FAITHFUL PRUNING: PRESERVING PREDICTIONS AND INTERCHANGE INTERVENTION RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training pruning usually makes a network smaller while preserving its outputs. A smaller network would also make mechanistic interpretability studies cheaper, but these studies depend on more than the outputs. They apply interchange interventions to internal states and measure how the outputs change, and a pruned network that matches the original's predictions can still respond differently to these interventions. We study feed-forward (FFN) channel pruning that aims to preserve both the predictions and these responses. Removed channels are replaced by constants that fold exactly into the layer's bias, so the pruned layer is physically smaller. We introduce a paired-response objective that penalizes the pruned network's output error without the intervention, its output error with the intervention, and its error in the change the intervention causes. A greedy selector fits this objective from a small calibration set and can prune a single layer or every FFN of a network. Against an otherwise identical objective without the paired term, pairing lowers intervention-response error by 7.8% to 19.6% in all six single-layer GPT-2 language settings we test, carries over unchanged to Pythia-410m, and persists when the retained weights are also trained. When every FFN of GPT-2 is pruned, pairing helps on pronoun continuation and gives mixed results on indirect object identification. Without the intervention, the effect of pairing on output error varies across settings.

1 INTRODUCTION

Mechanistic interpretability studies a trained network by intervening on it. An interchange intervention copies an internal state computed on a donor input into the forward pass on a base input and reads the resulting change in the outputs (Geiger et al., 2021; Wang et al., 2023). Running these studies on a pruned copy of the network would make them cheaper. Standard post-training pruning, however, fits the pruned copy only to the original network's outputs on ordinary inputs, and its objective says nothing about how the network responds to interventions. A pruned network can match those outputs closely and still respond differently when intervened on, because removed channels can carry part of the path from the intervened state to the output. For a pruned network to stand in for the original in an interpretability study, it should preserve both the original's predictions and its responses to the interventions the study will run, including responses that lead to wrong predictions.

We study this in a restricted post-training setting. In a chosen feed-forward (FFN) layer of a fixed network, selected channels are replaced at every token by constants, and their outgoing contributions fold exactly into a bias, so the layer becomes physically smaller while the rest of the network is unchanged. Replacing pruned channels by their mean activation and folding the result into the bias appears in earlier structured pruning (Berisha et al., 2025), and causal mechanism reduction (CMR) (Asiaee, 2026b) extends this compilation to other constants and to affine functions of the retained channels. Both select channels to preserve outputs on ordinary inputs. Here the intervention enters the fitting objective. We call the resulting procedure *paired-response reduction* (Figure 1).

A natural baseline fits the pruned network's outputs on both natural and intervened inputs. We call it the *same-data control*, because it sees exactly the data our method sees. Its loss, however, does not target what an interpretability study reads off, which is the change in output that the intervention

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.

causes. Two pruned networks with the same output errors in size can differ sharply in that change: if the natural and intervened outputs are off by the same amount, the errors cancel and the effect is preserved, and if they are off in opposite directions, the errors add. We therefore add a penalty on the difference between the pruned and dense networks' intervention-induced logit changes. Logits are centered first, which removes a common offset that does not affect probabilities. To fit this objective, a greedy selector removes channels one at a time using derivatives of the final outputs, and sets each removed channel's constant to offset the error left by the channels removed before it. Applied block by block, the same selector prunes every FFN of a network. The retained weights stay frozen, so the only fitted quantities are which channels to remove and one constant per removed channel.

We test this mainly on GPT-2 with two language tasks, indirect object identification (IOI) and pronoun continuation. In each, the intervention copies a residual-stream state from a donor prompt into a base prompt. The main measure is intervention-effect error, the gap between the pruned and dense networks' intervention-induced logit changes, and we also track KL divergence from the dense network on ordinary, unintervened inputs. The comparison is always against the same-data control. When a single FFN layer is pruned, pairing lowers intervention-effect error in all six settings we test (four IOI configurations and two data draws for pronoun continuation), by 7.8% to 19.6%, and every 95% interval excludes zero. The gain carries over unchanged to Pythia-410m, a second model family (the IOI result there is exploratory; Section 5.4), and it persists when the retained FFN weights are also trained on the same data. When every FFN of GPT-2 is pruned, the outcome depends on the task: on pronoun continuation pairing lowers effect error, KL, and both task-level errors, while on IOI it matches the control on effect error and lowers KL but worsens the task-level errors. Pairing does not reliably help on ordinary, unintervened inputs. With frozen weights, single-layer KL falls at the primary IOI setting but rises in the other settings, and these KL differences are not stable across refits. On a vision transformer, the gain is not distinguishable from zero.

Our contributions are:

1. A paired-response objective for post-training pruning. Besides the pruned network's output errors on natural and intervened inputs, e^0 and e^I, it penalizes their difference $e^I - e^0$, which is the error in the change the intervention causes (Figure 1A; Section 3.2).
2. A greedy procedure that fits this objective by removing channels one at a time and choosing a replacement constant for each, then compiles the result exactly into a smaller FFN. Applied layer by layer, it prunes a whole network (Sections 3.3–3.5).
3. A controlled evaluation against a same-data control on two language model families and a vision transformer, including whole-network pruning, a trained-student comparison, alternative fidelity metrics, and stability analyses, reporting where ordinary fidelity improves and where it degrades (Section 5).

2 Related work

Our method draws on three lines of work. Post-training pruning makes a network smaller while preserving its outputs on ordinary inputs. Distillation and causal abstraction ask a model to match another model's behavior under interventions, usually by training a new model. Circuit discovery prunes a network to explain a task rather than to deploy a smaller model. We take the compilation step from the first line, the use of interchange interventions as a fitting target from the second, and interpretability tasks such as IOI from the third.

Pruning and compensation. Early pruning methods rank parameters by how much the training loss would rise if they were removed, using a second-order approximation. Optimal Brain Damage uses a diagonal curvature estimate (LeCun et al., 1989), and Optimal Brain Surgeon uses the full inverse curvature and also adjusts the surviving weights to compensate (Hassibi & Stork, 1992); WoodFisher makes this inverse-curvature approach practical for modern networks (Singh & Alistarh, 2020). A related line transfers a removed unit's contribution to the units that remain, as in data-free neuron surgery (Srinivas & Babu, 2015) and Neuron Merging (Kim et al., 2020). Replacing a removed channel by its mean activation, rather than by zero, and folding that constant into the next layer's bias is the simplest form of such compensation; variance-based pruning (VBP) pairs it with an activation-variance ranking (Berisha et al., 2025), and CMR extends the folding to other constants and to affine replacements (Asiaee, 2026b). NISP instead propagates importance scores backward from

the network's output before pruning and fine-tuning (Yu et al., 2018). At the scale of large language models, SparseGPT and Wanda remove individual weights (Frantar & Alistarh, 2023; Sun et al., 2024), and LLM-Pruner removes larger structures such as attention heads and FFN channels (Ma et al., 2023). All of these methods aim to preserve behavior on ordinary inputs, and none places an intervention in the pruning objective. We use two of them as pruning baselines: VBP, and CMR-local, which deletes FFN channels using activation variance and outgoing-weight magnitude. Our selector also replaces channels by bounded constants, but chooses the channels and constants with final-logit derivatives and the paired-response objective.

Distillation and causal abstraction. Distillation trains a smaller student to match a teacher. Jacobian matching adds the teacher's input–output derivatives to the target and connects derivative matching to distillation on noisy inputs (Srinivas & Fleuret, 2018). We also use Jacobians, but only to approximate the effect of replacing a channel; the quantity we fit is a finite change caused by an intervention. Interchange intervention training asks a network to behave like a high-level causal model when internal states are swapped between inputs (Geiger et al., 2022), and DIITO uses the same idea to train a distilled student under aligned interchange interventions (Wu et al., 2022). Causal-abstraction theory gives these methods a common footing, allowing mechanism transformations and graded notions of fidelity (Geiger et al., 2025). Fidelity under intervention is itself an estimate from finite samples, and Asiaee (2026a) gives confidence intervals and anytime-valid confidence sequences for such interventional evaluations. Unlike DIITO, we do not train a student: the reference network stays fixed, and we fit only which channels to remove and the constants that replace them.

Circuit discovery and intervention-sensitive pruning. Circuit discovery looks for a small part of a network that accounts for its behavior on a task. Attribution patching estimates each component's importance from activation differences and downstream gradients (Syed et al., 2024). Edge Pruning learns a sparse set of edges using counterfactual replacements and measures faithfulness through ablation effects (Bhaskar et al., 2024), and DiscoGP jointly prunes graph edges and parameters to find task modules (Yu et al., 2025). CausalGate, closer to compression, distills static module gates from the output KL divergence observed after deleting branches (Nair et al., 2026). Our setting differs in what is held fixed and what is produced: the intervention port is the same for every method, the objective is the full-output paired contrast, and the result is a compiled, physically smaller layer. Our contribution is this procedure and the controlled comparison; we do not claim priority for intervention-sensitive pruning in general.

3 METHOD

3.1 SETUP AND EXACT FFN COMPILATION

Let D be a fixed reference network with final logits $z_D(x) \in \mathbb{R}^q$. For a target FFN of activation width p and residual width d, write

$$a = \sigma(W_1 h + b_1), \qquad u = W_2 a + b_2, \tag{1}$$

where $W_1 \in \mathbb{R}^{p\times d}$, $W_2 \in \mathbb{R}^{d\times p}$, and σ acts coordinatewise. The same FFN operates at every token. Partition its channels into a removed set S and retained set K, with $|K| = k$. Following Asiaee (2026b), replacing a_S by a constant vector c compiles to

$$W_1' = W_{1,K,:}, \quad b_1' = b_{1,K}, \quad W_2' = W_{2,:,K}, \quad b_2' = b_2 + W_{2,:,S}c. \tag{2}$$

Compilation changes only the size of the network, not the function it computes. The original network run with a_S clamped to c and the compiled network $R_{S,c}$ are two implementations of the same function, one smaller than the other: the FFN output u is identical for every input, so every later layer, including the nonlinear ones, receives identical inputs. Unlike the clamped network, it needs no clamping operation at run time. The fold is exact because W_2 is the only layer that reads a; if another layer also read the removed activations, its bias would need the same substitution. All networks run in evaluation mode, with dropout disabled.

