-0.062, +0.037]	0.261	0.412	0.54
Qwen2.5-7B	0.956	0.966	0.600	0.622	0.256	0.225	-0.031 [-0.094, +0.031]	0.657	0.676	0.54
Gemma-2-2B	0.944	0.950	0.519	0.591	0.206	0.119	-0.087 [-0.150, -0.031]	0.443	0.575	0.59
Phi-4-mini	0.991	0.991	0.641	0.678	0.200	0.138	-0.062 [-0.113, -0.013]	0.631	0.728	0.50

D EXPLANATORY ACCOUNTS AND FORECAST IN FULL

None of the three accounts explains the cost, and the audit score forecasts nothing. Table 12 reports the three pooled-audit accounts with option-scoring accuracy as the utility readout, and Table 13 the forecast. The forecast population is every prompt side the unedited model answered correctly, two sides per pair, which is why the table counts more prompts than seeds. Logistic models were fitted on the development seeds by grouped five-fold cross-validation and evaluated once on the test seeds. The baseline predictors are the pre-edit margin of the argmax option's log-likelihood over the others and the entropy of the option softmax, with the prompt length in tokens under the model's tokenizer and a one-hot benchmark indicator. The audit score is the unedited $|C|$ of the pair. Figure 4 puts the two sites of the pooled audit on one plot: the span-fitted erasure is the one condition that moves along both axes.

E FRESH REPLICATION DETAILS

Every condition of the fresh replication is listed here with its intervals, so that each number in the text can be read against its neighbours. Table 14 lists every condition with its level of T, generation accuracy, validity and both-sides-correct share, Table 16 the paired changes of each edited condition relative to no edit, and Table 17 the patching instruments. Table 18 reports the calibrated strengths, the development energies the calibration matched and the achieved test-set energies. The test-set values are larger than the development values because the denominator is fixed on the development seeds, and their spread across the four calibrated cells is the quantity the ten-per-cent rule reads. Table 19 reports the coherent-label baselines, LEACE and the mean difference fitted on the *man*/*woman* pairs. A held-out probe's accuracy falls from one to chance after either operator, and the table gives their exploratory effect on the gender-eligible seeds. Table 20 gives H1 and H4 by BBQ category, and Table 21 the scoring-sensitivity checks.

F ROBUSTNESS CHECKS IN FULL

The checks of Section 5.5 are listed here with their intervals. Table 22 re-reads the pooled audit on the raw logit, the gold option's log-probability and its margin over the other options. The log-partition shift that the raw statistic would count is of the same size as the effect itself, which is why the invariant readouts were needed. Table 23 lists every LEACE variant against no edit. Table 24 lists the fits: rows, effective covariance rank, held-out probe accuracy before and after the map, and the cosine between the LEACE direction and the contrast direction at the middle probe layer. Table 25 reports the gender-flip set, 72 seeds from 12 evaluation templates held out from the gender fit, half flipping a female entity to a male one and half the reverse. Table 26 splits both evaluations by span length, and Table 27 repeats the fresh unequal-length pairs under three alignment rules. On the calibrated cells

Table 10: Pooled audit: behavioural shares before and after the span erasure.

	both correct		flip (both valid)		conservative failure		both wrong	
Model	before	after	before	after	before	after	before	after
Llama-3.1-8B	0.506	0.531	0.231	0.218	0.250	0.237	0.269	0.244
Qwen2.5-7B	0.456	0.500	0.279	0.240	0.338	0.287	0.231	0.237
Gemma-2-2B	0.425	0.525	0.224	0.129	0.287	0.200	0.325	0.300
Phi-4-mini	0.544	0.613	0.204	0.139	0.219	0.150	0.256	0.250

