BEYOND RECONSTRUCTION: AUDITING SPARSE FEATURES FOR PHYSICAL CHANNEL PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) describe activations, but structured pruning must choose physical channels. We audit whether SAE features improve that decision in Qwen3-4B. Matched SAE and TopK-PCA representations generate fixed-width masks with the same joint search; statistical selectors receive common recovery rules. At 35.53% MLP-channel removal, SAE improves seven-task mean accuracy by 0.285 points and lowers WikiText perplexity by 0.834 versus RMS across three calibration seeds. PCA has comparable accuracy (49.52% versus 49.50% for SAE). Larger SAEs trained on over one million activation positions improve reconstruction and some local risk rankings, but neither greater sparsity nor active decoder subspaces consistently beat direct-gradient deletion scores. Common mean recovery changes rankings; FLAP-score and RCPU-score adaptations exceed SAE accuracy in the measured setting. A separate larger-calibration audit favors SAE on perplexity but favors a random dictionary on task accuracy. These results identify a limited reconstruction-proxy benefit without establishing SAE-specific functional preservation or superior physical deletion decisions.

1 INTRODUCTION

A sparse feature and a removable channel are different objects. Superposition allows a function to span multiple neurons and a neuron to participate in multiple functions (Elhage et al., 2022). Sparse autoencoders (SAEs) supply candidate feature coordinates (Cunningham et al., 2023; Gao et al., 2024a), but deployment requires a concrete decision: remove coupled MLP channels while retaining model quality. Accurate activation reconstruction does not by itself identify which channels can be deleted together.

The missing step is a mapping from representation to deletion damage. Interactions between removed channels affect their joint output error; reconstruction residuals introduce further interactions, and downstream computation changes the effect of a local error. Even a useful mask can change rank against another mask after recovery. We therefore ask: **under matched search and a specified recovery rule, do sparse features improve physical channel decisions, and where does the evidence stop?**

We make this question testable with a controlled sequence: representation, deletion-risk proxy, fixed-budget mask, common recovery, and end-to-end evaluation. A signed TopK SAE reconstructs MLP activations; reconstructed second moments and a diagonal residual penalty define a joint deletion proxy. Equal-cost swaps refine an RMS-initialized mask. A PCA-coordinate control with the same feature count and TopK truncation receives the same samples, projections, initialization and search limits. One-shot RMS and MLP-only FLAP/RCPU score adaptations supply statistical comparisons. The SAE–PCA contrast tests representation under matched search; either joint selector versus RMS changes both representation and optimization.

Our contributions are three linked audits. **First, a matched representation-to-mask comparison** measures three budgets and three calibration seeds at the largest budget, with complete seven-task evaluation, WikiText perplexity and physical export. SAE improves mean accuracy by 0.285 points and PPL by 0.834 over RMS at 35.53% MLP removal, while TopK PCA attains comparable accuracy. A larger-calibration component study favors the full SAE proxy on PPL, while a random

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

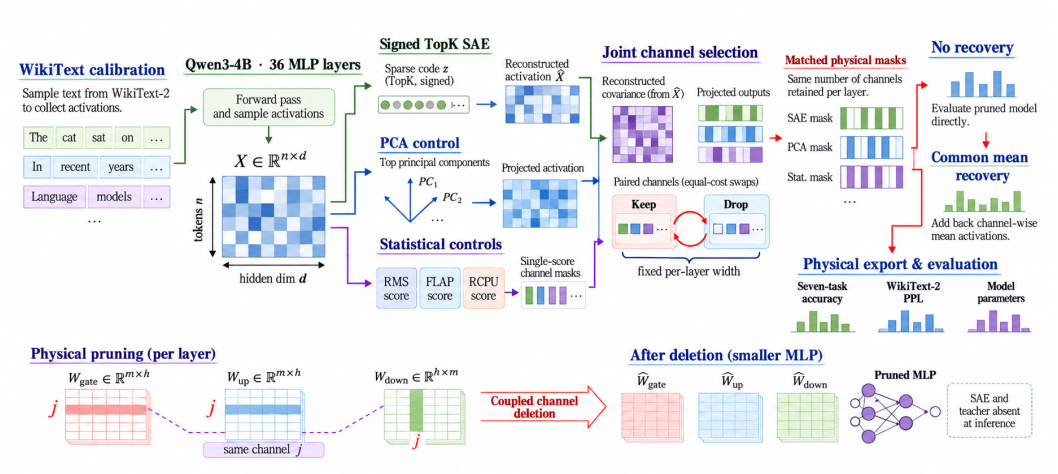

Figure 1: **Executed controlled study.** Shared data and physical budgets compare candidate selectors and common recovery. The PCA branch uses TopK-truncated PCA coordinates, as specified in Section 3.2. SAE and PCA share joint search; statistical baselines are one-shot adaptations, so these comparisons do not by themselves separate representation from search. Physical export is verified for the selected largest-budget RMS/SAE/PCA masks, not every recovery variant. The exploratory response pipeline is in Appendix C.

dictionary attains higher task accuracy under both recovery rules. **Second, a direct deletion-risk audit** compares frozen, equal-cost channel groups with actual single-layer damage: coarse discrimination does not establish reliable ordering near the pruning boundary. **Third, a recovery and behavior audit** shows that common mean correction changes selector rankings, while an exploratory feature-screened response objective does not improve the measured arithmetic behavior.

The final physical mask is the primary outcome; reconstruction and local deletion-risk correlations are explanatory diagnostics. We first compare complete masks, then ask why a learned representation did or did not change deletion decisions, and finally test recovery and behavior. The later million-position SAEs and active decoder subspaces improve parts of the diagnostic but do not overturn the mask comparison. The factorial response branch is a separate stress test, not part of the joint reconstruction selector. A stronger SAE-guided method must improve both actual deletion decisions and complete masks under the same information, budget and recovery rule.

2 Related Work and the Decision-Level Distinction

Compression units and recovery. LLM-Pruner removes coupled structures using gradient-based importance and applies LoRA recovery (Ma et al., 2023). SliceGPT rotates representations before slicing dimensions (Ashkboos et al., 2024), while ShortGPT ranks and removes transformer layers (Men et al., 2025). These change different structural units from our fixed-width MLP channel selection. Wanda combines weight magnitudes and input activations (Sun et al., 2024); Optimal Brain Compression and SparseGPT use layer reconstruction and second-order updates (Frantar et al., 2022; Frantar & Alistarh, 2023). FLAP combines fluctuation scoring, adaptive allocation and mean compensation (An et al., 2024); RCPU uses variance-aware selection and rotation-constrained recovery (Haruta et al., 2026). Our uniform MLP-only adaptations isolate components rather than reproduce these complete systems.

Features and their evaluation. Dictionary learning extracts interpretable features from a small transformer's MLP (Bricken et al., 2023), and Gemma Scope provides open SAEs across larger models and layers (Lieberum et al., 2024). Neither establishes the quality of our separately trained dictionary. SAEBench evaluates feature quality beyond reconstruction (Karvonen et al., 2025); feature absorption shows that a latent can miss inputs containing its apparent feature (Chanin et al., 2025). We therefore distinguish reconstruction, feature interpretation and useful deletion decisions.

Function-aware compression. FANG groups neurons by functional context, weights relevant tokens and preserves cross-context contributions (Yu et al., 2026); COLA studies calibration curation (He et al., 2025). FineScope uses SAE-selected data for gradient-based structured pruning and teacher-guided distillation (Bhattacharyya et al., 2025). Our proposed use instead scores candidate channel deletions directly; a static pruned model is a shared objective. Recent preprints study feature changes under weight pruning (Borobia et al., 2026), fixed-SAE robustness under unstructured pruning via activation-perturbation bounds (Gupte et al., 2026), and SAE-derived capability-density budget allocation (Gupta, 2026). Its GPT-2 Medium experiment reports worse perplexity than uniform budget allocation. These works already connect interpretability and compression; our contribution is the matched physical-channel selection and recovery study, not the first use of SAE information in pruning.

Interventions and matched evidence. Causal distillation matches teacher dynamics under interventions (Wu et al., 2022); ACDC automates circuit-edge discovery (Conmy et al., 2023). Activation-patching conclusions depend on corruption and metric choices (Zhang & Nanda, 2024). CausalGate studies module-ablation importance and CoAx conditional co-ablation (Nair et al., 2026; Gong et al., 2026). These motivate our separate feature-validation and physical-deletion tests, not a claim that they are interchangeable. Comparisons must share the checkpoint, calibration information, retained widths, recovery and evaluation protocol. Published scores on other backbones are not matched evidence; reconstruction gains alone do not validate functional deletion risk. The earlier association diagnostic remains in Appendix H.

3 Auditing the representation-to-deletion map

3.1 Physical budget and deployment setting

For layer l, let $x_l \in \mathbb{R}^{C_l}$ be the input to the down projection $W_l \in \mathbb{R}^{d_l \times C_l}$ and $m_l \in \{0,1\}^{C_l}$ the retained-channel mask. Deletion removes the same index from gate/up rows and down columns. For a bias-free gated MLP the parameter count is

$$P(m) = P_{\text{fixed}} + \sum_l 3d_l \|m_l\|_0. \tag{1}$$

All comparisons fix per-layer widths and retain attention; the budget isolates channel identities rather than layer allocation. Qwen3-4B (Qwen Team, 2025) has 36 MLPs of width 9,728. Requested removal r retains $64\lfloor 9728(1-r)/64 \rfloor$ channels per layer. At the largest budget, 6,272 channels remain: 35.53% MLP removal corresponds to 23.75% whole-model parameter removal. Recovery parameters are counted separately.

Common mean recovery defines the corrected-model comparison, with unrecovered masks reported alongside it to expose selection–recovery interaction. This deployment comparison is a retrospective extension, not a preregistered confirmation. Each constant correction is fitted from the same dense-input calibration activations, without selecting different recovery rules for different methods. Other recovery rules are secondary controlled interventions, not candidates from which to report a test-selected winner.

3.2 Executed selector: reconstructed joint deletion energy

We fit dictionaries at all 36 MLPs using standard WikiText-2 training text. Sixty-four disjoint 256-token windows provide 32 sampled activations per window; 51 windows fit the dictionary and 13 supply reconstruction diagnostics only. The signed TopK dictionary has 512 features, 32 active coordinates and 500 optimizer steps. It is an undercomplete learned representation, not a semantically validated SAE.

For column activation x and code z, let $\hat{x} = \mu + \sigma \odot (D^\top z)$, with decoder $D \in \mathbb{R}^{F \times C}$ and $C = 9728$. The frozen main SAE uses normalized activation reconstruction, $C^{-1}\mathbb{E}\|(x-\hat{x})/\sigma\|_2^2$. Development comparisons also considered output-aware fitting and diagonal mixtures; these are not the objective selected for the main tables. The matched PCA control uses the same standardized samples and feature count, with randomized low-rank PCA (three iterations). It also retains the 32 largest-magnitude coordinates per activation before reconstruction. Throughout the measured

main comparison, PCA denotes this TopK PCA-coordinate control, not a dense rank-512 linear reconstruction. Thus SAE–PCA compares learned dictionaries at matched coordinate sparsity; it does not isolate sparsity itself.

Selection retains interactions between deleted channels. Construct B with $B^\top B = \mathbb{E}[\hat{x}\hat{x}^\top]$, retaining mean and latent cross terms. For channel j, let b_j be its column and $u_j = Q^\top W_{:j}$. At a fixed number of retained channels, minimize

$$J(M) = \left\| \sum_{j \notin M} b_j u_j^\top \right\|_F^2 + \sum_{j \notin M} \mathbb{E}[(x_j - \hat{x}_j)^2] \|u_j\|_2^2. \tag{2}$$

The residual penalty omits cross-channel residual terms and reconstruction–residual cross terms, so this is an approximate layerwise proxy. Starting from matched RMS, each iteration screens 32 channels on each side by their individual restore/drop costs, evaluates all resulting swap pairs with their cross terms, and accepts only a proxy-decreasing swap, for at most 64 iterations per layer and budget. Each step preserves the budget. SAE and PCA use identical rank-64 independently seeded Q, initialization and search limits. A lower local proxy does not guarantee a lower end-to-end loss. The full standard-task evaluation tests that implication after development-only selection.