A constant replaces each removed channel by a single value and discards how that channel varies across inputs. When a removed channel is strongly correlated with retained channels in the same layer, a better replacement predicts it from them. CMR (Asiaee, 2026b) allows such affine replacements,

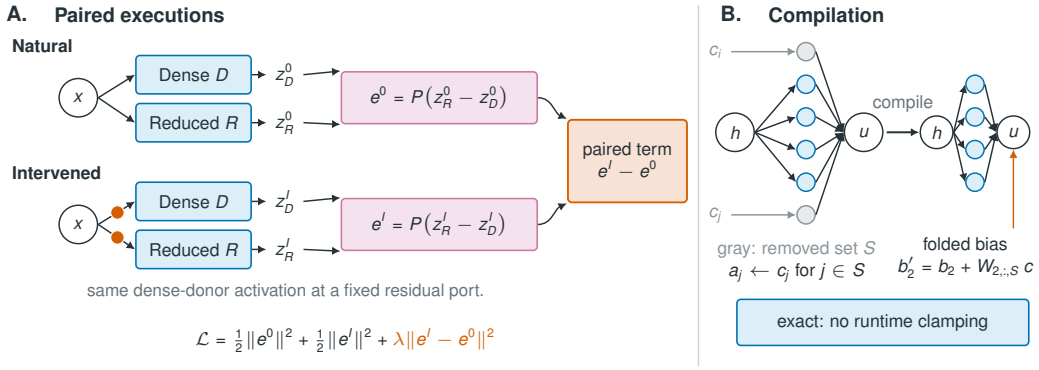


Figure 1: **Paired-response reduction compares a reduced and a dense model under the same natural input and the same upstream intervention. A.** In the natural execution, the dense network D and the reduced network R receive the same input x and produce logits z_D^0 and z_R^0. In the intervened execution, both also receive the same dense-donor activation at a fixed residual port (orange markers) and produce z_D^I and z_R^I. The matrix $P = I_q - \mathbf{1}\mathbf{1}^\top/q$ centers the logits, removing a common offset that does not change softmax probabilities, so e^0 and e^I are the centered output errors of the two executions. Their difference $e^I - e^0$ is the error in the logit change that the intervention causes. The loss $\mathcal{L}$ adds this paired term, weighted by λ, to the two output errors; $\lambda = 0$ gives the same-data control. **B.** Gray channels form the removed set S: they no longer receive the FFN input h, and each is replaced by its own bounded constant (c_i, c_j) at every token. Because these channels feed the affine output map, their fixed contribution $W_{2,:,S}c$ can be added to the output bias once. This compilation step yields a smaller dense FFN that computes exactly the same function as the clamped one, with no clamping at run time.

$a_S = \beta + Ba_K$ with $\beta \in \mathbb{R}^{p-k}$ and $B \in \mathbb{R}^{(p-k)\times k}$, which keep the same W_1', b_1' and fold into the output weights as well as the bias:

$$W_2' = W_{2,:,K} + W_{2,:,S}B, \qquad b_2' = b_2 + W_{2,:,S}\beta. \tag{3}$$

This is the compensation algebra of Optimal Brain Surgeon and Neuron Merging, and it reaches no network that unrestricted adjustment of the outgoing weights could not. Our experiments use only constants, so the retained weights, including $W_{2,:,K}$, stay unchanged. Appendix A proves both folds.

Each calibration pair $(x_i, \tilde{x}_i)$ specifies a base input, a donor, and a fixed intervention I_i. Let n be the number of calibration pairs, and let $\mathbb{E}_i$ below denote the empirical average $n^{-1}\sum_{i=1}^n$. We cache the dense donor's residual at a common uncompressed port before the target block. Natural and intervened executions have superscripts 0 and I. Every reduced model receives the same donor tensor at that port, which remains available independently of the chosen channel mask.

3.2 Paired-response centered-logit objective

Let $P = I_q - \mathbf{1}\mathbf{1}^\top/q$. Define centered fidelity errors $e_i^0 = P(z_R^0 - z_D^0)$, $e_i^I = P(z_R^I - z_D^I)$. For a width budget $|S| = p - k$, our target loss is

$$\mathcal{L}_\lambda(S, c) = \frac{1}{q}\mathbb{E}_i\left[\frac{1}{2}\left\|e_i^0\right\|_2^2 + \frac{1}{2}\left\|e_i^I\right\|_2^2 + \lambda\left\|e_i^I - e_i^0\right\|_2^2\right], \qquad c_j \in [\ell_j, h_j]. \tag{4}$$

The bounds are the minimum and maximum observed activation of channel j over natural and patched calibration executions. They fix the permitted replacement range before test evaluation. Each constant applies at all valid token positions. The loss preserves the reference's relative logits, including its errors; no label enters the selector.

Centering removes a common logit offset, which leaves softmax probabilities unchanged, so the loss ignores differences that cannot change the model's predicted probabilities. The first two terms fit outputs on the two executions. The third fits the response to the intervention because

$$e_i^I - e_i^0 = P\big[(z_R^I - z_R^0) - (z_D^I - z_D^0)\big].$$

Equivalently, the loss per pair is

$$\frac{1}{4q}\left\|e_i^0 + e_i^I\right\|_2^2 + \frac{1 + 4\lambda}{4q}\left\|e_i^I - e_i^0\right\|_2^2.$$

Thus pairing reweights contrast error relative to common error. The same-data control sets $\lambda = 0$ while using the same paired observations. The loss does not constrain the KL divergence between the pruned and dense networks' predictions on ordinary inputs, so a lower loss can come with a higher KL. We therefore report this KL separately as a check on ordinary-input fidelity (Section 4).

3.3 Cross-channel greedy selection with conditional constants

Let $J^b_{i,t,j} = \partial z^b_D / \partial a^b_{i,t,j} \in \mathbb{R}^q$ for $b \in \{0, I\}$. We compute it by backpropagation through every layer after the FFN, so it captures how the rest of the network (attention, normalization, and residual connections) transforms a change in the channel, not only the effect through the next layer's weights. Define

$$u^b_{i,j} = \sum_t P J^b_{i,t,j}, \qquad v^b_{i,j} = \sum_t P J^b_{i,t,j} a^b_{i,t,j}, \qquad e^b_i \approx \sum_{j \in S} (c_j u^b_{i,j} - v^b_{i,j}). \tag{5}$$

Token contributions are summed before squaring, so a channel's cross-token interactions remain. Stacking the calibration pairs and loss components gives

$$U_j = \frac{1}{\sqrt{nq}} \operatorname{stack}_i \begin{bmatrix} u^0_{i,j}/\sqrt{2} \\ u^I_{i,j}/\sqrt{2} \\ \sqrt{\lambda}(u^I_{i,j} - u^0_{i,j}) \end{bmatrix}, \quad V_j = \frac{1}{\sqrt{nq}} \operatorname{stack}_i \begin{bmatrix} v^0_{i,j}/\sqrt{2} \\ v^I_{i,j}/\sqrt{2} \\ \sqrt{\lambda}(v^I_{i,j} - v^0_{i,j}) \end{bmatrix}. \tag{6}$$

The quadratic proxy retains cross-channel terms: $Q_\lambda(S, c) = \left\| \sum_{j \in S} (c_j U_j - V_j) \right\|_2^2$. We use projected versions of these vectors in practice (Section 3.4). Starting from $S = \varnothing$ and residual $r = 0$, each remaining candidate uses

$$c^{\mathrm{alg}}_j(r) = \begin{cases} \operatorname{clip}_{[\ell_j, h_j]} \dfrac{U_j^\top (V_j - r)}{\|U_j\|_2^2}, & \|U_j\|_2^2 > \tau, \\ \operatorname{clip}_{[\ell_j, h_j]}(\bar{a}_j), & \|U_j\|_2^2 \le \tau, \end{cases} \qquad \Delta_j = \left\| r + c^{\mathrm{alg}}_j(r) U_j - V_j \right\|_2^2 - \|r\|_2^2 . \tag{7}$$

Here $c^{\mathrm{alg}}_j(r)$ is the best constant for channel j given the error r left by the channels already selected, clipped to the channel's observed range, and Δ_j is the resulting change in the proxy. If $\|U_j\|_2^2$ falls below a small threshold τ, we use the calibration mean $\bar{a}_j$ instead; this fallback was never triggered in the reported fits. At each step we remove the channel with the smallest Δ_j, add its contribution to r, and repeat until $p - k$ channels are removed. Δ_j can be negative, because a new channel's contribution can cancel error left by earlier ones. After selection, two passes of coordinate descent refit each constant with the others held fixed. The derivatives are computed once and cached, so selection and refitting need no further passes through the model. Appendix B gives pseudocode.

The procedure is a heuristic. Choosing S and c together is a combinatorial problem, and neither the selected channels nor the constants are guaranteed to be globally optimal. The proxy is also a local approximation, so lowering it need not lower the exact loss; every selected model is therefore evaluated with exact forward passes on held-out pairs. The proxy is the standard Gauss–Newton approximation: the loss and its gradient are zero before any channel is replaced, because the reduced network then equals the dense one, so the curvature comes entirely from the output Jacobians (Appendix A).

3.4 Sketch approximation via random output projections

The vectors $u^b_{i,j}$ and $v^b_{i,j}$ in Equation (5) have one entry per output logit. Computing them exactly takes one backward pass per logit, which is impractical when the outputs are a language model's vocabulary of tens of thousands of tokens. We therefore estimate the proxy from a few random combinations of the logits: each combination, or probe, costs one backward pass, and for any fixed choice of channels and constants the probes give an unbiased estimate of the proxy. For each pair and probe s, draw independent Rademacher signs $\xi_{i,s} \in \{-1, +1\}^q$ and differentiate $\xi_{i,s}^\top P z^b_D / \sqrt{q}$. The same probe is used for both executions and all objective variants, so one vector–Jacobian product supplies projected sensitivities for every candidate and preserves the paired cross term. Summing over valid tokens and stacking as in Equation (6), with normalization $1/\sqrt{nM}$, gives for any fixed error vector e

$$\mathbb{E}_\xi \left[(\xi^\top P e / \sqrt{q})^2 \right] = \|Pe\|_2^2 / q. \tag{8}$$

This estimate is unbiased for a fixed candidate independent of the probes, while selected-model evaluation uses exact held-out logits. We use $M = 16$ probes and $n = 128$ pairs; Appendix A.4 gives the variance bound and rank ceilings.

3.5 WHOLE-MODEL REDUCTION

For a model with L blocks, we apply the selector sequentially from block 0 through $L - 1$. At each block, derivatives are taken around the current partially reduced model while the original dense model remains the output target. The greedy procedure therefore starts from r_0, the stacked centered projected reduced-minus-dense error of the current chain, instead of $r = 0$. We compile the chosen replacements after each block before linearizing the next one; Appendix C records the frozen block order and whole-model protocol.

4 EXPERIMENTAL SETUP

Models, tasks, and interventions. We evaluate fixed DeiT-Tiny (Touvron et al., 2021) and GPT-2 (Radford et al., 2019) checkpoints on ImageNet-100 and indirect object identification (IOI) (Wang et al., 2023). The primary single-block setting removes 25% of block 8's FFN channels; 50% removal tests a heavier budget. A transferred occupation-pronoun task (Zhao et al., 2018; Vig et al., 2020) uses GPT-2 block 8, 25% removal, and $\lambda = 4$, while the whole-model study reduces all 12 GPT-2 FFNs on IOI and pronoun continuation. To test transfer across model families, we apply the same frozen setting to Pythia-410m (Biderman et al., 2023) (Section 5.4).

For DeiT, a cross-class donor supplies the block-input residual at four image tokens; for IOI, a role-swapped donor supplies the S2-position residual; for pronoun continuation, a donor differing in one occupation supplies that position's residual; every port precedes the target block and is shared across methods. This assay measures the final-output response to an upstream residual patch; it does not intervene on removed channels or identify an internal causal abstraction. The donor state always comes from the dense model, so intervention fidelity here means fidelity under a dense-model donor supplied at one port; in the whole-model study, the patch at block 8 also overrides the reduced blocks 0–7 at the patched position.