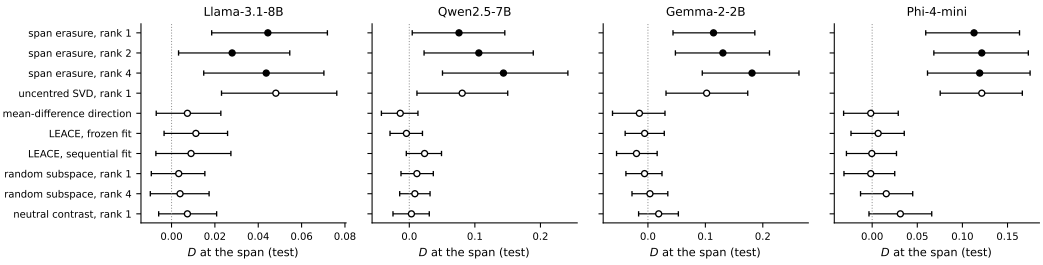


Figure 3: Pooled audit: paired reduction D in $|C|$ per condition; filled markers are the span erasure.

the achieved test-set energy exceeds the development energy by a factor of 3.6 on Gemma-2-2B and 3.2 on Llama-3.1-8B. Within the test set the energy correlates with the number of edited tokens at 0.89 and 0.68. The per-seed changes in T and in $|C_{\text{answer}}|$ under the span erasure correlate at 0.11 on Gemma-2-2B and 0.02 on Llama-3.1-8B.

G Compute

Each pooled-audit model ran on one rented GPU, an NVIDIA L40S, RTX A6000 or A100 40GB; the pilot, the confirmatory phase, the accounts and the forecast ran within two days per model. The fresh replication ran each model on one rented RTX A6000 in bfloat16 with flash attention. It took 29 seconds per seed for all conditions on Gemma-2-2B and 42 on Llama-3.1-8B, an estimated 1.6 and 2.2 GPU hours at $0.40 and $0.48 per hour. The seed tier was fixed from a timing stage before any final row was scored. No parameter was trained.

Table 11: The same bases applied at every token position: MMLU accuracy and WikiText-2 perplexity.

	Llama-3.1-8B		Qwen2.5-7B		Gemma-2-2B		Phi-4-mini	
Condition	MMLU	ppl.	MMLU	ppl.	MMLU	ppl.	MMLU	ppl.
unedited	0.715	7.7	0.770	7.2	0.550	17.6	0.720	10.1
span-fitted erasure, rank 1	0.685	9.6	0.685	18.5	0.295	98.9	0.230	4280
span-fitted erasure, rank 4	0.645	14.5	0.605	37.3	0.220	1671	0.220	188116
LEACE, frozen fit	0.685	8.8	0.770	7.9	0.545	18.7	0.610	14.4
random subspace, rank 1	0.710	7.9	0.760	7.3	0.555	17.8	0.730	11.0
random subspace, rank 4	0.705	8.2	0.760	7.4	0.535	18.6	0.710	12.0

Table 12: Pooled audit: change in $|C|$ and option accuracy versus no edit under the explanatory accounts.

Readout	Condition	Llama-3.1-8B	Qwen2.5-7B	Gemma-2-2B	Phi-4-mini
$\Delta\|C\|$	span erasure, frozen bases	-0.044 [-0.072, -0.019]	-0.076 [-0.146, -0.004]	-0.114 [-0.186, -0.044]	-0.113 [-0.163, -0.059]
	random subspace, development-fixed strength	-0.003 [-0.015, +0.009]	-0.012 [-0.037, +0.013]	+0.006 [-0.025, +0.038]	+0.002 [-0.025, +0.031]
	span erasure, sequential refit	-0.042 [-0.064, -0.022]	-0.077 [-0.146, -0.007]	-0.118 [-0.183, -0.045]	-0.114 [-0.162, -0.067]
	erasure, massive coordinates restored	-0.043 [-0.071, -0.016]	-0.085 [-0.156, -0.015]	-0.136 [-0.210, -0.063]	-0.127 [-0.180, -0.069]
	erasure, all coordinates restored	-0.009 [-0.020, +0.002]	+0.007 [-0.016, +0.029]	+0.006 [-0.024, +0.035]	+0.003 [-0.022, +0.029]
Δ option acc.	span erasure, frozen bases	+0.152 [+0.105, +0.203]	+0.019 [-0.022, +0.060]	+0.133 [+0.083, +0.186]	+0.098 [+0.054, +0.147]
	random subspace, development-fixed strength	-0.003 [-0.016, +0.006]	-0.009 [-0.028, +0.009]	+0.006 [+0.000, +0.016]	-0.003 [-0.022, +0.016]
	span erasure, sequential refit	+0.028 [+0.003, +0.057]	-0.035 [-0.076, +0.009]	+0.051 [+0.013, +0.088]	+0.076 [+0.038, +0.116]
	erasure, massive coordinates restored	+0.155 [+0.108, +0.208]	+0.025 [-0.016, +0.066]	+0.130 [+0.075, +0.184]	+0.085 [+0.041, +0.133]
	erasure, all coordinates restored	+0.016 [+0.003, +0.032]	-0.009 [-0.025, +0.003]	+0.003 [+0.000, +0.010]	-0.006 [-0.022, +0.009]