3.3 What proxy descent certifies

Equation 2 retains reconstructed cross-channel interactions, unlike summing feature energies independently. For fixed inputs, its first term equals expected projected reconstruction-induced deletion error. A same-cost swap has an exact quadratic cost difference, so accepting a strictly decreasing swap preserves channel count and decreases the proxy. These are local algebraic properties; a restricted candidate pool and a 64-step cap do not guarantee a local or global optimum. Proofs, the omitted-residual error bound and the mean-recovery derivation appear in Appendix E.

Independent feature interventions and improved full-mask decisions are required to connect this local objective to functional preservation. Appendix F specifies that unexecuted validation protocol.

3.4 A testable mechanism: proxy ranking versus deletion damage

Reconstructed cross terms can change a joint deletion ranking, but omitted reconstruction–residual and residual–residual cross terms can reverse that ranking again. Thus lower reconstruction error need not improve true local deletion-error ranking; even exact local ranking need not preserve nonlinear downstream behavior. Appendix E gives the exact decomposition, approximation bound and a margin counterexample. These statements predict a failure mode, not SAE superiority.

Our exploratory diagnostic freezes equal-cost deletion groups, compares SAE/PCA proxy scores with local projected errors, and measures single-layer dense-model damage on new calibration-disjoint windows from the same training corpus (Section 5.2). This separates representation fidelity, deletion-risk estimation and final mask quality. A constant mean correction is optimal only for local squared error on its fitting inputs; it can change downstream rankings. We report common recovered and unrecovered conditions rather than selecting the better condition for each method.

4 Experimental protocol

4.1 Complete standard-task evaluation and frozen revision

We extend every frozen pilot mask to seven complete labeled evaluation splits: BoolQ (Clark et al., 2019) (3,270), PIQA (Bisk et al., 2019) (1,838), HellaSwag (Zellers et al., 2019) (10,042), WinoGrande (Sakaguchi et al., 2019) (1,267), ARC-e (2,376), ARC-c (1,172) (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018) (500). Each configuration answers 20,465 questions. The zero-shot likelihood adapter follows inspected lm-eval 0.4.9.1 templates from the Language Model Evaluation Harness (Gao et al., 2024b), without chat formatting; it is a custom implementation, not a harness invocation. BoolQ and WinoGrande use accuracy; the other five use length-normalized accuracy. The earlier standalone PIQA raw-accuracy audit uses a different primary metric and is not mixed with these columns.

WikiText-2 (Merity et al., 2016) raw test uses all 299,078 tokens, joined with two newlines between rows. Non-overlapping 512-token blocks reset context and include the final 70-token block: 298,493 targets contribute to the token-weighted mean negative log likelihood before exponentiation. These results cannot be mixed with the earlier document-level prefix evaluation. Dataset hashes, prompts, masks, per-choice likelihoods and block losses are retained. CPU audits verify all 69,106 candidate boundaries and independently reconstruct the PPL token count.

The revision uses WikiText training calibration throughout all 36 layers. We compare the original frozen arithmetic-calibrated masks separately from the revised matched-calibration controls. Development selects among 13 candidates using 128 randomly sampled training examples per task and 64 WikiText validation blocks. Eligibility requires validation PPL no more than 2% above matched RMS; eligible candidates are ranked by seven-task mean, with lower PPL and simpler methods breaking ties. The selected activation-reconstruction SAE with joint swaps improves development mean by 1.23 points and validation PPL by 5.45%; these are selection statistics, not test evidence. Its settings and the PCA control are frozen before testing three budgets and two additional calibration seeds at the largest budget.

5 Experiments: representation, risk, recovery, and behavior

5.1 Experiment I: Do sparse representations improve the final mask?

Every pruned row of Table 1 uses the same checkpoint, 6,272 retained channels per layer, 1,632 fitting activations from 51 WikiText windows, and complete evaluation protocol. The 13 diagnostic windows do not fit scores or recovery. Mean correction adds 92,160 bias parameters uniformly. WANDA-sp and RMS retain exactly the same indices in all layers, so a verified run represents both labels.

The source methods are Wanda (ICLR 2024; Sun et al., 2024), FLAP (AAAI 2024; An et al., 2024), and RCPU (ICLR 2026; Haruta et al., 2026). FLAP uses the preserved public WIFV accumulator recurrence in FP32. RCPU uses an independent implementation of its variance-weighted score. Attention pruning, global allocation and sequential recalibration from the original systems are outside these adaptations' scope. SAE and PCA use the same projected joint-swap procedure. All extended baseline results are seed-zero exploratory comparisons on public tests already observed during the project.

Table 1: Direct Qwen3-4B comparison, seed zero: identical 6,272 retained MLP channels per layer (35.53% removed), calibration activations and full test protocol. All values are measured here. †Uniform MLP-only adaptations, not complete published systems. Seven accuracies and mean are percentages (↑); PPL ↓. Panels differ only in a common mean-recovery rule. Best pruned value per column within each panel is bold. Source venues: Wanda (ICLR 2024), FLAP (AAAI 2024), RCPU (ICLR 2026); venue labels refer to the original methods, not our adaptations.

Method	BQ	PIQA	HS	WG	ARC-e	ARC-c	OBQA	Mean	PPL
Dense	85.02	74.86	68.46	65.90	78.37	53.75	40.00	66.62	18.42
A. No recovery									
RMS / WANDA-sp	62.17	63.66	47.90	57.54	45.50	31.23	**32.40**	48.63	25.92
FLAP score†	56.27	62.57	43.91	54.06	43.39	26.45	30.60	45.32	28.63
RCPU score†	**65.90**	64.47	**48.71**	56.91	**48.40**	31.40	31.60	**49.63**	26.12
PCA joint control	63.12	64.58	48.47	**58.09**	44.82	31.23	31.40	48.82	25.21
SAE joint (ours)	63.98	**64.69**	48.58	56.20	45.62	**31.48**	31.40	48.85	**24.85**
B. Same mean recovery for every mask									
RMS / WANDA-sp	62.20	65.02	48.31	**57.70**	49.20	31.06	31.80	49.33	26.07
FLAP score†	**72.94**	64.31	48.04	56.27	**51.26**	**31.66**	**33.20**	**51.10**	26.44
RCPU score†	69.69	64.96	48.88	56.04	50.72	30.72	31.60	50.37	**26.03**
PCA joint control	63.43	64.96	**48.94**	57.38	47.98	30.63	32.40	49.39	26.47
SAE joint (ours)	64.37	**65.23**	48.86	56.43	48.44	31.31	32.60	49.61	26.23

Without recovery, SAE has lower PPL than every displayed pruned mask but lower accuracy than the RCPU-score adaptation. With common mean correction, RCPU improves both metrics over SAE, while FLAP has higher accuracy and slightly worse PPL. These results do not support a uniformly superior SAE selector. Mean correction changes FLAP accuracy by $+5.78$ points, compared with $+0.76$ for SAE and $+0.75$ for RCPU. BoolQ supplies about 41% of FLAP's macro-mean increase. The reversal is a mask–recovery interaction, not a reason to choose a different recovery condition for each row.

Ten thousand paired-item bootstrap draws, stratified by task, give SAE–RMS unrecovered accuracy difference $+0.22$ points with exploratory 95% interval $[-0.32, 0.76]$. Under mean correction, SAE–FLAP is -1.49 points $[-2.21, -0.78]$ and SAE–RCPU is -0.76 $[-1.48, -0.04]$. Recovered SAE's PPL is 0.77% higher than RCPU, with paired-block interval $[0.37, 1.17]\%$. These unadjusted intervals describe fixed items; they are not calibration-seed uncertainty, and shared contexts may narrow them.

Table 2: No-recovery budget comparison, calibration seed zero. Accuracy is the seven-task macro mean (%, ↑); PPL ↓. Bold denotes the best pruned value within each budget. Complete task values appear in Table 12.

Selector	MLP removal (%)	Accuracy	PPL
Dense	0.00	66.62	18.42
RMS	10.53	61.70	19.01
PCA joint	10.53	**62.02**	**18.69**
SAE joint (ours)	10.53	61.69	18.77
RMS	20.39	58.16	20.80
PCA joint	20.39	57.81	20.35
SAE joint (ours)	20.39	**58.32**	**20.29**
RMS	35.53	48.63	25.92
PCA joint	35.53	48.82	25.21
SAE joint (ours)	35.53	**48.85**	**24.85**

Table 3: Three-calibration-seed comparison at 35% requested MLP removal (35.53% actual). PCA uses TopK-truncated coordinates. Mean ± sample SD across seeds; differences pair each candidate with its same-seed RMS. Repeated benchmark questions are not independent new tests. Best means are bold (accuracy and its delta ↑; PPL and its delta ↓); SD is not ranked.

Method	Seven-task mean	PPL	Δ accuracy (pp)	Δ PPL
Matched RMS	49.22 ± 1.02	26.15 ± 0.26	0.00 ± 0.00	0.00 ± 0.00
PCA joint control	$\mathbf{49.52} \pm 1.22$	25.47 ± 0.24	$\mathbf{0.30} \pm 0.20$	-0.68 ± 0.10
SAE joint (ours)	49.50 ± 1.02	$\mathbf{25.31} \pm 0.42$	0.28 ± 0.07	$\mathbf{-0.83} \pm 0.20$

The matched representation contrast: SAE versus PCA. Across the three requested budgets, seed-zero SAE changes accuracy relative to RMS by $-0.01, +0.16, +0.22$ points and relative PPL by $-1.23, -2.42, -4.12\%$. Table 2 summarizes the cost–quality trade-off; Appendix D retains every task, including weaker WinoGrande and OpenBookQA outcomes. Figure 2 plots the actual MLP removal, not the smaller whole-model parameter removal.

At the largest budget, the paired three-seed accuracy improvement is 0.285 ± 0.066 points and absolute PPL improvement is 0.834 ± 0.205 (sample SD). Directions agree across all three calibration seeds. However, PCA reaches mean accuracy 49.52% versus SAE's 49.50%; mean PPL is 25.47 versus 25.31. The SAE–PCA contrast isolates representation under the same search; its small numerical PPL difference does not establish a sparse-feature mechanism. The joint-selector–RMS contrast changes both representation and search, so it cannot isolate an optimization benefit. FLAP/RCPU have only one matched calibration seed here, and the repeated benchmark questions are not independent test sets.

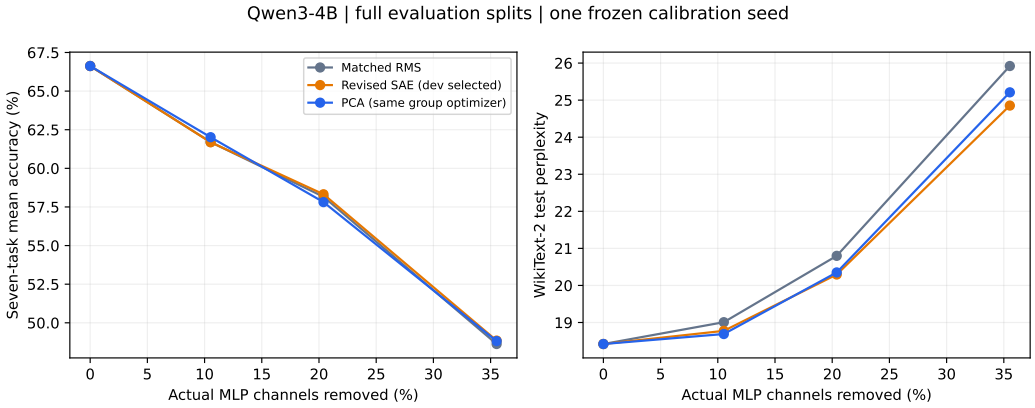

Figure 2: Frozen seed-zero accuracy and PPL against actual MLP-channel removal. All points use complete evaluation splits; the dense point is shared. Three calibration seeds are available only at the largest budget. Task-level curves are in Appendix D.