Calibration and separation. Each draw uses 128 calibration, 64 diagnostic, and 256 held-out pairs, with disjoint vision image pools and disjoint language name, occupation, clause, context, and template pools as applicable; Appendix C gives the indices. All selection rules were fixed before test evaluation.

Comparators. All methods share checkpoint, intervention tensors, site, data, and retained width. VBP uses activation variance and mean replacement; CMR-local (Asiaee, 2026b) adds outgoing-weight magnitude; E2E-natural fits natural outputs; E2E-augmented is the same-data control ($\lambda = 0$); and E2E-paired adds the paired penalty. Augmented-reconstruction refits local outgoing FFN weights on the same natural and intervened pairs. Retained learned weights remain frozen outside that reconstruction diagnostic and the explicitly post hoc trained-student analysis.

Metrics and uncertainty. The primary endpoint is $M_{\text{effect}} = \mathbb{E}\left\|e^I - e^0\right\|_2^2/q$, and ordinary fidelity is $K_0 = \mathbb{E}\,\mathrm{KL}(\mathrm{softmax}(z_D^0)\|\,\mathrm{softmax}(z_R^0))$ in nats. Task-effect MAE measures error in the intervention-induced IO-minus-subject or he-minus-she contrast; IIA is full-vocabulary argmax agreement under the common patch. We also report normalized response error, $\mathrm{NRE} = \sum_i \|\Delta_{R,i} - \Delta_{D,i}\|_2^2 \,/\, \sum_i \|\Delta_{D,i}\|_2^2$ with $\Delta_{M,i} = P(z_{M,i}^I - z_{M,i}^0)$, and decision-token log-probability-effect error, the error in intervention-induced log-probability changes on the union of the dense model's top-10 natural and intervened tokens. Intervals use 1,000 paired bootstrap resamples stratified within language templates. They are conditional on the checkpoint, draw, site, and input family, are not adjusted for multiplicity, and do not resample repeated names or occupations as clusters.

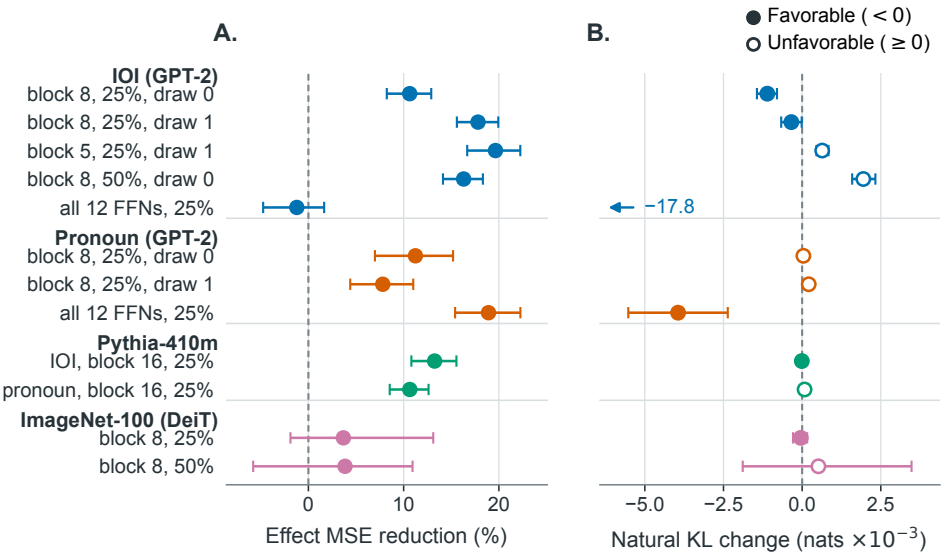


Figure 2: **Pairing lowers intervention-effect error in every single-layer language setting, while its effect on ordinary KL varies. A.** Percentage reduction in centered effect MSE relative to the same-data control ($\lambda = 0$); positive values favor pairing. **B.** Paired-minus-control natural KL in nats $\times 10^{-3}$; negative values favor pairing, and filled markers mark favorable point estimates. The whole-model IOI point (-17.8) is clipped at the axis edge. Rows cover single-block GPT-2 settings, 25% removal in all 12 GPT-2 FFNs, Pythia-410m at block 16, and DeiT-Tiny; bars are conditional 95% paired bootstrap intervals. The Pythia IOI row is exploratory because its prespecified derivative-stability check failed (Section 5.4).

Table 1: **The paired objective helps after matched-data student training.** Frozen rows retain the selector's compiled FFN; trained rows optimize the retained FFN on the same 128 pairs. Effect MSE and natural KL are multiplied by 10^3; task MAE is unscaled.

		Draw 0			Draw 1		
Regime	λ	Effect	KL	Task MAE	Effect	KL	Task MAE
Frozen	0	0.581	4.83	0.0890	0.606	5.52	0.0476
Frozen	4	0.519	3.72	0.0718	0.498	5.18	0.0454
Trained	0	0.521	4.23	0.0747	0.556	4.76	0.0483
Trained	4	0.439	2.27	0.0543	0.439	4.14	0.0435

5 Results

5.1 Single-block comparison

Figure 2 summarizes the comparison; Appendix Tables 5, 17, 6, and 10 give the raw values. Pairing lowers centered effect MSE in all four IOI configurations and both pronoun draws, with every interval excluding zero, whereas both DeiT intervals cross zero.

Natural KL falls at the primary IOI setting on both draws, by 1.11×10^{-3} nats (22.9%) and 3.43×10^{-4} nats (6.2%), and rises in the other four single-block language configurations: 6.38×10^{-4} (34.6%) at block 5, 1.95×10^{-3} (31.4%) at 50% removal, and 3.94×10^{-5} (3.1%) and 2.10×10^{-4} nats (16.8%) on the two pronoun draws.

Augmented reconstruction, which refits the FFN's own outputs rather than the network's final outputs, is much worse than pairing on language and competitive on vision (Appendices G, J, and F).

Appendix I reports how closely the derivative-based proxy tracks the exact loss, along with implementation checks, including that each compiled network reproduces its clamped counterpart to numerical tolerance.

5.2 Is the gain specific to frozen selection?

This post hoc analysis asks whether the gain survives when the retained FFN weights are also trained on the same 128 calibration pairs. After training, $\lambda = 4$ beats $\lambda = 0$ on effect MSE by 15.9% (95% interval 13.6–18.1%) and 20.9% (18.7–22.9%), and on natural KL by 46.4% (41.6–50.9%) and 12.9% (7.1–18.3%) for draws 0 and 1. Trained students also improve effect MSE and natural KL over frozen selection at both λ values; Appendix Table 7 gives absolute and relative intervals. Each trained cell inherits the mask selected under its own objective, so the comparison covers mask selection and training together and does not isolate the training loss at a fixed mask.

Across a range of λ, nearly every $\lambda > 0$ lowers both effect MSE and KL relative to $\lambda = 0$ on both draws. The only exceptions are draw 0 at $\lambda = 0.25$, where KL rises 1.1%, and $\lambda = 0.5$, where effect MSE rises 6.1%; draw 1 has no exception (Appendix Figure 3).

5.3 Whole-model reduction

Removing 25% of every FFN removes 14.2 million parameters, or 11.4% of GPT-2. On IOI, paired and the same-data control tie on effect MSE, 7.13×10^{-3} versus 7.05×10^{-3}, with difference $+8.60 \times 10^{-5}$ (95% interval $[-1.20, 3.29] \times 10^{-4}$); paired lowers natural KL by 0.0178 nats (31.6%; interval $[-0.0215, -0.0142]$) but has worse contrast MAE, 0.345 versus 0.232. On pronoun continuation, pairing lowers effect MSE from 0.0216 to 0.0176, a difference of -4.09×10^{-3} (18.9%; interval $[-5.07, -3.16] \times 10^{-3}$), and lowers KL by 3.94×10^{-3} nats (10.3%; interval $[-5.53, -2.36] \times 10^{-3}$).

Table 2: **Paired effect MSE and natural KL remain below VBP after sequential reduction of all GPT-2 FFNs.** Effect MSE and natural KL are multiplied by 10^3; task-contrast MAE is unscaled. Reconstruction rows appear in Appendix Table 6.

Task	Removed	Method	Effect	KL	MAE
IOI	25%	E2E-paired	7.13	38.7	0.345
		E2E-augmented	7.05	56.5	0.232
		VBP	10.2	66.0	0.341
	50%	E2E-paired	24.6	949	0.536
		E2E-augmented	27.9	675	0.552
		VBP	29.4	1,610	0.606
Pronoun	25%	E2E-paired	17.6	34.2	0.110
		E2E-augmented	21.6	38.1	0.128
		VBP	39.5	71.7	0.185
	50%	E2E-paired	62.1	122	0.260
		E2E-augmented	66.1	124	0.272
		VBP	119	151	0.308

At 50% FFN removal, which removes 22.8% of GPT-2, pairing lowers effect MSE by 11.8% on IOI and 6.0% on pronoun continuation, with both intervals excluding zero. IOI KL rises by 0.274 nats (40.6%; interval $[0.221, 0.329]$), while pronoun KL falls by 2.94×10^{-3} nats (2.4%; interval $[-5.96, 0.0457] \times 10^{-3}$) and its interval crosses zero. Paired effect MSE is well below VBP in every task-budget cell.

Augmented-reconstruction effect MSE spans 0.0278–0.180 and its natural KL spans 0.279–3.03, exceeding paired in all four cells; Appendix Table 6 reports every row.

5.4 Second model family: Pythia-410m

We apply the GPT-2 setting unchanged ($\lambda = 4$, 25% removal, no retuning) to Pythia-410m (Biderman et al., 2023), a 24-block GPT-NeoX model with parallel attention and FFN branches, at block 16, the same relative depth as GPT-2 block 8, with IOI and pronoun pairs regenerated for its tokenizer. Pairing lowers effect MSE by 13.3% on IOI (interval 10.8–15.5%) and 10.6% on pronoun continuation (8.5–12.6%). Natural KL is unchanged on IOI (-1.0×10^{-5} nats, -0.8%; interval -7.0% to $+5.2\%$) and rises by 7.8×10^{-5} nats (8.8%) on pronoun continuation. On IOI, E2E-paired has the lowest effect MSE, natural KL, and task-contrast MAE of the six methods (Appendix F). On pronoun continuation, the finite-difference stability check passes (relative RMS 6.0–7.8%). On IOI it exceeds our prespecified 10% threshold (13.4–14.4%), so we report the Pythia IOI result as exploratory.

5.5 Task-level and probability-sensitive metrics

Task-contrast MAE, the error in the intervention-induced change of the task's logit contrast (Section 4), falls with pairing in all four IOI cells, although VBP has a lower MAE than paired at the primary IOI cell (0.0536 versus 0.0718). IIA moves in both directions. On pronoun continuation, MAE rises on

Table 3: **Task-level and probability-sensitive metrics for the single-layer language settings.** Entries are same-data control / paired; NRE and decision-token log-probability-effect error are multiplied by 10^3, task MAE is unscaled, and IIA is a percentage.