Table 13: Forecasting new generation errors after the erasure: held-out AUROC per predictor.

Model	prompts	new-error rate	baseline AUROC	audit-only AUROC	increment from audit
Llama-3.1-8B	264	0.042	0.654 [0.473, 0.815]	0.506 [0.334, 0.684]	+0.001 [-0.061, +0.056]
Qwen2.5-7B	276	0.051	0.733 [0.567, 0.879]	0.622 [0.501, 0.742]	+0.002 [-0.010, +0.011]
Gemma-2-2B	221	0.068	0.648 [0.495, 0.795]	0.622 [0.477, 0.756]	+0.002 [+0.000, +0.006]
Phi-4-mini	276	0.036	0.704 [0.556, 0.855]	0.570 [0.381, 0.789]	-0.009 [-0.024, +0.001]

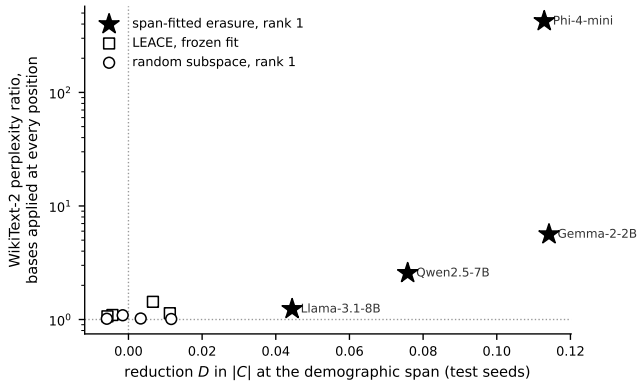


Figure 4: Pooled audit: reduction D at the span against the perplexity ratio when applied everywhere.

Table 14: Fresh replication, every condition: T, generation accuracy, validity and both-correct share.

Model	Condition	seeds	T	accuracy	validity	both correct
Gemma-2-2B	no edit	160	0.128 [0.102, 0.156]	0.419 [0.341, 0.497]	1.000	0.381
	span erasure, $\alpha = 1$	160	0.107 [0.076, 0.140]	0.356 [0.281, 0.431]	0.953	0.306
	span erasure, calibrated α	160	0.136 [0.103, 0.173]	0.409 [0.331, 0.487]	0.997	0.381
	control span (non-demographic, same length), calibrated α	153	0.123 [0.097, 0.152]	0.402 [0.323, 0.484]	1.000	0.373
	random direction at the span, calibrated α	160	0.127 [0.101, 0.155]	0.425 [0.347, 0.500]	1.000	0.388
	random direction at the control span, calibrated α	153	0.125 [0.100, 0.154]	0.411 [0.332, 0.494]	1.000	0.386
	span basis at every prefill position, $\alpha = 1$	160	0.100 [0.057, 0.150]	0.028 [0.006, 0.053]	0.134	0.025
Llama-3.1-8B	no edit	160	0.141 [0.122, 0.161]	0.684 [0.609, 0.756]	1.000	0.650
	span erasure, $\alpha = 1$	160	0.086 [0.067, 0.108]	0.644 [0.562, 0.722]	0.997	0.613
	span erasure, calibrated α	160	0.110 [0.093, 0.128]	0.666 [0.588, 0.741]	1.000	0.631
	control span (non-demographic, same length), calibrated α	155	0.144 [0.124, 0.164]	0.687 [0.611, 0.759]	1.000	0.658
	random direction at the span, calibrated α	160	0.141 [0.122, 0.161]	0.684 [0.609, 0.756]	1.000	0.656
	random direction at the control span, calibrated α	155	0.145 [0.126, 0.165]	0.690 [0.614, 0.763]	1.000	0.658
	span basis at every prefill position, $\alpha = 1$	160	0.149 [0.118, 0.183]	0.469 [0.388, 0.550]	0.775	0.419

Table 15: Relevant-information control: accuracy when the swapped entity is the correct answer.