5.2 Experiment II: Does the proxy predict actual deletion damage?

We freeze equal-cost groups of 64 channels at layers 6, 18 and 30, then compare training-only risk predictions with actual deletion damage. A broad sanity population contains 33 unique groups from the low/high tails of six scores. A narrower population contains 36 groups near the RMS deletion boundary: among 128 randomly drawn groups per layer, six maximize SAE/PCA rank disagreement and six are randomly selected. Candidate construction uses calibration seed zero only. Every predictor evaluates the same frozen groups; three calibration seeds provide alternative predictions, not independent downstream repetitions.

Local outcomes use each seed's 13 unfitted diagnostic windows. Behavioral damage is signed ΔNLL from actual single-layer deletion in the dense model, measured on 16 new WikiText training windows excluding all three seeds' calibration windows. It is neither task accuracy nor damage to a fully pruned model. Correlations are computed within layer and seed, then averaged; intervals resample the same text windows jointly across candidates and methods. Appendix D.2 records selection, dependence and data-identity checks.

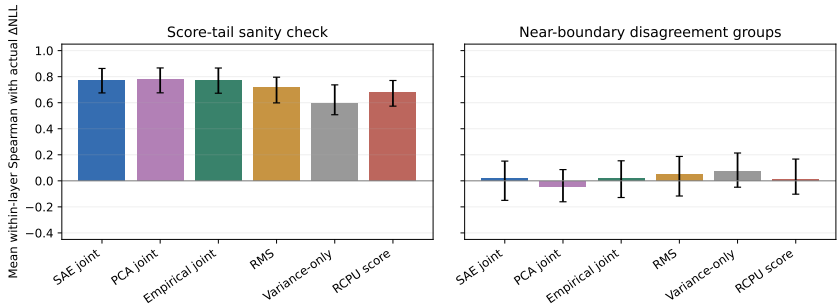

Figure 3: **Fixed-candidate deletion-risk diagnostic.** Mean within-layer/seed Spearman correlation with actual ΔNLL. Bars compare the same candidates within each population; 95% paired-window bootstrap intervals condition on the selected groups, layers and calibration fits. Broad-tail correlation alone does not establish near-threshold prediction. Variance-only is a diagnostic score, not official FLAP WIFV.

On broad tails, SAE/PCA/empirical joint risks correlate with actual ΔNLL at 0.774/0.776/0.775. Near the boundary, their correlations are $0.019/-0.044/0.019$; all six predictors' intervals include zero. SAE's local-error correlation also falls from 0.953 to 0.102. The near-boundary SAE–PCA correlation difference is 0.063 with paired interval $[-0.030, 0.108]$. Thus this BF16 diagnostic pro-

vides no stable evidence of an SAE-specific advantage in the selected threshold neighborhood. A full-float32 repeat gives SAE/PCA/empirical point correlations of .020/−.047/.016, retaining the same limited conclusion. Range restriction, finite text samples and dense single-layer interventions limit interpretation. The experiment neither isolates reconstruction residuals as the cause nor validates swaps within a fully pruned mask.

Training scale and alternative uses of the learned dictionary. We subsequently captured 1,048,576 training activation positions per layer at layers 6, 18 and 30, with a disjoint 32,768-position development cache. At fixed 8,192 features and TopK 32, increasing available fitting data and training to 64,000 steps lowers reconstruction MSE, but its deletion-risk correlation remains below the uncompressed first-order gradient in all four development panels (Appendix B). Increasing TopK from 32 to 128 improves some correlations: without recovery on broad candidates, the layer-mean Spearman correlation rises from 0.170 to 0.299; the direct-gradient control is 0.436. With no recovery near the decision boundary, TopK 128 reaches -0.021 versus 0.294 for direct gradient. Both sparsity settings, a random dictionary, and PCA remain reported in the appendix. These are three layers of one model, not three training seeds.

We also use each token's active SAE decoder vectors as a physical output subspace, retaining the true channel-deletion direction rather than differentiating through the SAE reconstruction. On broad candidates without recovery, this raises the SAE correlation from 0.170 to 0.356, still below the direct-gradient 0.436; near the boundary it is -0.002 versus 0.294. On previously inspected 32-window development outcomes, selecting low-risk groups using this projection gives higher average damage than direct gradient in three of four recovery/candidate panels. The fixed 11-rule comparison includes matched million-position PCA, random dictionary, shuffled SAE supports and direct gradient (Appendix B). No new standard benchmark or physical mask result is inferred from these local diagnostics.

Table 4: Additional development-only deletion-risk audit. Entries are equal-layer mean Spearman correlations with actual single-layer ΔNLL on the same 32 previously inspected windows; higher is better. K32 and K128 use the 64,000-step SAE checkpoints; "active" projects the physical output gradient onto active SAE decoder directions. The direct-gradient row retains the entire physical direction. These are fixed local groups, not completed masks.

Rule	None/broad	None/boundary	Mean/broad	Mean/boundary
Direct gradient	**0.436**	**0.294**	**0.282**	**0.133**
SAE TopK 32	0.170	−0.188	0.042	−0.138
SAE TopK 128	0.299	−0.021	0.237	0.103
SAE active subspace	0.356	−0.002	0.101	0.105

5.3 Experiment III: Does recovery change the selector ranking?

Holding masks fixed separates selector quality from weight or bias correction (Appendix Table 13). Mean recovery can improve task accuracy while worsening PPL; rank-eight ridge also helps some aggregate scores without establishing semantic preservation. Rank-64 projected rotation is a local control, not full RCPU recovery. The complete rotation/scale and correction-location comparisons are in Appendix D.

A fixed-mask diagnostic finds that the same fitted mean lowers full-output local MSE in all 180 layer–method settings on the 13 historical diagnostic windows. FLAP's median removed error fraction is 7.99%, versus 3.97/4.39/4.36/4.17% for RMS/SAE/PCA/RCPU; its larger locally removable mean component is consistent with the larger recovery gain, but does not causally explain downstream rankings. Appendix D.1 reports the fit-only prediction, per-layer correlations and limitations.

Full 2,560-dimensional rotation, fitted from only 1,632 rows, lowers training MSE but worsens diagnostic MSE in every layer of the three tested masks. Appendix D reports the degradation and rank-deficiency analysis; this local adaptation is distinct from the full RCPU system.

Physical export reduces the largest-budget RMS/SAE/PCA models from 4.022B to 3.067B parameters. Reloaded PPL shifts stay below 0.05%, but BF16 rounding reverses the small SAE/PCA

accuracy ordering. Appendix Table 18 gives the measurements; controlled throughput remains unmeasured.

5.4 Experiment IV: Do feature associations transfer to behavior?

A separate task-defined arithmetic/language experiment screens stable feature effects and ranks channel candidates by teacher-response matching (FRP). Its frozen three-seed masks obtain 24.22/29.69/15.62% correct completed answers on 128 held-out prompts, versus 32.03% for the unchanged RMS mask under the same 128-token generation cap. Stable association is not intervention-validated functionality. This result does not support upgrading reconstruction selection into a functional pruning claim. The complete protocol, response equations, SAE/PCA/random controls and fixed-pool sensitivities appear in Appendix C; the latter are exploratory re-rankings, not independent optimization runs.

5.5 Component ablations: representation, selection, and recovery

The larger, separately matched seed-zero audit uses 2,048 features, TopK 32 and 13,088 fitting activations; all masks retain 6,272 channels per layer. The full SAE proxy reaches PPL 24.77 without recovery versus 25.62 for diagonal selection under the same 64-swap budget, but the random TopK control has higher seven-task accuracy: 50.59/51.90% without/with recovery versus SAE's 49.53/50.58%. An exploratory equal-weight SAE–random objective mixture fails to improve either metric over SAE on the same host. The complete measured table, recovery protocols and all mixture weights are in Appendix S; these results do not establish a functional SAE advantage.

6 Discussion and limitations

The original matched TopK PCA comparison does not identify an SAE-specific accuracy gain. The larger-calibration ablation favors the full SAE proxy on PPL, but a random dictionary achieves higher accuracy under both recovery rules; the two calibration regimes also differ in several training settings. The later million-position training audit improves reconstruction, and TopK 128 improves one local risk panel, yet direct gradient remains stronger in the fixed development comparisons. Active SAE decoder subspaces likewise do not produce a reliable boundary-selection gain. A directional-curvature probe greatly improves same-input loss approximation but leaves the observed cross-window choice unchanged; its NLL finite differences remain precision-sensitive. Boundary-risk uncertainty, recovery-dependent rankings and the negative factorial test distinguish reconstruction benefit from validated functional preservation.

The study covers one backbone and uniform MLP widths. The original compared dictionaries use 512 or 2,048 features versus 9,728 physical channels; later 8,192-feature training covers only three layers. Three-seed evidence covers RMS/SAE/PCA only at the largest original budget; the expanded component study and FLAP/RCPU adaptations have one seed. Public tests were repeatedly inspected. The later risk diagnostics reuse previously inspected development windows and fixed dense single-layer interventions, not swaps in a fully pruned model. Feature associations lack independent causal validation.

7 Conclusion

The larger-calibration SAE proxy improves PPL over empirical moments and linear PCA, but random features yield higher accuracy with and without recovery. The original three-seed study finds comparable SAE/TopK-PCA accuracy. More extensive SAE training and a new active-subspace score do not establish a better physical deletion decision. The evidence supports continued direct damage and final-mask audits under matched search and recovery; a learned feature basis remains an unproven source of additional pruning benefit.

AI USAGE DISCLOSURE

Generative AI tools were used for research ideation, literature discovery, formulation of hypotheses and mathematical claims, experimental design, implementation and debugging, analysis, proof drafting, visualization and manuscript writing. The arithmetic test families are programmatically constructed; their design and generation code were AI-assisted. Benchmark measurements come from executed model evaluations, not language-model predictions of experimental outcomes. Automated checks reconstruct metrics from raw outputs, compare masks and data hashes, verify export tolerances and compile citations. Mathematical arguments received same-family model checks, which are not independent scientific validation. These checks do not replace human author verification of the claims, proofs, citations and artifacts before submission; the human authors retain responsibility for the final contents.

REPRODUCIBILITY STATEMENT

The appendix gives the fixed calibration/evaluation protocol, complete chain and recovery controls, proof assumptions and artifact inventory. Raw predictions, mask hashes, retained indices, calibration identities and physical export records are preserved in the research artifacts. All later comparisons on previously inspected public benchmarks are identified as exploratory. Unexecuted functional-risk tests are explicitly separated from measured results.

REFERENCES

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Fluctuation-based adaptive structured pruning for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10865–10873, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i10.28960. URL http://dx.doi.org/10.1609/aaai.v38i10.28960.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vXxardq6db.

Chaitali Bhattacharyya, Hyunsei Lee, Junyoung Lee, Shinhyoung Jang, Il Hong Suh, and Yeseong Kim. FineScope: SAE-guided data selection enables domain specific LLM pruning and finetuning. *arXiv preprint arXiv:2505.00624*, 2025. URL https://arxiv.org/abs/2505.00624v3.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. arXiv preprint, 2019. URL https://arxiv.org/abs/1911.11641.

Hector Borobia, Elies Seguí-Mas, and Guillermina Tormo-Carbó. How pruning reshapes features: Sparse autoencoder analysis of weight-pruned language models. *arXiv preprint arXiv:2603.25325*, 2026. URL https://arxiv.org/abs/2603.25325.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82318–82355, 2025. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv preprint, 2018. URL https://arxiv.org/abs/1803.05457.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. arXiv preprint, 2023. URL https://arxiv.org/abs/2309.08600.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337, 2023. URL https://proceedings.mlr.press/v202/frantar23a.html.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/1caf09c9f4e6b0150b06a07e77f2710c-Paper-Conference.pdf.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. arXiv preprint, 2024a. URL https://arxiv.org/abs/2406.04093.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, July 2024b. URL https://zenodo.org/records/12608602.