Setting	Task MAE	IIA (%)	NRE	Log-prob. error
IOI b8, 25%, d0	0.0890 / 0.0718	98.4 / 98.0	9.29 / 8.30	2.43 / 1.98
IOI b8, 25%, d1	0.0476 / 0.0454	96.1 / 97.7	9.66 / 7.94	1.58 / 1.32
IOI b5, 25%, d1	0.147 / 0.120	98.0 / 98.8	6.38 / 5.13	10.8 / 7.47
IOI b8, 50%, d0	0.0961 / 0.0713	98.0 / 96.9	24.8 / 20.7	4.84 / 3.95
Pronoun b8, d0	0.0292 / 0.0314	96.9 / 97.7	2.54 / 2.25	0.633 / 0.888
Pronoun b8, d1	0.0286 / 0.0258	98.0 / 96.9	2.20 / 2.03	0.418 / 0.391

draw 0 and falls on draw 1. NRE favors pairing in all six cells, and decision-token log-probability error favors it in five, the exception being pronoun draw 0.

5.6 Stability and ordinary fidelity

In all five sketch, calibration, and refinement perturbations, pairing lowers effect MSE, with four intervals excluding zero. The frozen selector's natural-KL change ranges from -41.8% to $+127.6\%$, while removed-set Jaccard overlap is about 0.50–0.54 across sketch seeds and calibration subsets (Appendix Table 8). Single-run KL differences of the frozen selector are therefore unstable, whereas the trained student lowers KL with pairing on both draws. The selector thus finds one of several functionally similar reduced networks, so the membership of individual channels in the removed set should not be given a mechanistic reading.

6 Discussion and limitations

What the evidence supports. Weighting the intervention-induced change, in addition to the two outputs, lowers response error at matched data and width in all six single-layer GPT-2 language settings and in the fixed-setting transfer to Pythia-410m. The same direction appears when the retained weights are also trained (post hoc) and in three of four whole-network cells; the 25% whole-network IOI cell is a tie with worse task-facing errors. The absolute errors are small at a single layer: NRE is below 1% for both methods in every 25%-removal GPT-2 cell and 2.1–2.5% at 50% removal. Under 25% whole-network pruning, paired NRE is 11.4% on IOI and 4.1% on pronoun continuation. The paired term is simple to add, and its ingredients have precedents in causal distillation, Jacobian matching, and compensated pruning; our contribution is the objective, a post-training procedure for fitting it, and a controlled study of when it helps.

Limitations. The objective does not constrain ordinary-input fidelity; it trades it against response fidelity through the single weight λ. Adding an explicit KL constraint or a probability-sensitive term is a natural next step. We study one intervention family per task, a dense-model donor state patched at a single residual port, and we do not claim that the pruned network implements the same internal mechanism as the dense one. Ordinary fidelity is measured on each task's prompt families rather than on general text. Each model family is represented by one checkpoint, and on vision the gain is not distinguishable from zero. Finally, we compare against a trained student only at small scale, not against a fully trained causal-distillation model.

7 Conclusion

Paired-response reduction selects bounded constant replacements for FFN channels of a fixed network to reduce errors in both its ordinary outputs and its final-output responses to a declared residual-patch intervention, then compiles the replacements exactly into smaller layers. Adding the paired term to an otherwise identical objective lowers intervention-response error in every single-layer GPT-2 language setting we test and in most whole-network settings, while ordinary and task-facing fidelity vary by setting. When a pruned network will be studied under a known intervention, including that intervention in the pruning objective is a simple change that improves fidelity to the measured response.

AI USE STATEMENT

We used generative AI assistants to help design, implement, and run the experiments, to draft parts of the text, especially the experimental sections, and to edit the whole manuscript. We did not use generative AI to fabricate data or results; every reported number is computed by our code from saved experiment outputs. We directed the research, reviewed all AI-assisted work, and take responsibility for the final content.

ETHICS STATEMENT

The pronoun-continuation task probes occupation-pronoun associations of a pretrained language model. Our objective preserves the reference model's responses, including stereotyped ones; it neither measures nor reduces bias, and a reduced model inherits the reference model's biases. No human subjects or new personal data are involved.

REPRODUCIBILITY STATEMENT

Section 4 and Appendices C and B specify models, data splits, frozen protocols, and the full procedure. Every configuration uses fixed seeds, saved input manifests with content hashes, and saved selections, constants, and compiled layers. We will release the code and these artifacts.

REFERENCES

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 210–227, 2026a.

Amir Asiaee. Causal mechanism reduction: Mechanism replacement for neural network pruning and abstraction, 2026b. arXiv preprint, v2.

Uranik Berisha, Jens Mehnert, and Alexandru Paul Condurache. Variance-based pruning for accelerating and compressing trained networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0587.

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

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337. PMLR, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 2022.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah D. Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5 (NeurIPS)*, pp. 164–171, 1992.

Woojeong Kim, Suhyun Kim, Mincheol Park, and Geonseok Jeon. Neuron merging: Compensating for pruned neurons. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems 2 (NeurIPS)*, pp. 598–605, 1989.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. *Advances in Neural Information Processing Systems*, 2023.

Kiran Nair, Smriti Regmi, and Rodrigue Rizk. CausalGate: Causal importance distillation for transformer module pruning, 2026. Preprint, version 2.

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

Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient second-order approximation for neural network compression. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Suraj Srinivas and R. Venkatesh Babu. Data-free parameter pruning for deep neural networks. In *British Machine Vision Conference*, 2015.

Suraj Srinivas and François Fleuret. Knowledge transfer with Jacobian matching. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4723–4731, 2018.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021.

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*, volume 33, 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*, 2023.

Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah Goodman. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4288–4295, 2022. doi: 10.18653/v1/2022.naacl-main.318.

Lei Yu, Jingcheng Niu, Zining Zhu, Xi Chen, and Gerald Penn. Sheaf discovery with joint computation graph pruning and flexible granularity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8822–8837, 2025. doi: 10.18653/v1/2025.emnlp-main.446.

Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I. Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S. Davis. NISP: Pruning networks using neuron importance score propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. doi: 10.18653/v1/N18-2003.

A Derivations and scope of exactness

A.1 Constant and affine bias folding

Partitioning the FFN output in Equation (1) gives

$$u = W_{2,:,K} a_K + W_{2,:,S} a_S + b_2.$$

Under $a_S \leftarrow c$, substitution gives

$$u^{S \leftarrow c} = W_{2,:,K} \sigma(W_{1,K,:} h + b_{1,K}) + \underbrace{b_2 + W_{2,:,S} c}_{b_2'}.$$

This is exactly the FFN with Equation (2)'s parameters. Equality is pointwise in the incoming residual h, so it also holds when h is produced by an upstream intervention. Adding the same residual branch and applying any identical deterministic suffix preserves equality. The identity relates the compiled network to the clamped one; it says nothing about how either compares with the dense network.

For an affine replacement, substitute $a_S = \beta + B a_K$:

$$u^{S \leftarrow \beta + B a_K} = (W_{2,:,K} + W_{2,:,S} B) a_K + b_2 + W_{2,:,S} \beta.$$

Neither fold depends on the choice of activation function.

For fixed K, S, write $\rho = \operatorname{rank}(W_{2,:,S})$. Constant folding changes the bias within a subspace of dimension ρ; affine folding changes the outgoing weights and bias within a parameter family of dimension $\rho(k+1)$. Both are contained in free outgoing compensation, which has dimension $d(k+1)$, with equality for the affine family if $\rho = d$.

Removing $m = p - k$ channels removes md incoming coefficients, m incoming biases, and dm outgoing coefficients. The output bias remains, giving $m(2d+1)$ fewer parameters for these FFNs. Affine folding has the same compiled shapes.

A.2 Centering and the paired-error identity

The projection satisfies $P^2 = P = P^\top$ and $P\mathbf{1} = 0$. Since $\operatorname{softmax}(z + \alpha \mathbf{1}) = \operatorname{softmax}(z)$, centering discards only the common-offset degree of freedom. Matching centered logits to zero error implies equal probabilities; a small centered Euclidean error and a small KL are different quantitative criteria.

Let $s = e^0 + e^I$ and $d_e = e^I - e^0$. Then $e^0 = (s - d_e)/2$ and $e^I = (s + d_e)/2$. Expanding squared norms yields

$$\tfrac{1}{2} \left\| e^0 \right\|_2^2 + \tfrac{1}{2} \left\| e^I \right\|_2^2 + \lambda \left\| d_e \right\|_2^2 = \tfrac{1}{4} \left\| s \right\|_2^2 + (\tfrac{1}{4} + \lambda) \left\| d_e \right\|_2^2,$$

which proves the displayed paired-error identity in Section 3.2. Also,

$$\left\| e^I - e^0 \right\|_2^2 \le 2 \left\| e^I \right\|_2^2 + 2 \left\| e^0 \right\|_2^2.$$

Thus sufficiently accurate same-data output fitting already controls effect error. Pairing changes the allocation of error under restricted fitting capacity. For a centered vector v, the error pairs (v, v) and $(v, -v)$ have identical separate squared errors but effect errors zero and $4 \left\| v \right\|_2^2$, respectively.

A.3 Conditional optimization

Fix residual r from the already selected channels. For candidate j,

$$f_j(c) = \left\| r + c U_j - V_j \right\|_2^2, \qquad f_j'(c) = 2 U_j^\top (r + c U_j - V_j).$$

For $\left\| U_j \right\|_2^2 > 0$, the unconstrained minimum is $U_j^\top (V_j - r) / \left\| U_j \right\|_2^2$. Clipping this scalar to the closed observed interval gives the bounded minimum. If $U_j = 0$, f_j is independent of c, although $\left\| r - V_j \right\|_2^2$ can differ from $\left\| r \right\|_2^2$. The implemented update uses the calibration mean when $\left\| U_j \right\|_2^2 \le \tau$; when the denominator is positive but below the threshold, this fallback need not minimize f_j. Apart from that

branch, each coordinate update weakly decreases the quadratic, although two sweeps need not reach its joint minimum.

The expansion in Equation (5) is exact for an affine output map on the relevant activation perturbations. For the squared fidelity discrepancy, the zero dense residual eliminates terms multiplying second derivatives of the outputs at the expansion point; the local Hessian has the usual Jacobian Gram form. For a large replacement the nonlinear suffix can still depart from this local approximation, which is why every selected model is evaluated exactly.

For independent signs, $\mathbb{E}[\xi\xi^\top] = I_q$; hence

$$\mathbb{E}[(\xi^\top Pe)^2] = e^\top P\mathbb{E}[\xi\xi^\top]Pe = e^\top Pe.$$

The same ξ on both executions is needed to estimate the cross term in $\left\|e^I - e^0\right\|_2^2$. Independent probes for the two executions would lose that cross term in expectation.

A.4 Sketch variance and rank ceilings

For M independent Rademacher probes and fixed $x = Pe$, the averaged estimator has variance at most $(2/M)(\|x\|_2^2/q)^2$, which shrinks as the number of probes grows. We use $M = 16$ probes and $n = 128$ pairs. The same-data design has $2nM = 4{,}096$ rows and therefore row rank at most $\min(p, 2nM)$. The paired design has $3nM = 6{,}144$ rows, but its third block is a rowwise linear combination of the first two, so its rank is also at most $2nM$. With four probes the ceiling would be 1,024, below the FFN width, which is why we use 16.