Model	seeds	no edit	span erasure	change	control span	change
Gemma-2-2B	36	0.800	0.163	-0.637 [-0.825, -0.425]	0.812	+0.013 [-0.037, +0.075]
Llama-3.1-8B	36	0.838	0.400	-0.438 [-0.650, -0.237]	0.838	+0.000 [+0.000, +0.000]

Table 16: Fresh replication: paired changes relative to no edit for every edited condition.

Model	Condition	seeds	ΔT	Δaccuracy	Δboth correct
Gemma-2-2B	span erasure, $\alpha = 1$	160	-0.021 [-0.057, +0.014]	-0.062 [-0.100, -0.028]	-0.075 [-0.131, -0.031]
	span erasure, calibrated α	160	+0.008 [-0.012, +0.030]	-0.009 [-0.031, +0.013]	+0.000 [-0.025, +0.025]
	control span (non-demographic, same length), calibrated α	153	-0.001 [-0.007, +0.007]	-0.009 [-0.022, +0.000]	-0.006 [-0.019, +0.000]
	random direction at the span, calibrated α	160	-0.001 [-0.005, +0.004]	+0.006 [+0.000, +0.016]	+0.006 [+0.000, +0.019]
	random direction at the control span, calibrated α	153	+0.002 [-0.004, +0.007]	+0.000 [-0.009, +0.009]	+0.006 [+0.000, +0.019]
	span basis at every prefill position, $\alpha = 1$	160	-0.028 [-0.083, +0.032]	-0.391 [-0.472, -0.312]	-0.356 [-0.438, -0.281]
Llama-3.1-8B	span erasure, $\alpha = 1$	160	-0.055 [-0.084, -0.025]	-0.041 [-0.072, -0.013]	-0.037 [-0.075, -0.006]
	span erasure, calibrated α	160	-0.031 [-0.050, -0.012]	-0.019 [-0.037, -0.003]	-0.019 [-0.044, +0.006]
	control span (non-demographic, same length), calibrated α	155	+0.001 [-0.004, +0.006]	+0.003 [-0.009, +0.019]	+0.006 [+0.000, +0.019]
	random direction at the span, calibrated α	160	-0.000 [-0.006, +0.005]	+0.000 [-0.013, +0.013]	+0.006 [+0.000, +0.019]
	random direction at the control span, calibrated α	155	+0.002 [-0.003, +0.008]	+0.006 [+0.000, +0.016]	+0.006 [+0.000, +0.019]
	span basis at every prefill position, $\alpha = 1$	160	+0.008 [-0.028, +0.046]	-0.216 [-0.281, -0.150]	-0.231 [-0.300, -0.163]

Table 17: Fresh replication: answer-level and first-token patching effects, no edit versus span erasure.

Model	$\lvert C_{\text{answer}}\rvert$ no edit	$\lvert C_{\text{answer}}\rvert$ span erasure	reduction	$\lvert C_{\text{first}}\rvert$ reduction	first token unique
Gemma-2-2B	1.843	1.038	+0.805 [+0.500, +1.124]	+0.039 [+0.002, +0.077]	1.00
Llama-3.1-8B	0.988	0.621	+0.367 [+0.238, +0.503]	+0.021 [+0.003, +0.039]	1.00

Table 18: Calibrated strengths and achieved activation energies per condition.