Zhiren Gong, Zihao Zeng, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. *arXiv preprint arXiv:2607.01940*, 2026. URL https://arxiv.org/abs/2607.01940v1.

Rishaank Gupta. Capability-guided compression: Toward interpretability-aware budget allocation for large language models. *arXiv preprint arXiv:2603.16440*, 2026. URL https://arxiv.org/abs/2603.16440.

Suchit Gupte, Xueru Zhang, and Mohammad Mahdi Khalili. When pruning meets interpretability: Preserving sparse autoencoder robustness in LLMs. *arXiv preprint arXiv:2608.25941*, 2026. URL https://arxiv.org/abs/2608.25941.

Shuichiro Haruta, Kazunori Matsumoto, Zhi Li, Yanan Wang, and Mori Kurokawa. RCPU: Rotation-constrained error compensation for structured pruning of large language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.07782.

Bowei He, Lihao Yin, Huiling Zhen, Shuqi Liu, Han Wu, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. Preserving LLM capabilities through calibration data curation: From analysis to optimization. *arXiv preprint arXiv:2510.10618*, 2025. URL https://arxiv.org/abs/2510.10618.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024. doi: 10.48550/arXiv.2408.05147. URL https://arxiv.org/abs/2408.05147.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/44956951349095f74492a5471128a7e0-Abstract-Conference.html.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1035. URL https://aclanthology.org/2025.findings-acl.1035/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint, 2016. URL https://arxiv.org/abs/1609.07843.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Kiran Nair, Smriti Regmi, and Rodrigue Rizk. CausalGate: Causal importance distillation for transformer module pruning. *arXiv preprint arXiv:2607.22720*, 2026. URL https://arxiv.org/abs/2607.22720v2.

Qwen Team. Qwen3-4B model card. Hugging Face model repository, 2025. URL https://huggingface.co/Qwen/Qwen3-4B. Official model card, accessed 2026-09-19.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. arXiv preprint, 2019. URL https://arxiv.org/abs/1907.10641.

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*, 2024. URL https://arxiv.org/abs/2306.11695.

Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D. Goodman. Causal distillation for language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4288–4295, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.318. URL https://aclanthology.org/2022.naacl-main.318/.

Tao Yu, Yongqi An, Kuan Zhu, Guibo Zhu, Ming Tang, and Jinqiao Wang. Improving generalization in LLM structured pruning via function-aware neuron grouping. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(33):27969–27977, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i33.40021. URL http://dx.doi.org/10.1609/aaai.v40i33.40021.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

A Appendix guide and evidence inventory

The entries below classify evidence by its role, not by the number of configurations. Reusing the same public benchmark or re-ranking a frozen candidate pool does not create independent confirmation. No local theorem establishes semantic pruning gains.

Table 5: Evidence map. Completed and unexecuted components are separated throughout the appendix.

Status	Contents	Interpretation
Main measurements	Matched masks, budget curves, three calibration seeds for RMS/SAE/PCA, common recovery, physical export	Controlled single-backbone evidence; public tests repeatedly observed
Exploratory completed	Fixed-group risk ranking, local recovery, response chains, random/PCA controls, frozen-pool weights and sensitivities	Hypothesis probes; retain negative and adverse outcomes
Historical completed	Early arithmetic, compensation and correlation diagnostics	Different scoring/calibration protocols; not pooled with standard-suite results
Unexecuted protocol	Independent feature interventions and validated functional-risk mask design	Falsifiable future tests, not measured benefits
Mathematical results	Local objective identities, swap descent, residual bounds, recovery limitations	Conditional algebraic properties, not guarantees of task accuracy

B Expanded SAE training and physical deletion diagnostics

This appendix reports the subsequent algorithm-development campaign separately from the original matched-mask and public-test results. All local outcomes below reuse a 32-window WikiText training split whose labels have already been inspected during method development. The campaign did not produce a new full-model SAE mask, a new standard-suite result, or a second model. Consequently none of these tables is an independent confirmation or evidence of deployment benefit.

What was trained. The original all-layer study trained 36 separate SAEs with 2,048 features, TopK 32, 52,416 sampled activations per layer and 16,000 optimizer steps. The later capacity study trained only three representative layers (6, 18 and 30), not another 36-layer collection. It captured 4,096 training windows of 256 positions each (1,048,576 positions), 128 development windows (32,768 positions), and reserved 256 additional windows without evaluating them. In the large-data comparison, each layer has an 8,192-feature signed TopK 32 SAE trained for 16,000 and 64,000 steps; the small-data 16,000-step control has 65,536 available positions. A separate TopK 128 run uses the same data, initialization, normalization, learning-rate and minibatch schedule at each preset checkpoint. It changes the number of active coordinates and computation. These windows are not independent documents, and the old 52,416-row capture used a different position-sampling protocol; the comparison across those two campaigns cannot isolate data quantity alone.

Table 6: Raw development reconstruction MSE at the preset 64,000-step checkpoint, 8,192 features and 1,048,576 available fitting positions. K32/K128 share the data and optimizer schedule. Smaller MSE does not imply a better deletion decision.

Layer	TopK 32	TopK 128	Relative change
6	0.188427	0.205392	$+9.0\%$
18	0.052298	0.039214	-25.0%
30	0.692572	0.501832	-27.5%

At layer 6, the first sequence position has a much larger activation norm than ordinary positions. Its share of centered output energy reaches 99.57% on the development cache under the large-run training mean, yet its median share of the absolute first-order deletion contribution across tested

groups is only about 1.5–2.6%. At 64,000 steps, TopK 128 reduces ordinary-position reconstruction MSE by 23.2% relative to TopK 32 but increases first-position MSE by 147.4%, explaining the worse aggregate in Table 6. This is a position-specific measurement, not proof that the first token is functionally unimportant or an attention sink.

Training scale does not transfer monotonically to deletion ranking. Each method scores the same 64 physical channel groups at each of three layers under no recovery and common mean recovery. The 32 broad and 32 near-boundary groups per layer are fixed before the later predictors are compared. Prediction uses 64 training windows (2,048 sampled positions); damage comes from actual single-layer deletion measured on the 32 previously inspected development windows. Correlations are computed within layer, then averaged with equal layer weight. The bootstrap resamples the 32 windows jointly across candidates, conditional on these selected groups and learned weights. This does not provide uncertainty over training seeds or over method selection after repeated exploration.

The 64,000-step TopK 32 SAE has layer-mean Spearman correlations $0.170/-0.188/0.042/-0.138$ in the four panels of Table 4, while direct gradient has 0.436/0.294/0.282/0.133. TopK 128 improves the first panel by 0.129 (conditional paired interval $[0.016, 0.278]$) but its other three improvements over TopK 32 have intervals including zero. In the none/boundary panel its deficit versus direct gradient is -0.315 (interval $[-0.427, -0.089]$). A same-normalization random TopK 128 dictionary reaches 0.076 and 0.196 on the two boundary panels, both above the trained TopK 128 point estimates. All six K128 checkpoints and both position-weighting analyses remain in the experiment record.

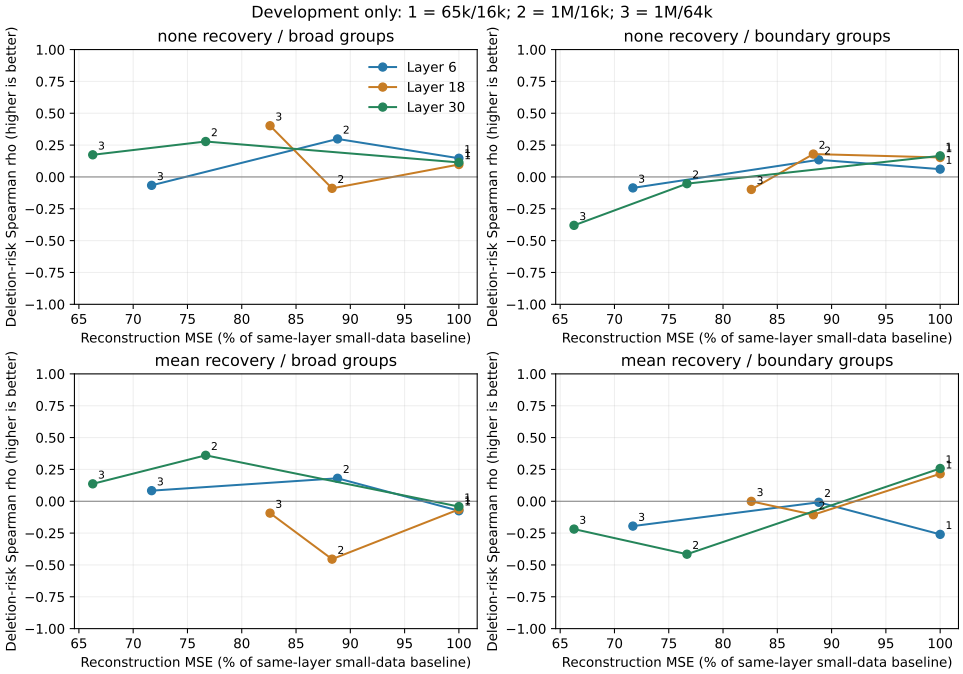

Figure 4: Reconstruction improvement versus deletion-risk ranking on development data. Each point is one of three layers at one of three preset TopK 32 checkpoints, not an independent model. The horizontal axis is development raw reconstruction MSE divided by that layer's small-data 16,000-step MSE; left is better. The vertical axis is within-layer Spearman correlation between a fixed candidate score and actual deletion ΔNLL; up is better. Separate panels retain both candidate populations and recovery rules. Lower reconstruction error does not consistently improve the ranking.

Active decoder directions and matched controls. For each selection token, let D_A be the decoder rows of its 32 active SAE features and let S undo training normalization. The active-subspace

rule scores a physical output deletion δ by $g^\top P_{\mathrm{span}(SD_A^\top)}\delta$, using a numerically rank-checked orthogonal projector. It retains the original physical δ and does not pass the deletion through the nonlinear SAE reconstruction map. Frozen-support Jacobian and dynamic reconstruction scores are separate diagnostics. We fit the PCA control on the same million-position training cache and normalization, keep 32 active directions, and also evaluate a same-size random dictionary, permuted SAE supports and unprojected direct gradient. PCA has at most 2,560 independent directions, so its candidate dictionary size is not 8,192. All 11 scoring rules, two recovery conditions and both candidate families are retained in Figure 5. The projector passes numerical rank and orthogonality checks, but its expected retained gradient energy is small; the output is a local prediction, not a functional feature validation.

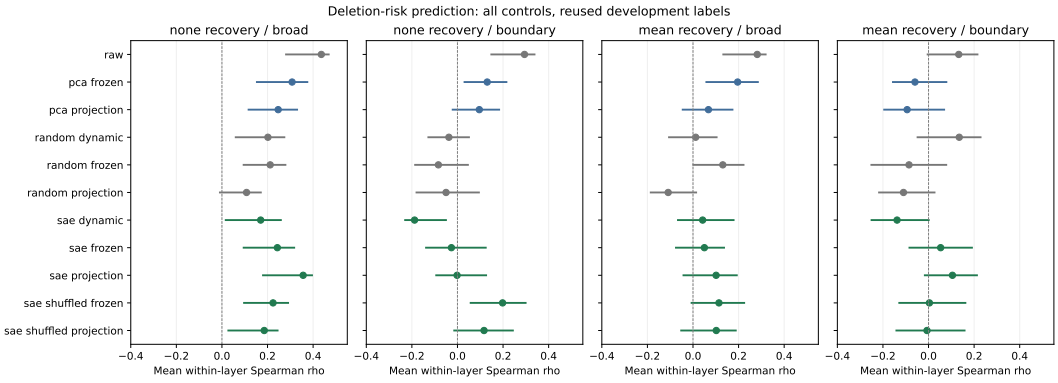

Figure 5: Every active-subspace development control. Horizontal position is the equal-layer mean Spearman correlation with actual single-layer deletion damage (higher is better); horizontal bars are conditional paired-window bootstrap intervals. Rows are scoring rules, not separate models. "Shuffled" permutes token support while retaining the trained dictionary. Direct gradient remains strongest in point estimate in all four panels; intervals do not include uncertainty from training or repeated method exploration.