B Full single-block procedure and pseudocode

This appendix gives the full single-block procedure. Whole-model reduction (Section 3.5) repeats it block by block with four changes: derivatives are evaluated around the current compiled chain, the original dense model remains the target, the quadratic is initialized with the chain's current reduced-minus-dense natural and intervened residuals instead of zero, and each block is compiled before the next is fitted. The single-block protocol uses centered outputs, 16 probes, diagnostic selection from $\{1, 4\}$, and two coordinate sweeps. The reference is in evaluation mode throughout; retained weights are frozen.

B.1 Common inputs and derivative collection

```
freeze(D.parameters); D.eval()
freeze(calibration_pairs, diagnostic_pairs, test_pairs, intervention_rule)
for each pair (x, donor) in each split:
    fixed_donor_state[pair] = D.capture(donor, common_port).detach()
    cache dense natural and patched logits using this donor state

Mprobes = 16
for calibration microbatch:
    for b in [0, I]:
        patch = None if b == 0 else fixed_donor_state[microbatch]
        # Prefix frozen; leaf activation feeds the actual nonlinear suffix.
        z[b], a[b] = forward_with_activation_leaf(D, microbatch, patch)
    update natural-only and pooled channel means, variances, min, max
    # All valid tokens contribute; padding does not.
    for s in 1..Mprobes:
        xi = reproducible_Rademacher_vectors(pair_ids, s, q)
        v = xi - mean(xi, output_axis, keepdim=True)
        for b in [0, I]:
            scalar = sum(v * z[b], output_axis) / sqrt(q)
            g = grad(sum(scalar), a[b], retain_graph=(s < Mprobes))
            uhat[b] = sum(g, valid_token_axis)
            vhat[b] = sum(g * a[b], valid_token_axis)
        save_atomic(pair_ids, s, uhat, vhat)
    free_graphs()
```

```
# Every end-to-end variant reuses these derivatives and paired probes.
if objective == natural:
    U, V = flatten(uhat[0]), flatten(vhat[0])
    bounds, means = natural_statistics
else:
    U = stack(uhat[0]/sqrt(2), uhat[I]/sqrt(2))
    V = stack(vhat[0]/sqrt(2), vhat[I]/sqrt(2))
    if lam > 0:
        append sqrt(lam)*(uhat[I]-uhat[0]) to U
        append sqrt(lam)*(vhat[I]-vhat[0]) to V
    bounds, means = pooled_natural_and_patched_statistics
U, V = U/sqrt(n*Mprobes), V/sqrt(n*Mprobes)
cache U.T@U, U.T@V, V.T@V in float64
```

B.2 Greedy selection, refinement, and compilation

For the saved implementation, the negligible-sensitivity threshold is $\tau = \max(10^{-12} \max_j \|U_j\|_2^2, 10^{-30})$. Means and bounds use the objective's indicated token pool.

```
def conditional_constant(j, residual):
    if norm2(U[j]) <= tau:
        return clip(mean[j], lo[j], hi[j])
    return clip(dot(U[j], V[j]-residual)/norm2(U[j]), lo[j], hi[j])

S = empty_set(); residual = zeros(number_of_sketch_rows)
constants = means.copy()
sweep = 0
for cursor in 1..p-k:
    for j not in S:
        proposed[j] = conditional_constant(j, residual)
        delta[j] = norm2(residual + proposed[j]*U[j] - V[j])
                   - norm2(residual)
    j = deterministic_argmin(delta, tie_break=channel_index)
    constants[j] = proposed[j]
    residual += constants[j]*U[j] - V[j]
    S.add(j)
    checkpoint_atomic(S, constants, U.T@residual, V.T@residual,
                      sweep, cursor, config_hash)

for sweep in 1..2:
    for cursor, j in enumerate(sorted(S)):
        without_j = residual - (constants[j]*U[j] - V[j])
        constants[j] = conditional_constant(j, without_j)
        residual = without_j + constants[j]*U[j] - V[j]
        save refinement state periodically

K = complement(S)
W1_new, b1_new = W1[K, :], b1[K]
W2_new = W2[:, K]
b2_new = b2 + W2[:, S] @ constants[S]
R = copy_with_physically_smaller_FFN(D, W1_new, b1_new, W2_new, b2_new)
assert_same_outputs(R, explicit_clamp(D, S, constants),
                    natural_and_patched_diagnostic_inputs,
                    atol=1e-4, rtol=1e-5)
freeze(S, K, constants, R.FFN_parameters)
```

B.3 Exact loss, selection of lambda, and held-out evaluation

```
def exact_centered_loss(D, R, pairs, lam):
    total = 0
    for pair in pairs:
```

```
        zD0, zDI = cached_dense_outputs(pair)
        zR0 = R(pair.x)
        zRI = R(pair.x, patch=pair.fixed_dense_donor_state)
        e0 = center(zR0-zD0); eI = center(zRI-zDI)
        total += (0.5*norm2(e0) + 0.5*norm2(eI)
                  + lam*norm2(eI-e0))/q
    return total/len(pairs)

for model in [DeiT, GPT2]:
    for keep in [0.75, 0.50]:
        fit VBP, CMR-local, E2E-natural, E2E-augmented
        fit E2E-paired(lam=1), E2E-paired(lam=4)
        fit augmented local reconstruction on E2E-augmented mask
        check approximation on diagnostic_pairs[0:8]
        compare candidates on diagnostic_pairs[8:64]
        KL_limit = 1.1 * min(KL_augmented, KL_reconstruction)
        eligible = paired candidates with natural_KL <= KL_limit
        diagnostic_feasible = eligible is not empty
        if diagnostic_feasible: choose lowest centered_effect_MSE among eligible
        else: choose lam=1
        save lam decision before test inference
        evaluate six named methods once on all 256 test pairs
# GPT-2 only: transfer keep=0.75 and lam=4 without retuning
refit all methods on a fresh draw1 at blocks 8 and 5
evaluate both blocks on their shared draw1 test pairs
report every configuration, component errors, conditional paired intervals,
       approximation error, and model sizes
```

C PROTOCOL AND DATA SPLITS

Data and preprocessing. Vision calibration uses the ImageNet-100 training split, while diagnostic and test base/donor IDs use disjoint validation pools. The model's saved 100-class ordering is retained. IOI prompts use the canonical role-swapped construction with 24 calibration, 16 diagnostic, and 24 test names and separate template pools. Calibration, diagnostic, and test template indices are $\{0, 1, 2, 5\}$, $\{6, 8\}$, and $\{3, 4, 7, 12\}$, respectively. Draw 1 excludes draw 0's texts, and its two GPT-2 sites share evaluation pairs. All valid token positions contribute to calibration statistics, output Jacobians, and reconstruction; next-token evaluation uses the final valid position. Each donor residual is computed by the dense model once and reused across methods. The pronoun task is described in Appendix J.

Single-block settings. We fix 16 probes, two refinement sweeps, and candidate $\lambda \in \{1, 4\}$. The first eight diagnostic pairs assess approximation. On the remaining 56, the candidate with lowest effect error is selected among those whose natural KL is at most 1.1 times that of each strong control. If neither is feasible, the declared fallback is $\lambda = 1$, flagged as infeasible. IOI selects $\lambda = 4$; neither vision candidate is feasible at either width, so vision uses the fallback.

For GPT-2, the 25% setting with $\lambda = 4$ is then refit on a second calibration draw and evaluated at block 8 and at block 5, without retuning. Together with 50% removal at block 8, this gives the four IOI configurations reported in Section 5.1.

Whole-model extension. For each task, method, and budget, blocks 0 through 11 are reduced in order. Every block targets original dense outputs and is linearized around the current compiled chain, with the current stacked reduced-minus-dense error as the initial residual. The extension fixes $\lambda = 4$ for paired reduction, $\lambda = 0$ for the same-data control, 16 projections, two refinement sweeps, pooled activation bounds, and 25% and 50% per-FFN removal. No whole-model outcome, hyperparameter, budget, block order, or projection count is selected from the held-out results. It reuses the draw-0 calibration and held-out pairs of the single-block study, so it is an untuned extension rather than a fresh replication.

Augmented local reconstruction. For retained activations A_K and dense local FFN outputs Y, pool valid tokens from natural and intervened calibration executions. Fit

$$\min_{C,b} \; N^{-1}\|A_K C + \mathbf{1}b^\top - Y\|_F^2 + \eta\|C\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Center both design and targets, solve for C, and recover the unpenalized intercept as $b = \bar{Y} - C^\top \bar{A}_K$. The ridge value is a multiplier from $\{10^{-4}, 10^{-2}, 1\}$ times the mean diagonal of the centered design covariance, floored at 10^{-15}. A fixed seed, 2,828, chooses 26 of the 128 pairs for validation. Both executions and all tokens of a pair remain together. Validation local-output MSE selects the multiplier, then all 128 pairs are used for the final fit. This baseline updates $d(k+1)$ coefficients while keeping incoming retained weights fixed. The table reports that fitting dimension; it does not add parameters to the compiled network.

Table 4: **Only a small fraction of the whole network is removed.** Dense sizes are 124,439,808 GPT-2 and 5,543,716 DeiT parameters. Bounds and fallback counts refer to the selected paired fit after refinement; fallback counts are over all candidate channels.

Setting	Removed channels	Parameters removed	Model (%) removed	Recon. fit coefficients	Bound hits	Fallbacks
GPT-2, b8, 25%, d0	768	1,180,416	0.9486	1,770,240	1	0
GPT-2, b8, 25%, d1	768	1,180,416	0.9486	1,770,240	0	0
GPT-2, b5, 25%, d1	768	1,180,416	0.9486	1,770,240	0	0
GPT-2, b8, 50%, d0	1,536	2,360,832	1.8972	1,180,416	10	0
DeiT, b8, 25%, d0	192	73,920	1.3334	110,784	17	0
DeiT, b8, 50%, d0	384	147,840	2.6668	73,920	50	0

Numerical execution. Models are evaluated in FP32 with TF32 disabled, and the selector's Gram matrices use float64. The single-block study ran on one RTX 4500 Ada GPU with 10 CPU cores, using 486 seconds of worker time and at most 0.60 GB of allocated GPU memory.

Reproducibility scope. For every study we save the pair manifests, checkpoint identities, output sketches, masks, constants, compiled weights, per-example outputs, λ decisions, and bootstrap intervals, together with a protocol file fixed before evaluation. These artifacts will be released with the code.

D Raw single-block primary values

Table 5: **Pairing improves IOI response fidelity with site-dependent ordinary-fidelity costs.** A is E2E-augmented and P is E2E-paired. Effect MSE and KL values, including absolute KL differences, are multiplied by 10^3; relative columns are percentages. Positive effect reduction favors P; positive KL change means worse ordinary fidelity. Blocks are zero-based, d0/d1 denote draws, and percentages in settings denote FFN channel removal. Each row has 256 test pairs.