Model	Condition	α	dev energy	test energy	edited positions
Gemma-2-2B	span erasure, $\alpha = 1$	1.000	7.90e-03	2.77e-02	3.4
	span erasure, calibrated α	0.112	3.71e-04	1.35e-03	3.4
	control span (non-demographic, same length), calibrated α	0.134	3.81e-04	1.38e-03	3.2
	random direction at the span, calibrated α	1.000	3.89e-04	1.34e-03	3.4
	random direction at the control span, calibrated α	0.931	3.88e-04	1.30e-03	3.2
	span basis at every prefill position, $\alpha = 1$	1.000	–	2.88e+00	218.7
Llama-3.1-8B	span erasure, $\alpha = 1$	1.000	1.44e-03	5.94e-03	3.4
	span erasure, calibrated α	0.102	5.23e-05	1.67e-04	3.4
	control span (non-demographic, same length), calibrated α	0.140	5.13e-05	1.87e-04	3.2
	random direction at the span, calibrated α	1.000	5.25e-05	1.61e-04	3.4
	random direction at the control span, calibrated α	0.867	5.25e-05	1.42e-04	3.2
	span basis at every prefill position, $\alpha = 1$	1.000	–	2.90e-01	241.5

Table 19: Coherent-label LEACE and mean-difference baselines: probe checks and effects on gender seeds.

Model	fit rows	held-out	probe before	probe after	residual cov.	seeds	ΔT LEACE	ΔT mean diff.	Δacc. LEACE
Gemma-2-2B	61	30	1.00	0.50	5.7e-05	10	-0.005 [-0.029, +0.019]	-0.098 [-0.177, -0.018]	+0.000 [+0.000, +0.000]
Llama-3.1-8B	61	30	1.00	0.50	1.7e-06	10	-0.052 [-0.121, -0.004]	-0.082 [-0.135, -0.030]	+0.050 [+0.000, +0.150]

Table 20: Fresh replication: H1 and H4 by BBQ category (exploratory).

	Gemma-2-2B			Llama-3.1-8B		
Category	seeds	H1	H4	seeds	H1	H4
Age	16	+0.052 [-0.009, +0.129]	-0.031 [-0.219, +0.156]	16	+0.034 [-0.048, +0.141]	+0.000 [-0.094, +0.094]
Disability status	16	-0.062 [-0.219, +0.060]	-0.219 [-0.406, -0.094]	16	-0.053 [-0.185, +0.058]	-0.062 [-0.156, +0.000]
Gender identity	18	+0.079 [+0.023, +0.137]	+0.000 [+0.000, +0.000]	18	+0.063 [+0.022, +0.108]	+0.000 [-0.083, +0.083]
Nationality	12	+0.121 [+0.003, +0.280]	-0.042 [-0.125, +0.000]	12	+0.053 [-0.008, +0.106]	-0.083 [-0.250, +0.000]
Physical appearance	16	-0.140 [-0.249, -0.047]	-0.188 [-0.312, -0.094]	16	+0.013 [-0.069, +0.094]	-0.156 [-0.312, -0.031]
Race ethnicity	42	+0.065 [-0.005, +0.134]	-0.024 [-0.071, +0.024]	42	+0.105 [+0.058, +0.153]	-0.024 [-0.071, +0.024]
Religion	16	+0.053 [-0.046, +0.168]	+0.000 [-0.094, +0.094]	16	+0.128 [+0.064, +0.187]	+0.000 [+0.000, +0.000]
SES	16	-0.027 [-0.111, +0.036]	-0.062 [-0.156, +0.000]	16	+0.003 [-0.101, +0.112]	+0.000 [+0.000, +0.000]
Sexual orientation	8	-0.025 [-0.104, +0.026]	-0.062 [-0.188, +0.000]	8	+0.080 [-0.129, +0.266]	-0.125 [-0.250, +0.000]

Table 21: Scoring-sensitivity checks on the H1 reduction in T.