The active SAE projection improves over dynamic SAE reconstruction in some panels, but does not dominate direct gradient. A post hoc decision-curve audit chooses the predicted lowest-risk group among every prefix size $k = 1, \ldots, 32$ and averages its actual damage. This is selection of one group from a k-group prefix, not simultaneous pruning of k groups. The projection's mean damage exceeds direct gradient in three of four panels, with the fourth unresolved by the conditional window interval. No active-subspace full-mask selection or physical export was run.

Directional curvature: a distinct, unsuccessful decision test. On one layer, eight fixed calibration windows and eight fixed channel groups, a full-FP32 model evaluates the same physical deletion direction at strengths $0, \pm 0.25, \pm 0.125, 1$. The first derivative uses the gradient of mean cross-entropy over 255 targets and all 256 intervention positions. A separate physical mask agrees with the strength-one intervention in loss to at most 1.44×10^{-6} nats per target across 64 pairs. On those same 64 pairs, adding the finite-difference NLL curvature at step 0.25 reduces absolute prediction error from 2.14×10^{-4} to 2.73×10^{-5} nats per target. However, the direct gradient already orders the eight candidates correctly by their calibration mean. On 32 previously observed windows re-evaluated in matching FP32, the gradient, both NLL-curvature steps, KL curvature and even direct calibration deletion all give the same eight-candidate ranking (Spearman 0.667) and select the same group. This mechanism did not yield a new physical pruning decision; no SAE-conditioned curvature model was trained.

The NLL finite-difference curvature changes sign across the two preset step sizes in eight of 64 pairs, with median relative change 35.6%; the KL curvature is much more stable but is not the gold-label NLL Hessian. Two attempted higher-precision tail checks exceeded the fixed 28 GB GPU memory cap and were stopped; neither provides a precision-validated Hessian. These numerical and sample-size limits preclude an exact second-order or broad generalization claim. All inputs, candidate IDs, full 64-pair outputs, matched-FP32 comparison and failed precision logs are preserved in the campaign artifacts.

A separate calibration-sampling proposal. A later protocol proposes using SAE codes only to stratify a fixed 64-window finite population before sampling 8 or 16 windows for the unchanged direct-gradient risk. It would compare population-weighted, candidate-centered risk variance against PCA, raw activation, RMS, random dictionary and simple random sampling under identical budgets. The estimator and its finite-population variance pass an exhaustive toy identity check, but the representation-based strata and model experiment have not been completed. No sampling or pruning benefit is claimed from this proposal.

C COMPLETED FACTORIAL-RESPONSE CHAINS AND FIXED-POOL DIAGNOSTICS

This section reports the completed end-to-end extension separately from earlier pilot experiments. Its purpose is to test whether dictionary-associated channel proposals and a factorial response objective improve a fixed structured mask. All reported numbers are measured; neither a favorable algorithm seed nor a post hoc test-optimal response weight replaces the predeclared primary result. The extension does not establish autonomous semantic discovery by an SAE.

C.1 RESPONSE OBJECTIVE AND PRIMARY CHAIN COMPARISON

Table 7: Executed four-round SAE-feature candidate chain at 20.39% actual MLP removal. Left: correct completed answers (%) on 128 new arithmetic prompts, fixed 128-token limit. Right: full seven-task mean and WikiText test, mean ± sample SD over three algorithm seeds sharing calibration data. RMS/dense are shared controls. Best pruned value per column is bold (accuracy ↑, PPL ↓).

Selection	Gen. s0	Gen. s1	Gen. s2	Seven-task mean	PPL
Dense	100.00	100.00	100.00	66.62	18.42
RMS initial mask	32.03	**32.03**	**32.03**	58.16	**20.80**
Same-data KL	28.91	29.69	23.44	$\mathbf{58.31 \pm 0.17}$	20.93 ± 0.05
SAE → FRP	24.22	29.69	15.62	58.29 ± 0.16	20.97 ± 0.01
Same-data score MSE	**46.09**	24.22	28.91	57.94 ± 0.27	21.11 ± 0.04

The response experiment asks whether preserving how a teacher changes its answer across controlled inputs gives a useful signal beyond ordinary prediction matching. A quartet holds operands fixed while crossing addition/subtraction and English/Chinese. Discovery, feature screening, calibration and held-out generation use 16/8/16/32 quartets with distinct operand pairs. Teacher qualification requires all four cells to be correct in the discovery, screening and calibration stages. Stable feature effects propose candidates; the selected masks are frozen before the held-out generation test. This operational separation limits leakage within the chain, but stable effects are not intervention-validated functions.

Let v_{ab} be four answer-option log scores at a forced answer position. The response coordinates are

$$\begin{aligned} d_A &= (v_{10} - v_{00} + v_{11} - v_{01})/2, \\ d_B &= (v_{01} - v_{00} + v_{11} - v_{10})/2, \quad d_{AB} = v_{11} - v_{10} - v_{01} + v_{00}. \end{aligned} \tag{3}$$

With $R = [d_A, d_B, d_{AB}]$, FRP ranks candidates by conditional full-vocabulary KL plus $\lambda\|(R_m - R_T)S\|_F^2$. The primary chain fixes $\lambda = 1$ and teacher-scale floor 0.1. Same-data KL, score MSE, diagonal weighting and removal of the interaction term receive the same candidates. Quartet matching is thus a weighting of already observed answer scores, not additional labels; its quadratic form is derived in Appendix J. The standard-task comparison and generated-answer comparison measure distinct outcomes, and both are retained.

The completed task-defined arithmetic/language chain screens features in all layers and performs four shared rounds of eight candidates. Its frozen three-seed FRP masks obtain 24.22/29.69/15.62% correct completed answers on 128 held-out prompts, versus 32.03% for the unchanged RMS mask. All use the same 128-token generation cap; unfinished answers count as wrong. The corresponding seven-task mean is about 58.29%, versus 58.31% for same-data KL, and FRP PPL exceeds RMS

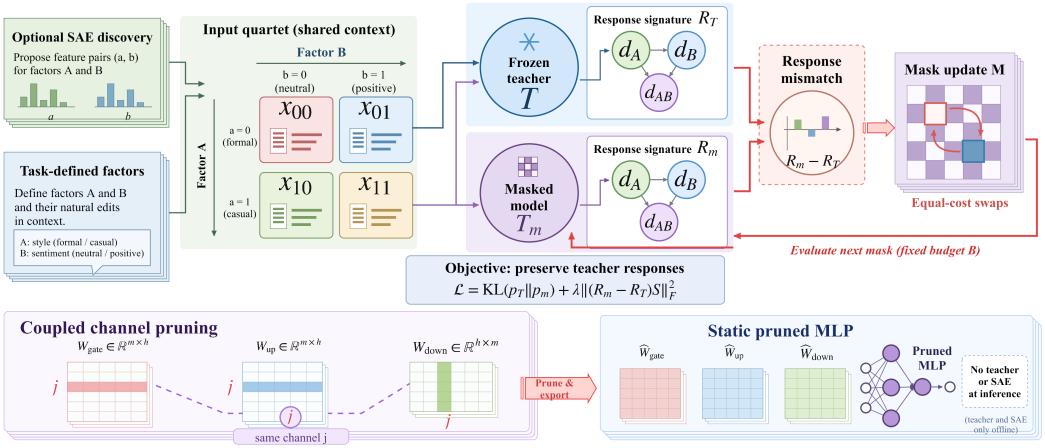

Figure 6: **Exploratory response-selection and export schematic.** The task-defined branch and feature-screened candidates are bounded diagnostics; optional SAE discovery is not independently validated semantic discovery. Response matching guides equal-cost swaps and coupled-channel export, with no teacher or SAE at inference. The figure does not claim functional superiority. The main quantitative study separately evaluates joint reconstruction selection.

at every seed. Execution of the full chain therefore does not establish a benefit from its response objective.

Completed PCA/random chains, all KL/FRP/MSE comparisons and the fixed-pool weight study are provided in Appendix C. The latter re-ranks frozen masks and is not a fresh solver at each hyperparameter. A favorable test-selected setting cannot replace the predefined result. Feature screening tests stable association, not the independent intervention-based validation required for a functional pruning claim.

C.2 Data, feature screening, and frozen selection

We use Qwen3-4B with a common RMS initialization in all 36 MLPs. The requested 20% width reduction is rounded to retain 7,744 of 9,728 channels per layer, for an actual width reduction of 20.3947%. Each mask removes the corresponding rows of the gate and up projections and columns of the down projection. The evaluations in this section use masks, so they do not constitute a standalone inference-speed measurement.

The factors are task-defined: addition versus subtraction and English versus Chinese. Each quartet holds operands and answer-option ordering fixed while crossing these two factors. Discovery, validation, calibration, and test use 16, 8, 16, and 32 quartets, respectively, with distinct operand pairs. Qualification requires the dense teacher to answer all four cells correctly and finish within 128 generated tokens. All discovery, validation, and calibration quartets passed in each completed chain. Test consists of 128 questions, and is not used to qualify calibration examples or choose masks.

The SAE is the previously fitted, frozen WikiText-training dictionary with 512 features and signed TopK sparsity 32. Features are screened using the last prompt position after a forced answer prefix. We require the mean effect to have the same sign in two discovery halves and validation, and a mean-to-RMS magnitude of at least 0.25 in each split. At most eight features per factor or interaction are retained per layer. Feature-effect second moments, squared raw decoder coefficients, and down-projection column norms produce channel proposal scores. This is a feature-associated heuristic, not a validated causal attribution or an exact decomposition of channel utility.

PCA uses the matched standardized calibration fit with 512 components and TopK 32. The random control uses the same mean and scale, a fixed orthogonal 512-feature encoder and transposed decoder, zero bias, and TopK 32; it receives no additional fitting examples. Thus the control changes the dictionary while retaining the factor definitions and screening/search protocol.

Table 8: Complete generated-answer objective controls (%, 128 questions). All methods in a row share calibration and candidate pools. Dense and RMS are 100.00 and 32.03 in every row. Seeds describe algorithm randomness only.

Dictionary	Seed	KL	FRP	MSE	No interaction	Diagonal
SAE	0	28.91	24.22	46.09	28.91	46.09
SAE	1	29.69	29.69	24.22	27.34	24.22
SAE	2	23.44	15.62	28.91	25.78	28.91
PCA	0	19.53	20.31	29.69	19.53	29.69
Random	0	9.38	9.38	37.50	9.38	9.38

C.3 Matched iterative objectives

The five objectives share the initial mask, calibration quartets, and a pool of eight candidate slots in each of four rounds. Candidate proposals exchange 2, 4, 8, 16, or 32 kept/deleted channels per layer, with randomized mixtures of factor-associated scores; all 36 layers are eligible. Each pool retains the distinct current incumbents before adding proposals. The current solution is therefore available to each objective in the next round. All runs preserve the per-layer budget and exhibit nonincreasing accepted calibration loss.

KL denotes full-vocabulary next-token teacher divergence after the forced answer prefix. FRP adds the normalized contrast error of four option log-probabilities, with the response weight fixed to one. The teacher-only calibration scales for the two main effects and interaction are 13.3157, 3.3739, and 4.1025; a scale floor of 0.1 is fixed in advance. MSE instead adds squared error of the raw four-cell option scores. The no-interaction variant omits the interaction term; the diagonal variant removes cross-cell terms from the FRP quadratic. These are distinct objective controls, although they may select identical masks.

Algorithm seeds 0, 1, and 2 vary proposal randomness, not data or calibration sampling. All runs allocate 32 candidate slots. Cached distinct mask forwards number 23, 23, and 22 for the three SAE seeds, 22 for PCA, and 25 for random; identical slots do not incur a second forward. Equal maximum slot budgets must therefore not be described as identical realized forward counts. A selection manifest is frozen before generated-answer test evaluation.