	Centered effect MSE			Natural KL (nats)			
Setting	A	P	Reduction (%)	A	P	Δ	Δ (%)
GPT-2, b8, 25%, d0	0.58055	0.51880	10.64	4.82778	3.72172	-1.10606	-22.91
GPT-2, b8, 25%, d1	0.60618	0.49823	17.81	5.52300	5.17996	-0.34304	-6.21
GPT-2, b5, 25%, d1	2.00876	1.61419	19.64	1.84180	2.47948	+0.63768	+34.62
GPT-2, b8, 50%, d0	1.54834	1.29599	16.30	6.20111	8.14625	+1.94515	+31.37
DeiT, b8, 25%, d0	0.42126	0.40578	3.67	0.77234	0.72871	-0.04362	-5.65
DeiT, b8, 50%, d0	4.55398	4.37887	3.85	10.40986	10.92721	+0.51735	+4.97

E Whole-model and post hoc sensitivity results

E.1 Full whole-model comparison

E.2 Matched-data student intervals

Table 6: **Whole-model reduction across both tasks and budgets.** Effect MSE and natural KL are multiplied by 10^3; contrast MAE is unscaled. Each budget removes the stated fraction from every GPT-2 FFN.

Task	Removed	Method	Effect MSE	Natural KL	Contrast MAE
IOI	25%	E2E-paired	7.13	38.7	0.345
		E2E-augmented	7.05	56.5	0.232
		Augmented-reconstruction	27.8	1,090	0.475
		VBP	10.2	66.0	0.341
	50%	E2E-paired	24.6	949	0.536
		E2E-augmented	27.9	675	0.552
		Augmented-reconstruction	42.4	3,030	1.05
		VBP	29.4	1,610	0.606
Pronoun	25%	E2E-paired	17.6	34.2	0.110
		E2E-augmented	21.6	38.1	0.128
		Augmented-reconstruction	173	279	0.545
		VBP	39.5	71.7	0.185
	50%	E2E-paired	62.1	122	0.260
		E2E-augmented	66.1	124	0.272
		Augmented-reconstruction	180	296	0.564
		VBP	119	151	0.308

Table 7: **Conditional paired intervals for the matched-data trained-student analysis.** T and F denote trained and frozen regimes; subscripts denote λ. Effect-MSE and KL differences are multiplied by 10^3, task-MAE differences are unscaled, and negative values favor the first term in each comparison.

Draw	Comparison	Metric	Absolute difference [95% interval]	Relative (%) [95% interval]
0	$T_4 - T_0$	Effect MSE	-0.0827 $[-0.0968, -0.0695]$	-15.9 $[-18.1, -13.6]$
0	$T_4 - T_0$	Natural KL	-1.96 $[-2.27, -1.69]$	-46.4 $[-50.9, -41.6]$
0	$T_4 - T_0$	Task MAE	-0.0204 $[-0.0248, -0.0154]$	-27.3 $[-32.0, -21.8]$
0	$T_0 - F_0$	Effect MSE	-0.0591 $[-0.0659, -0.0524]$	-10.2 $[-11.3, -9.1]$
0	$T_0 - F_0$	Natural KL	-0.594 $[-0.723, -0.477]$	-12.3 $[-14.7, -10.0]$
0	$T_0 - F_0$	Task MAE	-0.0143 $[-0.0177, -0.0110]$	-16.1 $[-19.8, -12.6]$
0	$T_4 - F_4$	Effect MSE	-0.0800 $[-0.0872, -0.0728]$	-15.4 $[-16.6, -14.2]$
0	$T_4 - F_4$	Natural KL	-1.45 $[-1.60, -1.32]$	-39.0 $[-41.2, -37.0]$
0	$T_4 - F_4$	Task MAE	-0.0175 $[-0.0205, -0.0147]$	-24.4 $[-27.9, -21.1]$
1	$T_4 - T_0$	Effect MSE	-0.116 $[-0.131, -0.100]$	-20.9 $[-22.9, -18.7]$
1	$T_4 - T_0$	Natural KL	-0.614 $[-0.904, -0.329]$	-12.9 $[-18.3, -7.1]$
1	$T_4 - T_0$	Task MAE	-0.00476 $[-0.00900, -0.000502]$	-9.9 $[-17.9, -1.1]$
1	$T_0 - F_0$	Effect MSE	-0.0507 $[-0.0578, -0.0430]$	-8.4 $[-9.5, -7.1]$
1	$T_0 - F_0$	Natural KL	-0.767 $[-0.930, -0.605]$	-13.9 $[-16.1, -11.4]$
1	$T_0 - F_0$	Task MAE	$+0.000654$ $[-0.00229, +0.00360]$	$+1.4$ $[-4.8, +7.7]$
1	$T_4 - F_4$	Effect MSE	-0.0590 $[-0.0650, -0.0532]$	-11.8 $[-12.9, -10.8]$
1	$T_4 - F_4$	Natural KL	-1.04 $[-1.19, -0.879]$	-20.0 $[-22.2, -17.7]$
1	$T_4 - F_4$	Task MAE	-0.00186 $[-0.00431, +0.000564]$	-4.1 $[-9.1, +1.3]$

E.3 Post hoc lambda frontier

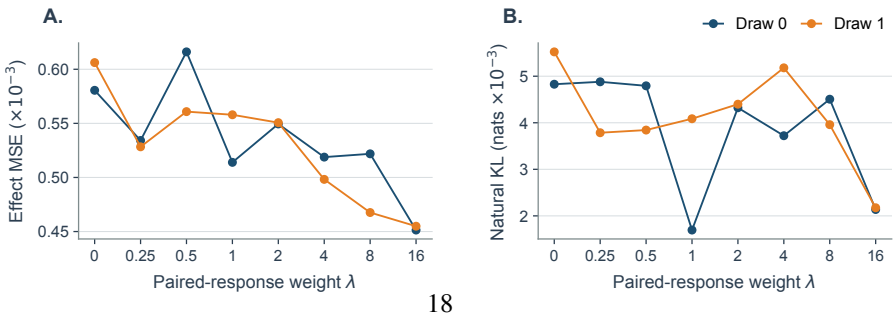


Figure 3: **The primary-setting held-out frontier is smooth overall but is not monotone. A.** Centered effect MSE across paired-response weights. **B.** Natural KL for the same frozen selections; lines connect weights within

Table 8: **Effect gains persist across all five primary-setting perturbations, while KL and selected sets vary.** Effect MSE and natural KL are multiplied by 10^3; Jaccard overlap compares each removed set with the original frozen selection at the same λ.

Variant	λ	Effect MSE	Natural KL	Task MAE	Jaccard
Calibration first 32	0	0.610	4.43	0.0567	0.512
	4	0.482	2.58	0.0471	0.516
Calibration first 64	0	0.568	2.68	0.0524	0.530
	4	0.548	4.98	0.0725	0.528
Sketch seed 1	0	0.616	3.73	0.0575	0.504
	4	0.556	2.63	0.0523	0.536
Sketch seed 2	0	0.488	2.46	0.0425	0.507
	4	0.464	5.59	0.0573	0.531
Zero refinement	0	0.576	4.84	0.0879	1.000
	4	0.520	3.57	0.0742	1.000

Table 9: **Paired-minus-control intervals for the five stability perturbations.** Absolute effect-MSE and KL differences are multiplied by 10^3; each cell also gives the signed relative change and its conditional 95% interval.

Variant	Effect MSE: difference [interval]; relative [interval]	Natural KL: difference [interval]; relative [interval]
Calibration first 32	-0.128 $[-0.142, -0.114]$; -21.0% $[-23.0, -19.0]$	-1.85 $[-2.29, -1.49]$; -41.8% $[-47.9, -35.5]$
Calibration first 64	-0.0200 $[-0.0397, +0.000523]$; -3.5% $[-6.8, +0.1]$	$+2.30$ $[+2.02, +2.56]$; $+85.7\%$ $[+74.8, +97.3]$
Sketch seed 1	-0.0599 $[-0.0764, -0.0436]$; -9.7% $[-12.2, -7.1]$	-1.10 $[-1.34, -0.881]$; -29.4% $[-33.6, -24.9]$
Sketch seed 2	-0.0246 $[-0.0383, -0.0115]$; -5.0% $[-7.8, -2.4]$	$+3.14$ $[+2.74, +3.51]$; $+127.6\%$ $[+109.7, +147.6]$
Zero refinement	-0.0566 $[-0.0706, -0.0426]$; -9.8% $[-12.1, -7.5]$	-1.27 $[-1.60, -0.967]$; -26.2% $[-31.6, -20.6]$

F PYTHIA-410M RESULTS

Pythia-410m uses snapshot `9879c9b5`, block 16 of 24, 25% FFN removal, $\lambda = 4$ transferred from GPT-2, 16 probes, and two refinement sweeps. IOI and pronoun pairs were regenerated with single-token names, occupations, and pronouns for its tokenizer, with the same split structure and counts as the GPT-2 study (128/64/256). The paired-versus-control effect intervals are $[-15.5\%, -10.8\%]$ for IOI and $[-12.6\%, -8.5\%]$ for pronoun continuation; the natural-KL intervals are $[-7.0\%, +5.2\%]$ and $[+6.6\%, +11.1\%]$. On pronoun continuation, E2E-natural has lower natural KL than E2E-paired and nearly equal effect MSE. Finite-difference relative RMS is 0.134–0.144 on IOI and 0.060–0.078 on pronoun continuation; linearization relative RMS error is 0.130–0.144 and 0.071–0.076. The protocol required finite-difference stability below 10% before held-out evaluation. Pronoun continuation passed. IOI exceeded the threshold, so we report the Pythia IOI cell as exploratory; its held-out numbers are still exact evaluations of the selected model.

Table 10: **All Pythia-410m methods at block 16 and 25% removal.** Effect MSE and natural KL are multiplied by 10^3; task MAE is in unscaled logit units; IIA is a percentage. Each row uses 256 held-out pairs.

Task	Method	Effect MSE	Natural KL	Task MAE	IIA (%)
IOI	VBP	0.293	1.752	0.0326	98.0
	CMR-local	0.254	1.837	0.0364	96.5
	E2E-natural	0.192	1.538	0.0285	96.1
	E2E-augmented	0.193	1.241	0.0333	96.9
	E2E-paired	0.167	1.231	0.0257	97.7
	Aug.-reconstruction	0.786	6.583	0.0652	96.1
Pronoun	VBP	0.594	2.177	0.0168	99.6
	CMR-local	0.477	1.869	0.0173	99.2
	E2E-natural	0.314	0.687	0.0161	98.8
	E2E-augmented	0.344	0.878	0.0209	100.0
	E2E-paired	0.308	0.955	0.0161	98.8
	Aug.-reconstruction	3.021	8.722	0.0778	98.4

G FULL SINGLE-BLOCK RESULTS

Tables 11–14 report every method in every single-block configuration. E2E-paired uses $\lambda = 4$ in all IOI cells and $\lambda = 1$ in vision. Every row averages 256 pairs. Accuracies and agreements below are fractions. Vision ECE uses 15 equal-width confidence bins. Dashes indicate metrics not defined for that task.

Table 11: **Centered fidelity and probability fidelity for all six methods.** All numeric entries are multiplied by 10^3; KL is in nats before scaling.