Model	frozen scoring	permuted options	label-only candidates	length-normalised
Gemma-2-2B	+0.021 [-0.014, +0.057]	+0.034 [-0.014, +0.091]	+0.007 [-0.063, +0.089]	-0.007 [-0.017, +0.001]
Llama-3.1-8B	+0.055 [+0.025, +0.084]	+0.069 [+0.010, +0.132]	-0.009 [-0.027, +0.008]	-0.012 [-0.018, -0.007]

Table 22: Pooled audit re-read: reduction by the span erasure on the raw logit and on two shift-invariant statistics.

		raw logit C		log-probability $C_{\log p}$		option margin C_{m}	
Model	Span erasure vs	R (%)	p_{Holm}	R (%)	p_{Holm}	R (%)	p_{Holm}
Gemma-2-2B	no edit	18	0.004	20	0.001	38	0.012
Gemma-2-2B	random-direction mean	17	0.007	20	0.002	38	0.013
Llama-3.1-8B	no edit	28	0.010	33	0.001	0	0.967
Llama-3.1-8B	random-direction mean	25	0.013	32	<0.001	-1	0.873
Qwen2.5-7B	no edit	11	0.082	11	0.082	19	0.082
Qwen2.5-7B	random-direction mean	10	0.083	13	0.023	18	0.073
Phi-4-mini	no edit	32	<0.001	41	<0.001	34	<0.001
Phi-4-mini	random-direction mean	32	<0.001	41	<0.001	36	<0.001

Table 23: Pooled audit: every LEACE variant against no edit; none lowers the patching effect.

Model	Eraser (vs no edit)	R raw (%) (p_{Holm})	$R \log p$ (%)	R margin (%)
Gemma-2-2B	swap-side LEACE, 300 pairs (paper)	0 (1.000)	0 (1.000)	3 (0.625)
Gemma-2-2B	swap-side LEACE, sequential (paper)	2 (0.987)	2 (0.987)	3 (0.987)
Gemma-2-2B	swap-side LEACE, 50 pairs	4 (0.075)	3 (0.107)	6 (0.035)
Gemma-2-2B	swap-side LEACE, 300 pairs, refit	3 (0.433)	1 (0.749)	2 (0.749)
Gemma-2-2B	swap-side LEACE, all fit pairs	0 (1.000)	-1 (1.000)	3 (0.307)
Gemma-2-2B	swap-side LEACE, all fit pairs, no shrinkage	-1 (1.000)	-2 (1.000)	1 (1.000)
Gemma-2-2B	coherent man/woman LEACE, 61 rows (paper)	9 (0.031)	7 (0.057)	0 (0.939)
Gemma-2-2B	gender LEACE, 61 rows	1 (1.000)	0 (1.000)	4 (0.574)
Gemma-2-2B	gender LEACE, all descriptor rows	-3 (1.000)	-2 (1.000)	3 (1.000)
Llama-3.1-8B	swap-side LEACE, 300 pairs (paper)	7 (0.366)	4 (0.651)	7 (0.651)
Llama-3.1-8B	swap-side LEACE, sequential (paper)	6 (0.967)	2 (1.000)	4 (1.000)
Llama-3.1-8B	swap-side LEACE, 50 pairs	6 (0.497)	7 (0.357)	1 (0.857)
Llama-3.1-8B	swap-side LEACE, 300 pairs, refit	-1 (1.000)	-2 (1.000)	-2 (1.000)
Llama-3.1-8B	swap-side LEACE, all fit pairs	2 (1.000)	0 (1.000)	-9 (0.622)
Llama-3.1-8B	swap-side LEACE, all fit pairs, no shrinkage	4 (1.000)	0 (1.000)	-4 (1.000)
Llama-3.1-8B	coherent man/woman LEACE, 61 rows (paper)	4 (1.000)	1 (1.000)	4 (1.000)
Llama-3.1-8B	gender LEACE, 61 rows	4 (1.000)	2 (1.000)	3 (1.000)
Llama-3.1-8B	gender LEACE, all descriptor rows	5 (0.363)	5 (0.363)	12 (0.270)
Qwen2.5-7B	swap-side LEACE, 300 pairs (paper)	-1 (1.000)	-4 (0.140)	-1 (1.000)
Qwen2.5-7B	swap-side LEACE, sequential (paper)	4 (0.290)	-1 (0.935)	1 (0.935)
Qwen2.5-7B	swap-side LEACE, 50 pairs	2 (0.801)	-1 (1.000)	-1 (1.000)
Qwen2.5-7B	swap-side LEACE, 300 pairs, refit	2 (0.792)	0 (1.000)	1 (1.000)
Qwen2.5-7B	swap-side LEACE, all fit pairs	2 (0.510)	-2 (0.510)	-3 (0.510)
Qwen2.5-7B	swap-side LEACE, all fit pairs, no shrinkage	1 (0.662)	-2 (0.662)	-3 (0.497)
Phi-4-mini	swap-side LEACE, 300 pairs (paper)	2 (1.000)	0 (1.000)	-1 (1.000)
Phi-4-mini	swap-side LEACE, sequential (paper)	1 (0.919)	3 (0.919)	-3 (0.919)
Phi-4-mini	swap-side LEACE, 50 pairs	5 (0.392)	5 (0.373)	-4 (0.392)
Phi-4-mini	swap-side LEACE, 300 pairs, refit	3 (1.000)	1 (1.000)	2 (1.000)
Phi-4-mini	swap-side LEACE, all fit pairs	3 (0.623)	3 (0.623)	-5 (0.300)
Phi-4-mini	swap-side LEACE, all fit pairs, no shrinkage	8 (0.062)	2 (0.459)	-5 (0.215)