C.4 Generated answers and standard benchmarks

Generated-answer accuracy requires both a correct final option and termination within 128 tokens; unfinished outputs count as incorrect. Dense accuracy is 128/128 and RMS accuracy is 41/128. RMS terminates on only 47/128 questions, so the metric reflects a substantial length/completion effect, not solely arithmetic correctness. The token audit found no missed alternate end token: all non-ended audited records reached the 128-token cap. These results are not interchangeable with earlier experiments using a longer generation cap.

For the standard suite we evaluate the full labelled splits of BoolQ, PIQA, HellaSwag, WinoGrande, ARC-easy, ARC-challenge, and OpenBookQA using zero-shot continuation likelihood. The primary metric is raw accuracy for BoolQ and WinoGrande and character-normalized accuracy for the other tasks. Perplexity uses the full WikiText-2 test split in nonoverlapping 512-token blocks, including the final partial block. Table 10 exposes the individual task scores rather than only their average.

The three SAE–FRP generated accuracies are 24.22%, 29.69%, and 15.62%, all below RMS. MSE reaches 46.09% at seed 0 but falls to 24.22% and 28.91% at the other two seeds; reporting only its first seed would conceal this instability. SAE–FRP averages 23.18% generated accuracy and 58.289% on standard tasks. Its standard-task average is close to RMS (58.160%), while all three FRP perplexities exceed the RMS value. This combination does not support a stable practical gain over the initialization.

For SAE seed 0, the FRP–RMS difference is -7.81 percentage points with a quartet-bootstrap 95% interval of $[-14.84, -0.78]$; seed 1 gives -2.34 $[-8.59, 3.12]$, and seed 2 gives -16.41 $[-25.78, -7.81]$. We resample the 32 whole quartets 10,000 times, preserving their four-cell struc-

Table 9: Every completed KL/FRP/MSE chain: generated accuracy, seven-task macro-average (%), and full WikiText-2 test perplexity. The same initial RMS mask yields 32.03%, 58.160%, and 20.7986, respectively. All 15 entries are complete; exact duplicate masks reuse identical standard evaluations, leaving 13 distinct standard-suite runs.

Dictionary	Seed	Objective	Generated ↑	Seven-task ↑	PPL ↓
SAE	0	KL	28.91	58.110	20.8911
SAE	0	FRP	24.22	58.322	20.9666
SAE	0	MSE	46.09	57.708	21.0864
SAE	1	KL	29.69	58.427	20.9799
SAE	1	FRP	29.69	58.427	20.9799
SAE	1	MSE	24.22	58.234	21.1584
SAE	2	KL	23.44	58.386	20.9064
SAE	2	FRP	15.62	58.118	20.9512
SAE	2	MSE	28.91	57.892	21.0947
PCA	0	KL	19.53	58.198	20.9867
PCA	0	FRP	20.31	58.052	20.9651
PCA	0	MSE	29.69	58.004	20.9293
Random	0	KL	9.38	58.198	20.9126
Random	0	FRP	9.38	58.198	20.9126
Random	0	MSE	37.50	58.300	20.9016

Table 10: Full standard-suite task breakdown (%). Columns use the primary metric described in the text. These are measured masked-model results, not literature values or physical-export speed measurements.

Dictionary	Seed	Objective	BoolQ	PIQA	HellaSwag	WinoG.	ARC-e	ARC-c	OBQA
SAE	0	KL	80.00	71.98	58.42	60.69	61.53	38.14	36.00
SAE	0	FRP	80.18	72.03	58.15	61.09	62.12	38.48	36.20
SAE	0	MSE	76.02	71.33	57.45	61.72	62.50	39.93	35.00
SAE	1	KL	80.03	71.60	58.17	61.88	62.21	38.91	36.20
SAE	1	FRP	80.03	71.60	58.17	61.88	62.21	38.91	36.20
SAE	1	MSE	79.91	71.71	57.97	61.25	61.70	38.91	36.20
SAE	2	KL	79.60	71.71	57.82	61.96	62.08	39.33	36.20
SAE	2	FRP	79.72	72.09	57.85	61.09	61.74	39.33	35.00
SAE	2	MSE	78.13	71.65	57.59	60.62	63.09	39.16	35.00
PCA	0	KL	80.55	72.03	58.10	60.85	61.03	38.82	36.00
PCA	0	FRP	80.12	71.38	57.96	61.01	61.70	38.40	35.80
PCA	0	MSE	80.43	71.33	58.16	60.46	61.24	38.82	35.60
Random	0	KL	79.20	71.22	57.72	62.67	60.98	39.59	36.00
Random	0	FRP	79.20	71.22	57.72	62.67	60.98	39.59	36.00
Random	0	MSE	78.81	71.33	57.99	61.80	62.42	39.16	36.60

ture. These intervals are unadjusted for multiple comparisons and do not quantify cross-model or independently resampled-calibration uncertainty.

SAE seed 0 exceeds PCA and random FRP on generated accuracy, but all three remain below RMS. One dictionary-control seed cannot establish consistent SAE necessity. Identical selections are also informative: SAE seed 1 KL and FRP coincide, and random KL, FRP, no-interaction, and diagonal coincide. We retain these equalities instead of forcing different masks.

C.5 Fixed-pool response-weight and candidate-prefix diagnostic

After the primary experiments, we perform an explicitly exploratory diagnostic. For each SAE seed, we retain the chronological unique candidates from its completed adaptive search and rerank prefixes of requested size 8, 16, and 32 using calibration KL plus response weights $\lambda \in \{0, 0.1, 0.3, 1, 3, 10\}$. The final available pools contain only 23, 23, and 22 distinct candidates, respectively. The 54 configurations select 13 distinct masks: seven reuse existing generated records and six receive new 128-question evaluations. All diagnostic choices are recorded before their additional test evaluation, but the original test outcomes had already been observed when this diagnostic was proposed.

This is *not* 54 independent multi-round searches. The pools come from the primary adaptive trajectories, so reranking a prefix neither regenerates proposals for a new weight nor isolates a causal

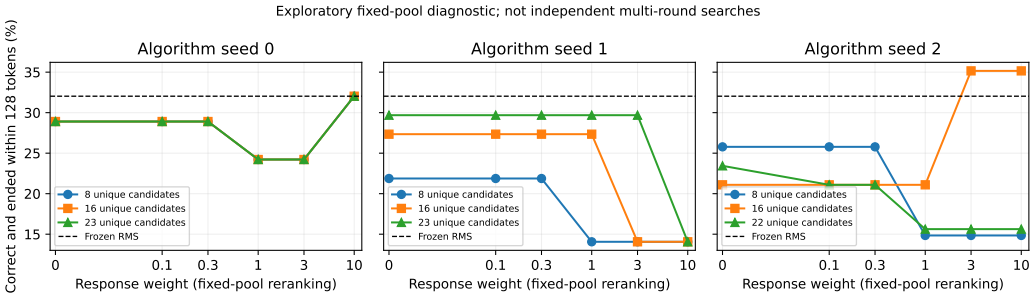

Figure 7: Exploratory fixed-pool reranking of already generated SAE candidate pools. Each panel varies the response weight for one algorithm seed; legends show actual unique prefix sizes. The dashed line is the frozen RMS mask. All points are shown, and no test-optimal point replaces the primary FRP result. The different curves are not independently rerun search solvers.

Table 11: All 54 fixed-pool diagnostic configurations: generated accuracy (%). The actual pool size is shown, not the requested upper bound of 32. Repeated scores often reflect repeated selected masks.

Seed	Pool	$\lambda = 0$	0.1	0.3	1	3	10
0	8	28.91	28.91	28.91	24.22	24.22	32.03
0	16	28.91	28.91	28.91	24.22	24.22	32.03
0	23	28.91	28.91	28.91	24.22	24.22	32.03
1	8	21.88	21.88	21.88	14.06	14.06	14.06
1	16	27.34	27.34	27.34	27.34	14.06	14.06
1	23	29.69	29.69	29.69	29.69	29.69	14.06
2	8	25.78	25.78	25.78	14.84	14.84	14.84
2	16	21.09	21.09	21.09	21.09	35.16	35.16
2	22	23.44	21.09	21.09	15.62	15.62	15.62

effect of search budget. The test set is now exposed: the best diagnostic test point cannot become a new confirmatory primary method without fresh held-out evaluation.

The diagnostic is nonmonotonic: seed 2 reaches 35.16% at a 16-candidate prefix with weights 3 or 10, but falls to 15.62% when the available pool expands to 22 candidates at the same weights. Seed 0 reaches only the RMS accuracy at weight 10, and seed 1 remains below RMS throughout. These observations show sensitivity and calibration/test mismatch rather than a validated improved operating point. They are retained as negative and exploratory evidence, with all configurations visible.

C.6 Integrity checks and scope of inference

The final audit verifies all 15 standard-suite entries are complete, their source-file hashes match the evaluated inputs, and all 36 layer-wise index arrays agree with the frozen masks. Registry identifiers hash canonical JSON, whereas evaluation configurations hash the exact source bytes; whitespace differences in a rewritten copy are checked by comparing parsed indices rather than conflating these two hash definitions. No results are rewritten to resolve a formatting-only hash difference.

The negative outcome does not by itself distinguish a poor response proxy from a poor proposal mechanism. The proposals prioritize dictionary-associated scores without an additional RMS guard or a general-language preservation constraint. KL is measured at a forced answer position, whereas deployment requires free generation. Both are plausible failure mechanisms, not proven causal explanations. The completed experiments establish an auditable factor-defined dictionary-to-mask pipeline and reveal its current limitations; they do not establish that SAE-guided pruning or FRP is effective.

D EXTENDED EVALUATION, RECOVERY AND DEVELOPMENT CONTROLS

The tables below use the masks and calibration described in the main paper. Full rotation always uses 2,560 output dimensions; the rank-64 projected control is a different procedure. Recovery variants were retained irrespective of test outcome. The correction-location offsets use dense-minus-uncorrected-masked outputs and include upstream drift, unlike the local mean correction used in the main comparison.

Table 12: Measured three-budget comparison, frozen calibration seed zero, no recovery. Actual MLP removal is shown. All seven full tasks and complete WikiText test are evaluated. These are the same masks as the budget curves. Best pruned value at each budget is bold (accuracy ↑, PPL ↓).

Method	MLP%	BQ	PIQA	HS	WG	ARC-e	ARC-c	OBQA	Mean	PPL
Dense	0.00	85.02	74.86	68.46	65.90	78.37	53.75	40.00	66.62	18.42
Matched RMS	10.53	**82.94**	73.50	**64.20**	**64.01**	66.92	43.94	36.40	61.70	19.01
PCA joint control	10.53	82.66	**74.27**	64.08	63.85	**68.14**	**44.71**	36.40	**62.02**	**18.69**
SAE joint (ours)	10.53	82.20	73.72	63.81	62.83	68.01	44.45	**36.80**	61.69	18.77
Matched RMS	20.39	**80.31**	**71.65**	58.20	61.40	**61.11**	38.65	35.80	58.16	20.80
PCA joint control	20.39	80.18	71.22	58.53	61.09	60.35	38.31	35.00	57.81	20.35
SAE joint (ours)	20.39	80.09	71.22	**58.72**	**62.51**	60.40	**38.91**	**36.40**	**58.32**	**20.29**
Matched RMS	35.53	62.17	63.66	47.90	57.54	45.50	31.23	**32.40**	48.63	25.92
PCA joint control	35.53	63.12	64.58	48.47	**58.09**	44.82	31.23	31.40	48.82	25.21
SAE joint (ours)	35.53	**63.98**	**64.69**	**48.58**	56.20	**45.62**	**31.48**	31.40	**48.85**	**24.85**

Table 13: Fixed-mask recovery controls at 35.53% MLP removal. Each cell is seven-task accuracy (%)/PPL. Fitting activations and evaluation data are shared. The projected rotation here is a rank-64 local control, distinct from full RCPU rotation. Within each recovery column, best accuracy (↑) and PPL (↓) are bold separately.