Method	Natural MSE	Patched MSE	Effect MSE	Natural KL	Patched KL
GPT-2, b8, 25%, d0					
VBP	11.443546	11.609194	0.803200	3.941705	4.593239
CMR-local	10.327119	10.461076	0.703324	5.599611	7.537119
E2E-natural	9.822801	9.933340	0.557659	4.984852	7.112743
E2E-augmented	9.340096	9.573154	0.580553	4.827777	7.893405
E2E-paired	8.678939	8.918080	0.518805	3.721722	5.696408
Augmented-reconstruction	31.766031	31.532521	2.116524	8.772803	13.193447
GPT-2, b8, 25%, d1					
VBP	11.626985	11.816226	0.808220	3.367110	4.065080
CMR-local	10.850966	10.982728	0.701477	4.067409	5.371140
E2E-natural	9.697598	9.925108	0.673542	3.126672	4.356384
E2E-augmented	12.113426	12.437153	0.606178	5.522996	6.878784
E2E-paired	9.262803	9.550223	0.498231	5.179960	6.683433
Augmented-reconstruction	27.555577	28.006610	1.794856	6.003803	7.214967
GPT-2, b5, 25%, d1					
VBP	14.787949	15.222887	3.676206	3.888387	6.063530
CMR-local	9.108547	9.176770	2.391743	2.739382	4.685702
E2E-natural	10.104280	11.756328	3.945561	2.208697	10.376756
E2E-augmented	8.969632	9.812766	2.008762	1.841802	4.355747
E2E-paired	11.550856	11.795518	1.614191	2.479484	3.954494
Augmented-reconstruction	45.166684	41.397691	9.595811	29.779983	21.185897
GPT-2, b8, 50%, d0					
VBP	34.498326	34.591363	2.370687	10.037512	13.886455
CMR-local	35.001046	35.094816	2.395935	14.861821	21.063200
E2E-natural	26.473733	26.672321	1.641522	7.558612	10.323572
E2E-augmented	21.740179	21.801057	1.548335	6.201106	9.483895
E2E-paired	22.799441	22.829347	1.295987	8.146253	11.467427
Augmented-reconstruction	76.201281	76.239175	4.546556	18.423915	24.460667

Table 11 continued

Method	Natural MSE	Patched MSE	Effect MSE	Natural KL	Patched KL
DeiT, b8, 25%, d0					
VBP	11.379037	12.926151	2.731496	4.153149	7.254211
CMR-local	3.928840	5.059899	1.609014	1.562810	4.839664
E2E-natural	2.548867	2.747030	0.424849	0.792469	0.719798
E2E-augmented	2.483326	2.679623	0.421258	0.772335	0.683939
E2E-paired	2.425292	2.614147	0.405783	0.728712	0.686472
Augmented-reconstruction	2.513222	2.708884	0.417712	0.763100	0.689799
DeiT, b8, 50%, d0					
VBP	57.220689	59.157568	7.857006	19.251181	21.879237
CMR-local	40.623720	42.387989	5.941366	16.796848	19.359002
E2E-natural	29.871064	31.588971	4.768510	11.301303	14.121023
E2E-augmented	27.401205	28.581781	4.553980	10.409861	12.742226
E2E-paired	26.041848	27.611349	4.378868	10.927214	13.434225
Augmented-reconstruction	26.202845	28.143779	4.390361	8.380126	11.510373

Table 12: **Task performance, agreement, and calibration for all single-block cells.** Agreement is natural argmax agreement with the dense model. Language accuracy compares the two IOI candidates; IIA is full-vocabulary agreement. NLL is natural true-target negative log probability; ECE is reported for vision.

Method	Accuracy	Agreement	IIA	NLL	Task MAE	ECE
GPT-2, b8, 25%, d0						
VBP	0.996094	0.976562	0.980469	0.780219	0.053599	–
CMR-local	0.996094	0.984375	0.972656	0.797983	0.059934	–
E2E-natural	1.000000	0.992188	0.984375	0.761711	0.061922	–
E2E-augmented	1.000000	0.996094	0.984375	0.760823	0.089028	–
E2E-paired	1.000000	0.992188	0.980469	0.758406	0.071789	–
Augmented-reconstruction	1.000000	0.984375	0.972656	0.741927	0.143668	–
GPT-2, b8, 25%, d1						
VBP	0.996094	0.980469	0.957031	0.784966	0.055303	–
CMR-local	0.996094	0.980469	0.960938	0.791602	0.062920	–
E2E-natural	0.996094	0.988281	0.976562	0.763742	0.056343	–
E2E-augmented	0.996094	0.992188	0.960938	0.782888	0.047619	–
E2E-paired	0.996094	0.992188	0.976562	0.772590	0.045380	–
Augmented-reconstruction	0.996094	0.984375	0.960938	0.741212	0.077362	–
GPT-2, b5, 25%, d1						
VBP	0.996094	0.984375	0.976562	0.711425	0.177185	–
CMR-local	0.996094	0.984375	0.984375	0.742643	0.177165	–
E2E-natural	0.996094	0.984375	0.972656	0.719209	0.385444	–
E2E-augmented	0.996094	0.992188	0.980469	0.727905	0.146617	–
E2E-paired	0.996094	0.984375	0.988281	0.736426	0.119714	–
Augmented-reconstruction	0.996094	0.980469	0.964844	0.839890	0.548119	–
GPT-2, b8, 50%, d0						
VBP	1.000000	0.980469	0.953125	0.767021	0.138448	–
CMR-local	1.000000	0.976562	0.968750	0.817323	0.140710	–
E2E-natural	1.000000	0.988281	0.976562	0.759341	0.095040	–
E2E-augmented	1.000000	0.984375	0.980469	0.760224	0.096063	–
E2E-paired	1.000000	0.984375	0.968750	0.767289	0.071261	–
Augmented-reconstruction	1.000000	0.984375	0.953125	0.756164	0.159579	–
DeiT, b8, 25%, d0						
VBP	0.875000	0.988281	0.988281	0.401317	–	0.038448
CMR-local	0.882812	0.996094	0.996094	0.404691	–	0.034734
E2E-natural	0.878906	0.988281	0.992188	0.400736	–	0.027852
E2E-augmented	0.878906	0.988281	0.992188	0.401112	–	0.025055
E2E-paired	0.878906	0.996094	0.992188	0.403940	–	0.023907
Augmented-reconstruction	0.882812	0.992188	0.984375	0.401319	–	0.025632

Table 12 continued

Method	Accuracy	Agreement	IIA	NLL	Task MAE	ECE
DeiT, b8, 50%, d0						
VBP	0.882812	0.949219	0.953125	0.436007	–	0.044729
CMR-local	0.871094	0.949219	0.953125	0.422261	–	0.032161
E2E-natural	0.882812	0.968750	0.960938	0.406752	–	0.033487
E2E-augmented	0.890625	0.976562	0.960938	0.404532	–	0.026591
E2E-paired	0.878906	0.964844	0.972656	0.402444	–	0.047392
Augmented-reconstruction	0.890625	0.976562	0.964844	0.405037	–	0.048926

Table 13: **Dense-reference behavior for each configuration.** Dense effect MSE is the squared centered intervention change divided by q, not a reduction error. Dense task effect is the mean signed change of the IO-minus-subject contrast. Repeated budgets share pairs and dense outputs.

Setting	Dense accuracy	Dense effect MSE	Dense task effect
GPT-2, b8, 25%, d0	1.000000	0.062497	-0.895905
GPT-2, b8, 25%, d1	0.996094	0.062783	-0.918035
GPT-2, b5, 25%, d1	0.996094	0.314839	-5.699345
GPT-2, b8, 50%, d0	1.000000	0.062497	-0.895905
DeiT, b8, 25%, d0	0.882812	0.135344	–
DeiT, b8, 50%, d0	0.882812	0.135344	–

Table 14: **Additional language metrics distinguish two-candidate and vocabulary-wide behavior.** Full-vocabulary top-1 uses the correct IO token; candidate IIA compares the two-name decisions under patching. Signed task error is reduced minus dense intervention contrast.

Method	Full-vocabulary top-1	Candidate IIA	Signed task error
GPT-2, b8, 25%, d0			
VBP	0.886719	0.996094	-0.014880
CMR-local	0.894531	0.996094	-0.033570
E2E-natural	0.902344	1.000000	-0.050797
E2E-augmented	0.914062	1.000000	-0.074923
E2E-paired	0.910156	1.000000	-0.064026
Augmented-reconstruction	0.917969	0.996094	-0.054054
GPT-2, b8, 25%, d1			
VBP	0.894531	0.988281	-0.017795
CMR-local	0.894531	1.000000	-0.038421
E2E-natural	0.917969	1.000000	-0.024449
E2E-augmented	0.914062	1.000000	-0.010880
E2E-paired	0.914062	1.000000	-0.023012
Augmented-reconstruction	0.925781	1.000000	0.025876
GPT-2, b5, 25%, d1			
VBP	0.929688	0.980469	-0.076131
CMR-local	0.914062	0.992188	-0.129112
E2E-natural	0.921875	0.968750	-0.242310
E2E-augmented	0.914062	0.988281	-0.059179
E2E-paired	0.906250	0.988281	-0.060541
Augmented-reconstruction	0.917969	0.976562	0.527646
GPT-2, b8, 50%, d0			
VBP	0.914062	0.996094	-0.102612
CMR-local	0.894531	1.000000	-0.087358
E2E-natural	0.914062	1.000000	-0.051583
E2E-augmented	0.910156	1.000000	-0.074227
E2E-paired	0.910156	1.000000	-0.038532
Augmented-reconstruction	0.917969	0.988281	-0.058529

H CONDITIONAL PAIRED INTERVALS

Table 15: **Paired-minus-control differences with conditional 95% intervals.** Differences are paired minus control; all differences and interval endpoints are multiplied by 10^3. Aug. and Recon. denote E2E-augmented and Augmented-reconstruction. Intervals are paired 95% bootstrap intervals with 1,000 resamples, conditional on the fixed checkpoint, draw, site, and input family. They are unadjusted.

Setting	Control	ΔM_{effect}	95% interval	ΔK_0	95% interval
GPT-2, b8, 25%, d0	Aug.	-0.06175	$[-0.07665, -0.04755]$	-1.10606	$[-1.43669, -0.79982]$
GPT-2, b8, 25%, d0	Recon.	-1.59772	$[-1.67514, -1.52350]$	-5.05108	$[-6.03006, -4.15776]$
GPT-2, b8, 25%, d1	Aug.	-0.10795	$[-0.12303, -0.09273]$	-0.34304	$[-0.66563, -0.01472]$
GPT-2, b8, 25%, d1	Recon.	-1.29663	$[-1.35244, -1.23996]$	-0.82384	$[-1.57462, -0.14377]$
GPT-2, b5, 25%, d1	Aug.	-0.39457	$[-0.46188, -0.33155]$	+0.63768	$[+0.45491, +0.83843]$
GPT-2, b5, 25%, d1	Recon.	-7.98162	$[-8.46626, -7.47365]$	-27.30050	$[-29.72149, -24.94772]$
GPT-2, b8, 50%, d0	Aug.	-0.25235	$[-0.29671, -0.21025]$	+1.94515	$[+1.59046, +2.32793]$
GPT-2, b8, 50%, d0	Recon.	-3.25057	$[-3.42674, -3.09099]$	-10.27766	$[-12.23223, -8.56699]$
DeiT, b8, 25%, d0	Aug.	-0.01547	$[-0.04273, +0.00741]$	-0.04362	$[-0.29020, +0.16045]$
DeiT, b8, 25%, d0	Recon.	-0.01193	$[-0.04428, +0.01678]$	-0.03439	$[-0.23330, +0.15113]$
DeiT, b8, 50%, d0	Aug.	-0.17511	$[-0.61950, +0.21217]$	+0.51735	$[-1.88917, +3.47737]$
DeiT, b8, 50%, d0	Recon.	-0.01149	$[-0.45522, +0.39738]$	+2.54709	$[+0.46426, +5.22919]$

The intervals describe sampling variation over the held-out pairs; Language resampling is stratified within the four held-out templates. The intervals do not account for variation across training checkpoints, and the two draw-1 sites share test pairs.