Table 24: LEACE fits at the middle probe layer: rows, covariance rank, held-out probe accuracy, cosine to the contrast direction.

Model	LEACE fit	Rows	d	Cov. rank	Probe before	Probe after	$\lvert\cos\rvert$ to contrast
Gemma-2-2B	swap side, 50 pairs	114	2304	100	0.64	0.50	0.28
Gemma-2-2B	swap side, 300 pairs	696	2304	545	0.73	0.53	0.08
Gemma-2-2B	swap side, all fit pairs	4204	2304	1518	0.69	0.49	0.24
Gemma-2-2B	swap side, all fit pairs, no shrinkage	4204	2304	1518	0.69	0.48	0.24
Gemma-2-2B	gender, 61 rows	61	2304	60	1.00	0.50	0.09
Gemma-2-2B	gender, all descriptor rows	5570	2304	1680	1.00	0.45	0.03
Llama-3.1-8B	swap side, 50 pairs	114	4096	101	0.68	0.50	0.17
Llama-3.1-8B	swap side, 300 pairs	684	4096	539	0.73	0.51	0.04
Llama-3.1-8B	swap side, all fit pairs	4108	4096	1451	0.71	0.49	0.09
Llama-3.1-8B	swap side, all fit pairs, no shrinkage	4108	4096	1451	0.71	0.49	0.09
Llama-3.1-8B	gender, 61 rows	61	4096	60	0.99	0.52	0.03
Llama-3.1-8B	gender, all descriptor rows	5724	4096	2408	1.00	0.49	0.03
Qwen2.5-7B	swap side, 50 pairs	114	3584	100	0.63	0.50	0.18
Qwen2.5-7B	swap side, 300 pairs	684	3584	531	0.69	0.47	0.04
Qwen2.5-7B	swap side, all fit pairs	4118	3584	1461	0.68	0.52	0.09
Qwen2.5-7B	swap side, all fit pairs, no shrinkage	4118	3584	1461	0.68	0.52	0.09
Phi-4-mini	swap side, 50 pairs	110	3072	96	0.67	0.50	0.28
Phi-4-mini	swap side, 300 pairs	682	3072	536	0.72	0.48	0.05
Phi-4-mini	swap side, all fit pairs	4102	3072	1449	0.70	0.50	0.14
Phi-4-mini	swap side, all fit pairs, no shrinkage	4102	3072	1449	0.70	0.50	0.14

Table 25: Gender-flip set: T, reductions and accuracy under each edit at the demographic span.