Mask	None	Mean	Rank8 ridge	Projected rotation
RMS	48.63/25.92	49.33/**26.07**	49.27/**25.66**	48.86/26.02
PCA joint	48.82/25.21	49.39/26.47	49.64/25.87	**48.93**/25.28
SAE joint (ours)	**48.85/24.85**	**49.61**/26.23	**49.84**/25.78	48.76/**24.89**

Table 14: Same full 2,560-dimensional rotation recovery, same original calibration activations and fixed masks as Table 1. The additional scale is one isotropic factor per layer. All six settings are reported; no test-based variant selection. Deltas are relative to the RCPU mask with the same recovery, not to the unrecovered model.

Mask selector	Recovery	Mean (%)	PPL	Δ mean	Δ PPL
RCPU score†	Full rotation	37.19	67.22	+0.00	+0.00
PCA joint control	Full rotation	37.31	65.80	+0.12	-1.42
SAE joint (ours)	Full rotation	38.13	65.83	+0.94	-1.39
RCPU score†	Rotation + scale	37.17	66.61	+0.00	+0.00
PCA joint control	Rotation + scale	37.10	65.31	-0.07	-1.30
SAE joint (ours)	Rotation + scale	38.26	65.68	+1.09	-0.94

D.1 PREDICTING THE LOCALLY REMOVABLE MEAN COMPONENT

We diagnose the five frozen seed-zero masks from Table 1 at 35.53% MLP-channel removal, retaining all 36 layers rather than selecting favorable examples. The calculation uses full-dimensional down-projection outputs in FP32, not the rank-64 projection used in joint selection. The original 51 WikiText fitting windows supply the correction; the original 13 diagnostic windows evaluate it. These diagnostic windows have been inspected previously, so this is a retrospective mechanism

analysis rather than independent task validation. Masks, window indices and numerical precision are fixed; no new hyperparameter is selected.

For each method and layer, let $r(x)$ denote the dense-minus-pruned output residual at fixed dense MLP inputs and let $c = \widehat{\mathbb{E}}_{\text{fit}} r(x)$. The fit-only predictor and diagnostic outcome are

$$q_{\text{fit}} = \frac{\|c\|_2^2}{\widehat{\mathbb{E}}_{\text{fit}} \|r\|_2^2}, \qquad q_{\text{diag}} = \frac{\widehat{\mathbb{E}}_{\text{diag}} \|r\|_2^2 - \widehat{\mathbb{E}}_{\text{diag}} \|r - c\|_2^2}{\widehat{\mathbb{E}}_{\text{diag}} \|r\|_2^2}. \tag{4}$$

The first quantity is the fraction of fitting residual energy removable by its sample mean. The second measures removal on different windows using exactly that fitted constant. The identity $\widehat{\mathbb{E}}\|r\|^2 - \widehat{\mathbb{E}}\|r - c\|^2 = 2c^\top \widehat{\mathbb{E}} r - \|c\|^2$ was checked for every setting; it does not imply that $q_{\text{fit}} = q_{\text{diag}}$.

Table 15: Completed fixed-mask local recovery diagnostic. Fractions are medians across 36 layers (%); Spearman correlation is within each method across layers. Bold highlights the largest fraction in each of the two fraction columns, not a claim of best overall pruning quality. All 180 layer–method settings improve diagnostic MSE.

Mask	Fit fraction	Diagnostic fraction	Spearman	Improved layers
RMS	4.88	3.97	0.983	36/36
SAE joint	5.03	4.39	0.985	36/36
PCA joint	5.05	4.36	0.990	36/36
FLAP score	**9.00**	**7.99**	0.992	36/36
RCPU score	5.32	4.17	0.992	36/36

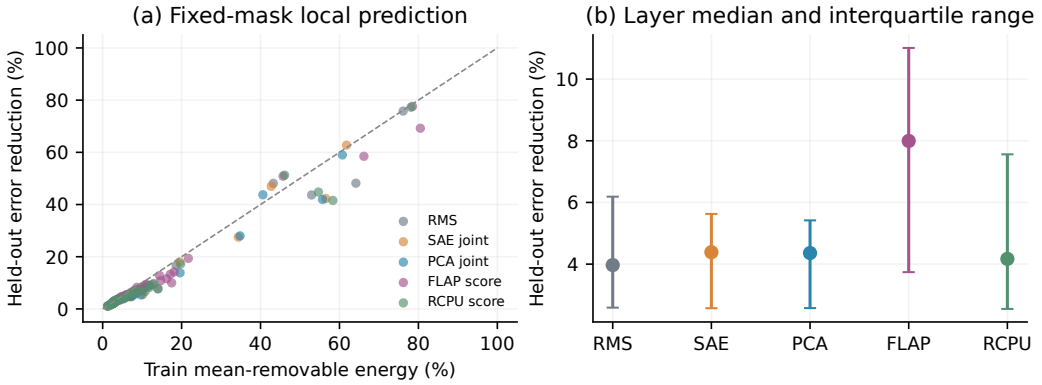

Figure 8: Fit-only removable mean energy versus measured diagnostic error removal for frozen masks. The 36 points per method are layers of one backbone, not independent model replications. Any displayed quartile range summarizes layer heterogeneity and is not a confidence interval.

Fit-only fractions track diagnostic fractions across layers (within-method Spearman 0.983–0.992), and the fitted correction lowers diagnostic error in all 180 settings. FLAP's higher removable fraction is consistent with its larger task-accuracy gain under mean recovery, but does not establish a causal explanation for that gain. Only five masks from one calibration seed enter this comparison. The local residual is evaluated at fixed dense inputs; accumulated upstream drift and nonlinear downstream decisions are absent. Consequently, this experiment characterizes local recoverability and cannot establish prediction of full-model task rankings.

D.2 Fixed-candidate deletion-damage ranking

All candidates delete exactly 64 intermediate channels at one of layers 6, 18 and 30 in Qwen3-4B. Predictions use the original 51 fitting windows and the shared rank-64 down-output projection. SAE/PCA joint risks use the reconstructed second moment plus diagonal residual from Equation 2;

empirical joint risk uses the actual training second moment under the same projection. RMS, population variance times full down-column norm, and the RCPU-score adaptation are additional predictors. The variance diagnostic is not the public FLAP accumulator used in the full-mask experiments.

For the coarse population, seed-zero training scores choose the lowest/highest 64 channels for each of the six methods; exact groups are deduplicated within layer, leaving 33 groups. The near-boundary population restricts channels to ± 486 ranks around the RMS 35.53% deletion boundary. From 128 fixed-seed random groups of 64 channels per layer, six have the largest absolute SAE/PCA risk-rank disagreements and six are sampled randomly; duplicates are removed, leaving 36 groups. No diagnostic outcome is used to construct either population. These populations answer different questions and are not pooled. The disagreement population is selected to probe representational differences, not to estimate average performance over all channel sets.

Local outcome is actual projected group-deletion energy on each calibration seed's 13 unfitted windows; these are historical diagnostics, not a newly untouched corpus. Actual end-to-end damage is measured with the same frozen groups on 16 new 256-token WikiText training windows. The token stream is reconstructed from the original parquet artifact, all three original calibration-window token hashes are verified, and every calibration block is excluded. Each window contributes 255 next-token targets. Full-vocabulary cross entropy is evaluated in float32 using the BF16 model. The intervention is a single-layer group deletion in a dense background; conclusions do not establish joint multi-layer damage prediction or the quality of a complete function-risk mask.

Table 16: Risk-ranking diagnostic, mean within-layer/seed Spearman. Local error uses historical unfitted windows; ΔNLL uses newly sampled training windows. Near-boundary intervals resample 16 windows jointly 1,000 times, keeping the candidates, layers and three fitted predictors fixed.

Predictor	Tail local	Tail ΔNLL	Boundary local	Boundary ΔNLL [95% interval]
SAE joint	.953	.774	.102	.019 $[-.150, .152]$
PCA joint	.941	.776	.033	$-.044$ $[-.161, .086]$
Empirical joint	.946	.775	.071	.019 $[-.128, .154]$
RMS	.685	.715	$-.002$	.050 $[-.117, .187]$
Variance-only	.631	.594	.095	.073 $[-.049, .214]$
RCPU score	.659	.678	$-.158$	.009 $[-.103, .167]$

The three calibration seeds predict the same measured outcomes; they are not three independent behavioral samples. Negative average ΔNLL is retained for 14/33 tail groups and 19/36 boundary groups. The paired SAE–PCA correlation difference is $-.002$ with interval $[-.017, .009]$ on broad tails and .063 with interval $[-.030, .108]$ near the boundary. No interval is a confidence statement over other models or candidate-selection procedures. Even measured local projected error correlates with end-to-end damage only at .763 on tails and .078 near the boundary (mean within layer/seed).

A separate input-distribution diagnostic transfers the frozen tail groups to eight 256-token BoolQ training passages without questions or labels, excluding passage hashes shared with validation. SAE/PCA/empirical behavioral correlations are .821/.820/.823. This is another coarse sanity check, not a second fitted calibration distribution or second backbone. The first WikiText measurement used a different historical token stream and could not establish exclusion; it is retained as superseded and excluded from every reported number above.

Raw candidate manifests, token hashes, per-window losses, prediction scores and stratified correlations are retained under `experiments/review_revision_20260925/mechanism`. The finite sample, narrow risk range, numerical precision and dependence among overlapping channel groups remain sensitivity limits. Independent semantic interventions and end-to-end functional-mask improvement remain untested.

Numerical sensitivity. The same 36 threshold-neighborhood candidates and 16 windows were rerun with the full model and cross entropy in float32, with TF32 disabled. Mean within-layer/seed risk correlations are SAE .020, PCA $-.047$, empirical joint .016, RMS .097, variance-only .146 and RCPU score .058. BF16–float32 damage-rank correlations are .860/.818/.531 at layers 6/18/30, so individual rankings are precision-sensitive. The maximum absolute change in a candidate's mean ΔNLL is .00137. The unchanged low SAE/PCA point correlations do not eliminate sampling or

range-restriction explanations. The intervals in the preceding table remain those of the original BF16 experiment; they are not reused as float32 confidence intervals.

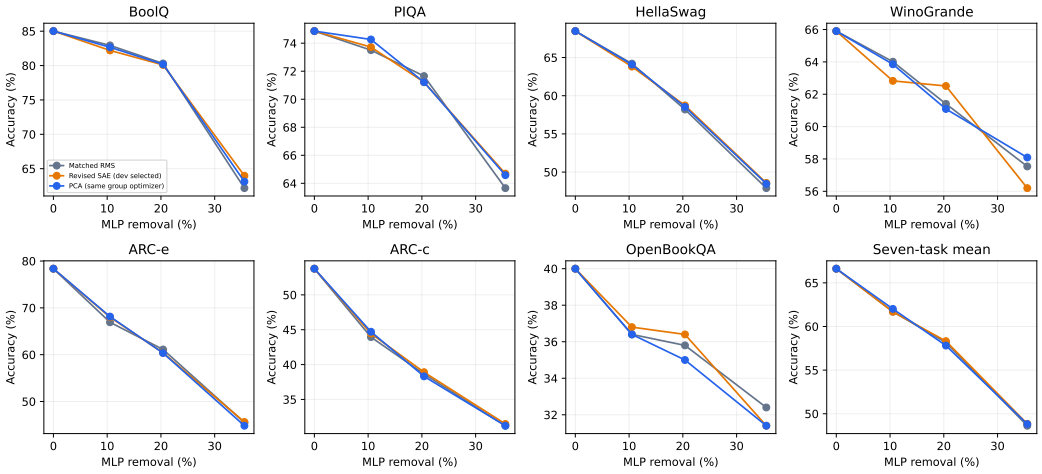

Figure 9: All seven task-level curves for the frozen three-budget comparison. Declines and flat results are retained.

D.3 Regularization and correction location

Table 17: Completed SAE fixed-mask recovery variants at 35.53% MLP removal. All use full evaluation splits. Ridge varies regularization of the rank-eight correction; site offsets are a different intervention. No test-best variant replaces the main comparison.