I APPROXIMATION DIAGNOSTICS AND IMPLEMENTATION CHECKS

Table 16: **Derivative stability, nonlinear approximation, and sketch error are separate checks.** Each row uses the first eight diagnostic pairs. G/D denote GPT-2/DeiT, and paired-1/4 denote fitted lambda candidates. FD is the relative RMS discrepancy between selected finite-difference estimates; Lin. is linearization relative RMS error. Exact and linearized losses use the same centered $\lambda = 1$ diagnostic for every row. The last two columns are independent projection estimates, not new fitted methods.

Setting	Fit	FD (%)	Lin. (%)	Exact loss	Linear loss	4 probes	16 probes
G, b8, 25%, d0	augmented	0.498	11.94	0.013863	0.014208	0.012223	0.016056
G, b8, 25%, d0	paired-1	0.446	11.81	0.016720	0.017630	0.013429	0.014112
G, b8, 25%, d0	paired-4	0.480	11.75	0.015141	0.015208	0.012961	0.017571
G, b8, 25%, d1	augmented	0.640	9.43	0.008713	0.008725	0.008703	0.009286
G, b8, 25%, d1	paired-4	0.689	8.82	0.007602	0.007513	0.007104	0.006967
G, b5, 25%, d1	augmented	0.703	14.87	0.007759	0.008044	0.007809	0.008245
G, b5, 25%, d1	paired-4	0.744	15.62	0.007019	0.006939	0.007284	0.007075
G, b8, 50%, d0	augmented	0.282	18.27	0.041119	0.044056	0.074593	0.052979
G, b8, 50%, d0	paired-1	0.282	18.02	0.041016	0.044228	0.059985	0.043544
G, b8, 50%, d0	paired-4	0.282	18.98	0.040553	0.043642	0.064544	0.052120
D, b8, 25%, d0	augmented	0.065	15.45	0.001467	0.001531	0.001633	0.001411
D, b8, 25%, d0	paired-1	0.064	15.50	0.001493	0.001566	0.001408	0.001370
D, b8, 25%, d0	paired-4	0.066	15.74	0.001392	0.001463	0.001486	0.001349
D, b8, 50%, d0	augmented	0.019	18.11	0.017705	0.017608	0.018165	0.018608
D, b8, 50%, d0	paired-1	0.020	20.30	0.013947	0.014088	0.014795	0.013416
D, b8, 50%, d0	paired-4	0.019	19.29	0.016524	0.016561	0.016557	0.016320

Finite differences use candidate steps 0.02, 0.01, and 0.005. The saved checks select stable estimates and compare full centered vectors for the completed replacement direction. They quantify how far the large deletion departs from the dense tangent approximation. The independent projection columns use the resulting linearized outputs and new probes. For example, at GPT-2 block 8, 25% removal, the same-data control's linearized loss is 0.014208; its four- and sixteen-probe estimates are 0.012223 and 0.016056. Increasing the rank ceiling does not imply a smaller error in every particular sketch.

Unit tests with affine-output oracles cover the natural, same-data ($\lambda = 0$), paired-1, and paired-4 losses, and invariance to a common logit shift is checked. Compilation checks use absolute tolerance

10^{-4} and relative tolerance 10^{-5} for natural and patched logits, and also compare prefixes. The independent CPU artifact audit checks folded parameters, selected sets, bounds, and quadratic construction and reproduces the saved primary metrics exactly from the stored outputs. Independent stock-model CPU checks cover one calibration input and projection in both executions for each evaluated site/draw.

J PRONOUN-CONTINUATION FOLLOW-UP

The pronoun task transfers block 8, 25% removal, and $\lambda = 4$ from the IOI setting, and refits masks/constants on two new calibration draws. The six methods, 16 projections, and two refinement sweeps are fixed. Diagnostic inputs serve only numerical and descriptive checks; no new λ selection is performed. It adds a second language behavior on the same GPT-2 checkpoint, evaluated under its own fixed protocol.

Input design and interpretation. Each pair changes exactly one single-token occupation and patches the dense donor's block-input residual at that occupation position. The 40 occupations split into 12 calibration, eight diagnostic, and 20 test occupations. There are 48/32/64 exact templates across these splits, formed from disjoint clause and context pools. The 144 templates instantiate 18 clause patterns. Each draw uses 128/64/256 pairs, and the two draws share no base/donor occupation-template combination. The stored audit verifies 1,792 unique texts/combinations, disjoint split pools, one-token changes, and fixed he/she token IDs.

Occupations are split into stereotype-associated pools by hand; these pools are heuristic, not measured demographic labels. They are balanced within each split and draw and are not used by the losses or the reported metrics. A lower contrast error means closer agreement with the dense model, which preserves its gender associations rather than reducing them.

Table 17: **Primary pronoun-task results for the same-data control and paired reduction.** A/P denote augmented/paired; MSE, KL, and absolute KL differences are multiplied by 10^3, while relative changes are percentages. KL is in nats before scaling.

Draw	Effect A	Effect P	Reduction (%)	KL A	KL P	Δ KL	Δ KL (%)
0	1.074727	0.953954	11.24	1.282610	1.322054	+0.039444	+3.08
1	0.913940	0.842596	7.81	1.248923	1.458708	+0.209786	+16.80

Table 18: **All 12 pronoun-task cells, with centered MSE and KL scaled by 10^3.** Each row uses 256 held-out pairs. KL is in nats before scaling.

Method	Natural MSE	Patched MSE	Effect MSE	Natural KL	Patched KL
Gender-occupation, draw 0					
VBP	10.195684	9.009980	3.366322	3.729206	3.328055
CMR-local	7.403531	6.543558	2.315723	2.191866	1.932501
E2E-natural	4.161314	4.093543	0.976692	1.219899	1.171846
E2E-augmented	3.878739	3.816345	1.074727	1.282610	1.273056
E2E-paired	4.485501	4.464376	0.953954	1.322054	1.292901
Augmented-reconstruction	36.939111	37.079523	12.075610	11.346893	10.998796
Gender-occupation, draw 1					
VBP	10.413402	9.328610	3.392061	3.702267	3.366546
CMR-local	7.630049	6.920770	2.542599	2.211866	1.956917
E2E-natural	3.863314	3.761553	0.941977	1.161662	1.122324
E2E-augmented	4.085763	4.040878	0.913940	1.248923	1.224599
E2E-paired	4.344465	4.270429	0.842596	1.458708	1.434946
Augmented-reconstruction	43.511243	42.583747	11.500109	14.161016	14.052383

The 1,000 paired bootstrap resamples preserve each of the 64 fixed held-out templates' sample sizes. Relative intervals use ratios of resampled paired means. They are unadjusted and conditional on the checkpoint, calibration draw, and fixed occupation/template pools; they do not cover new checkpoints or input families. Both draws share those split pools even though their exact texts are disjoint.

Table 19: **Pronoun-contrast fidelity and output agreement are distinct endpoints.** Task MAE and signed error use the intervention-induced he-minus-she logit change. Both error columns are in unscaled logit units; signed error is reduced minus dense intervention contrast. Agreements are fractions relative to the dense model; candidate IIA compares he/she decisions, while IIA uses all vocabulary logits.

Method	Task MAE	Signed error	Natural agreement	IIA	Candidate IIA
Gender-occupation, draw 0					
VBP	0.050866	0.004673	0.976562	0.953125	0.988281
CMR-local	0.050178	0.004318	0.996094	0.980469	0.992188
E2E-natural	0.029239	0.003950	0.996094	0.976562	0.992188
E2E-augmented	0.029239	0.000823	0.996094	0.968750	0.992188
E2E-paired	0.031371	-0.001432	0.988281	0.976562	0.992188
Augmented-reconstruction	0.093804	-0.007990	0.984375	0.957031	0.992188
Gender-occupation, draw 1					
VBP	0.042671	0.004334	0.976562	0.949219	1.000000
CMR-local	0.046832	0.004633	0.976562	0.976562	1.000000
E2E-natural	0.023932	-0.000551	0.980469	0.988281	1.000000
E2E-augmented	0.028636	-0.003059	0.976562	0.980469	0.996094
E2E-paired	0.025846	-0.002432	0.968750	0.968750	0.996094
Augmented-reconstruction	0.067244	-0.004320	0.960938	0.972656	0.992188

Table 20: **Pronoun-task paired-minus-control intervals.** Absolute differences are paired minus control and unscaled. Relative effect MSE is a reduction; relative KL and task MAE are changes, so positive values indicate worse fidelity for those two rows. Aug./Recon. are the two strong controls.

Draw	Control	Metric	Difference	95% interval	Relative (%)	95% interval (%)
0	Aug.	Effect MSE	-0.00012077	$[-0.00016923, -0.00007295]$	+11.24	$[+6.99, +15.19]$
0	Aug.	Natural KL	+0.00003944	$[+0.00000216, +0.00007306]$	+3.08	$[+0.16, +5.77]$
0	Aug.	Task MAE	+0.00213167	$[-0.00085474, +0.00534860]$	+7.29	$[-2.77, +19.43]$
0	Recon.	Effect MSE	-0.01112166	$[-0.01169673, -0.01054410]$	+92.10	$[+91.71, +92.46]$
0	Recon.	Natural KL	-0.01002484	$[-0.01031072, -0.00974558]$	-88.35	$[-88.73, -87.94]$
0	Recon.	Task MAE	-0.06243384	$[-0.06943015, -0.05526975]$	-66.56	$[-69.41, -63.29]$
1	Aug.	Effect MSE	-0.00007134	$[-0.00010187, -0.00003930]$	+7.81	$[+4.39, +11.01]$
1	Aug.	Natural KL	+0.00020979	$[+0.00017715, +0.00023872]$	+16.80	$[+13.99, +19.40]$
1	Aug.	Task MAE	-0.00279006	$[-0.00496049, -0.00065961]$	-9.74	$[-16.80, -2.34]$
1	Recon.	Effect MSE	-0.01065751	$[-0.01136841, -0.01001831]$	+92.67	$[+92.25, +93.06]$
1	Recon.	Natural KL	-0.01270231	$[-0.01322459, -0.01218908]$	-89.70	$[-90.12, -89.29]$
1	Recon.	Task MAE	-0.04139784	$[-0.04759292, -0.03594760]$	-61.56	$[-66.22, -56.67]$

Independent CPU audits reproduce all held-out primary, KL, and task-contrast metrics and verify compiled weights, constant bounds, quadratics, tokenization, and disjointness. One calibration forward pass and output-projection derivative per draw are rechecked against the unmodified model in both executions. On eight diagnostic pairs, finite-difference relative discrepancies span 1.45–1.79% and local linearization relative RMS errors span 4.26–5.88%. Hardware and numerical precision match the IOI study.