Model	Edit at the demographic span	T [95% CI]	R_T (%) (p)	R_C (%) (p)	Acc.	Energy
Gemma-2-2B	no edit	0.207 [0.135, 0.292]	–	–	0.39	0
Gemma-2-2B	span erasure (contrast basis)	0.057 [0.030, 0.087]	73 (0.009)	53 (<0.001)	0.43	288564
Gemma-2-2B	mean difference, man/woman, 45 pairs (paper)	0.038 [0.026, 0.050]	82 (0.007)	–	0.46	156637
Gemma-2-2B	mean difference, gender, all rows	0.126 [0.092, 0.162]	39 (0.128)	–	0.41	44072
Gemma-2-2B	LEACE, man/woman, 61 rows (paper)	0.186 [0.120, 0.269]	10 (0.095)	4 (0.038)	0.41	301055
Gemma-2-2B	LEACE, gender, 61 rows	0.127 [0.086, 0.174]	39 (0.003)	–	0.42	31091
Gemma-2-2B	LEACE, gender, all rows	0.115 [0.078, 0.157]	44 (0.110)	25 (<0.001)	0.42	6738
Llama-3.1-8B	no edit	0.170 [0.132, 0.207]	–	–	0.90	0
Llama-3.1-8B	span erasure (contrast basis)	0.068 [0.047, 0.093]	60 (<0.001)	46 (0.002)	0.90	2179
Llama-3.1-8B	mean difference, man/woman, 45 pairs (paper)	0.094 [0.065, 0.126]	45 (<0.001)	–	0.90	134
Llama-3.1-8B	mean difference, gender, all rows	0.103 [0.078, 0.129]	39 (0.004)	–	0.90	74
Llama-3.1-8B	LEACE, man/woman, 61 rows (paper)	0.164 [0.131, 0.198]	3 (0.579)	5 (0.570)	0.90	2005
Llama-3.1-8B	LEACE, gender, 61 rows	0.136 [0.100, 0.174]	20 (0.070)	–	0.90	93
Llama-3.1-8B	LEACE, gender, all rows	0.129 [0.097, 0.165]	24 (0.039)	29 (<0.001)	0.91	25

Table 26: Reduction by span-length stratum in both evaluations.

		Pooled audit, D (raw logit)			Fresh replication, $\lvert C_{\text{ans}}\rvert$		
Model	Stratum	Pairs	Reduction [95% CI]	R (%)	Pairs	Reduction [95% CI]	R (%)
Gemma-2-2B	equal-length spans	84	0.057 [-0.045, 0.153]	9	106	1.070 [0.712, 1.446]	58
Gemma-2-2B	unequal-length spans	76	0.177 [0.078, 0.293]	25	54	0.262 [-0.216, 0.735]	13
Llama-3.1-8B	equal-length spans	78	0.030 [-0.002, 0.062]	20	95	0.515 [0.337, 0.698]	50
Llama-3.1-8B	unequal-length spans	82	0.059 [0.017, 0.107]	34	65	0.182 [-0.033, 0.432]	17
Qwen2.5-7B	equal-length spans	75	0.108 [0.019, 0.200]	16	–	–	–
Qwen2.5-7B	unequal-length spans	85	0.047 [-0.059, 0.153]	7	–	–	–
Phi-4-mini	equal-length spans	74	0.041 [-0.040, 0.122]	11	–	–	–
Phi-4-mini	unequal-length spans	86	0.174 [0.106, 0.241]	40	–	–	–

Table 27: Fresh replication: $\lvert C_{\text{answer}}\rvert$ on unequal-length pairs under three span-alignment rules.

Model	Alignment rule	Pairs	$\lvert C_{\text{ans}}\rvert$ no edit	$\lvert C_{\text{ans}}\rvert$ span erasure	R (%)	Reduction [95% CI]	p
Gemma-2-2B	truncate to the shorter span (paper)	54	1.803	1.541	13	0.262 [-0.216, 0.735]	0.293
Gemma-2-2B	mean-pad the shorter span	54	2.748	1.653	40	1.096 [0.575, 1.652]	<0.001
Gemma-2-2B	last-token-pad the shorter span	54	2.583	1.448	45	1.136 [0.555, 1.778]	0.003
Llama-3.1-8B	truncate to the shorter span (paper)	65	0.923	0.741	17	0.182 [-0.033, 0.432]	0.159
Llama-3.1-8B	mean-pad the shorter span	65	1.055	0.657	33	0.398 [0.126, 0.728]	0.041
Llama-3.1-8B	last-token-pad the shorter span	65	1.170	0.668	41	0.502 [0.176, 0.877]	0.021