Recovery	Seven-task mean (%)	PPL
Ridge 10^{-4}	49.66	25.63
Ridge 10^{-2}	49.84	25.78
Ridge 1	49.71	25.98
Attention-site offset	48.56	45.10
MLP-site offset	49.41	32.51
Both-site offsets	49.37	67.03

The attention/MLP/both offsets give PPL 45.10/32.51/67.03 despite relatively modest task-mean changes. Upstream drift makes these offsets distinct from the dense-input local mean. The unrestricted full rotation in Table 14 also differs from rank-64 projected rotation in Table 13.

D.4 Development-only sensitivity

The 31 completed sensitivity evaluations vary fitting windows 8–512, dictionary size 64–512 and swap limits 16/64/256 (Figure 10). At $N = 64, F = 512$, increasing SAE search from 64 to 256 swaps raises development accuracy from 51.56 to 52.01%, but worsens validation PPL from 26.22 to 26.47. At $N = 512, F = 64$, SAE and PCA reach 54.80/54.35% with PPL 26.62/26.71. These are development observations, not additional final-test confirmation or proof that 500 steps converge at every sample size.

D.5 Physical export, precision and evidence scope

The largest-budget RMS/SAE/PCA checkpoints each contain 3,066,953,216 parameters versus 4,022,468,096 dense parameters. Weight files occupy about 6.134 GB versus 8.045 GB, counting weights rather than tokenizer/config files. Removing matching gate/up rows and down columns realizes the static mask without running a teacher or dictionary. Table 18 separates masked evaluation

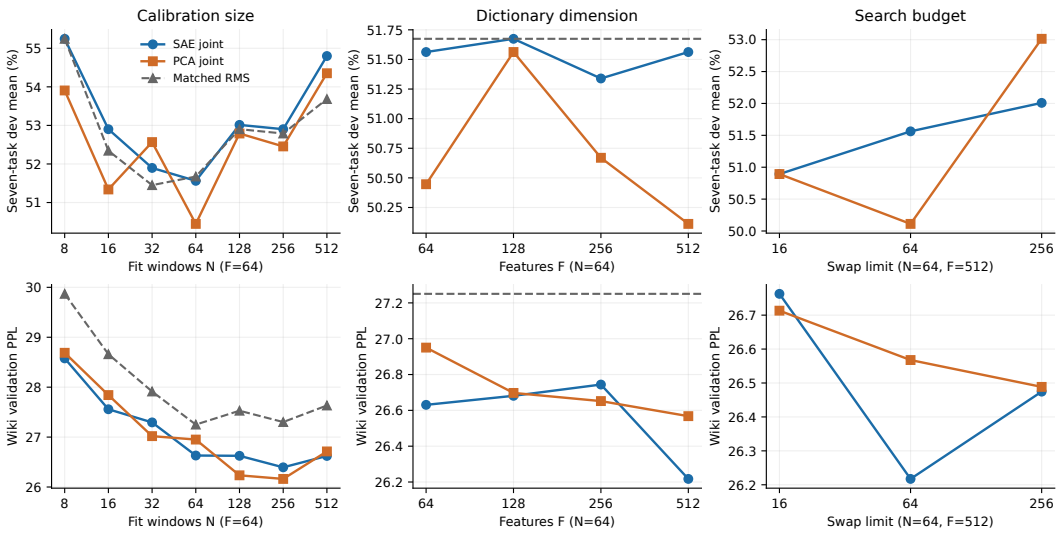

Figure 10: Completed **development-only** sensitivity study: seven task training subsets of 128 items each and full WikiText validation. Nested fit windows exclude the same 13 diagnostic windows. TopK8 is used throughout, unlike TopK32 in the frozen main configuration; SAEs train for 500 steps and PCA uses a fixed randomized fit. More fitting data, features or swaps need not improve both metrics. No point is selected using final test performance.

from actual saved-and-reloaded models. The exported FRP model has 3,473,931,776 parameters

Table 18: Complete evaluations before and after physical export. Each cell is seven-task mean (%)/PPL. The FRP row uses seed zero at the smaller 20.39% MLP-removal budget; the other rows use 35.53%.

Mask	Masked	Saved and reloaded
35.53% MLP removal		
RMS	48.627/25.921	48.725/25.921
PCA joint	48.815/25.211	48.825/25.217
SAE joint	48.849/24.853	48.814/24.864
Separate FRP branch: 20.39% MLP removal		
SAE–FRP seed zero	58.322/20.967	58.188/20.971

(13.64% whole-model removal), illustrating why requested MLP and total-parameter reductions cannot be equated. Its FP32 export diagnostic has maximum logit error approximately 1.04×10^{-4} with identical tested top-one predictions, whereas BF16 full-suite predictions can change near decision boundaries. For RMS/SAE/PCA, relative full PPL differences stay below 0.05%. These are tolerance checks rather than bitwise equivalence claims. Tensor and file-size reduction do not establish throughput; noisy shared-host timings are not treated as controlled deployment evidence.

D.6 Artifact and audit inventory

Each complete suite retains the configuration, split and checkpoint identifiers, input mask, per-choice likelihoods, correctness indicators and WikiText block losses. Reported seven-task means are unweighted means of the specified seven task metrics. Full-suite counts total 20,465 items, and full PPL uses 298,493 predicted tokens. Independent reconstruction checks verify these aggregates. Mask provenance checks compare the recorded input-file byte hash and exact retained indices; JSON formatting differences in copied masks are recorded rather than mistaken for changed selections. The completed chain has fifteen objective/seed rows but thirteen unique masks, with duplicates explicitly reused. Source mappings for the extended tables and proof checks accompany the source archive.

E LOCAL ALGEBRAIC PROPERTIES AND LIMITS

This appendix proves properties of fixed-input, single-layer proxies. It does not prove semantic interpretability, generalization of a fitted dictionary, optimal end-to-end pruning, or superiority of SAE features. Expectations below are either population expectations with finite second moments or uniform averages over a fixed calibration set; the identities apply to both, but empirical identities are not population guarantees.

E.1 RECONSTRUCTED SECOND MOMENTS AND THE RESIDUAL APPROXIMATION

Let $x \in \mathbb{R}^C$ be a column of intermediate activations, $W \in \mathbb{R}^{d\times C}$ the output projection, and $Q \in \mathbb{R}^{d\times r}$ a fixed matrix with orthonormal columns. Write $U = Q^\top W = [u_1, \dots, u_C]$. A retained set M has deletion indicator $a_j = \mathbf{1}\{j \notin M\}$ and $D_a = \mathrm{diag}(a)$. These results condition on W, Q, M and the fitted reconstruction map; they make no independence assumption between x and its reconstruction.

Absorb the channelwise normalization into a raw decoder $D_{\mathrm{raw}} \in \mathbb{R}^{F\times C}$, so that $\hat{x} = \mu + D_{\mathrm{raw}}^\top z$. Define

$$h = \begin{bmatrix} 1 \\ z \end{bmatrix}, \qquad A = \begin{bmatrix} \mu^\top \\ D_{\mathrm{raw}} \end{bmatrix}, \qquad H = \mathbb{E}[hh^\top], \qquad B = H^{1/2}A. \tag{5}$$

The symmetric positive-semidefinite square root exists even if H is singular. In particular, $B \in \mathbb{R}^{(F+1)\times C}$ and $B^\top B = \mathbb{E}[\hat{x}\hat{x}^\top]$. The leading row and column of H retain the mean and mean–latent cross terms.

Proposition 1 (exact reconstructed deletion energy). With $b_j = B_{:j}$,

$$A_a := \mathbb{E}\|UD_a\hat{x}\|_2^2 = \left\| \sum_j a_j b_j u_j^\top \right\|_F^2 = a^\top \left[(B^\top B) \odot (U^\top U) \right] a. \tag{6}$$

Proof. Since $\hat{x} = A^\top h$, its second moment is $A^\top H A = B^\top B$. Expanding the squared norm gives $\sum_{j,k} a_j a_k \mathbb{E}[\hat{x}_j \hat{x}_k] u_j^\top u_k$. Expanding the Frobenius norm gives the same sum because $\langle b_j u_j^\top, b_k u_k^\top \rangle_F = (b_j^\top b_k)(u_j^\top u_k)$. This also proves the final equality. No diagonal approximation is used here. □

Let $e = x - \hat{x}$, $v_j = \mathbb{E}[e_j^2]$, and $V_a = \sum_j a_j v_j \|u_j\|_2^2$. The implemented proxy is $J_a = A_a + V_a$, whereas the true fixed-input projected deletion energy is $L_a = \mathbb{E}\|UD_a x\|_2^2$.

Proposition 2 (explicit omitted terms and a bound). Put $R_a = \mathbb{E}\|UD_a e\|_2^2$. Then

$$L_a - J_a = 2\mathbb{E}\langle UD_a\hat{x}, UD_a e\rangle + \sum_{j\neq k} a_j a_k \mathbb{E}[e_j e_k] u_j^\top u_k, \tag{7}$$

$$|L_a - J_a| \le 2\sqrt{A_a R_a} + \sum_{j\neq k} a_j a_k |u_j^\top u_k| \sqrt{v_j v_k}, \tag{8}$$

$$R_a \le \left(\sum_j a_j \|u_j\|_2 \sqrt{v_j} \right)^2. \tag{9}$$

Proof. Substitute $x = \hat{x} + e$ into L_a. In the residual quadratic term, the diagonal is exactly V_a, since $a_j^2 = a_j$. The remaining terms yield Eq. 7. Cauchy–Schwarz in the space of square-integrable random vectors bounds the first term by $2\sqrt{A_a R_a}$. Scalar Cauchy–Schwarz gives $|\mathbb{E}[e_j e_k]| \le \sqrt{v_j v_k}$. Finally, the triangle inequality in that same square-integrable space bounds the norm of $\sum_j a_j u_j e_j$ by $\sum_j a_j \|u_j\| \sqrt{v_j}$. □

These bounds can be loose: small individual residual variances alone need not control the sum over many correlated deleted channels. If both omitted terms vanish, $J_a = L_a$; an SAE training objective does not impose that condition. Nor does a projected energy control all output directions: $\|Q^\top y\| \le \|y\|$, with potentially nonzero error in the orthogonal complement. Equation 6 depends

on the dictionary only through its reconstructed second moment; any other estimator with the same moment yields the same proxy.

E.2 Equal-cost swaps and the scope of descent

Set $T_j = b_j u_j^\top$, $c_j = v_j \|u_j\|^2$, and $S = \sum_{j \notin M} T_j$. Consider restoring deleted channel $i \notin M$ and deleting retained channel $k \in M$, giving $M' = (M \cup \{i\}) \setminus \{k\}$.

Proposition 3 (exact swap change). For $J(M) = \|S\|_F^2 + \sum_{j \notin M} c_j$,

$$\begin{aligned} J(M') - J(M) = & -2\langle S, T_i \rangle_F + \|T_i\|_F^2 - c_i + 2\langle S, T_k \rangle_F + \|T_k\|_F^2 + c_k \\ & - 2(b_i^\top b_k)(u_i^\top u_k). \end{aligned} \tag{10}$$

Proof. The new sum is $S - T_i + T_k$. Expand its squared norm and subtract $\|S\|_F^2$; the diagonal residual sum changes by $-c_i + c_k$. Use the outer-product inner-product identity from Proposition 1. □

Every such swap preserves $|M|$. It also preserves structural cost when all channels considered have equal cost, as in a fixed-width layer; it need not do so across unequal-cost groups. In exact arithmetic, accepting only strictly negative changes makes the proxy strictly decrease. There are only $\binom{C}{|M|}$ possible retained sets, so a sequence of strict improvements cannot revisit a set and has at most $\binom{C}{|M|} - 1$ accepted swaps. A prescribed iteration cap gives a stronger operational limit.

This is a termination statement, not global optimality. With exhaustive one-swap search and no iteration cap, stopping because no strict improvement exists establishes only one-swap local optimality. The implementation instead screens candidates, applies a numerical tolerance, and caps accepted iterations; its stopping rule does not establish even exhaustive one-swap local optimality. Floating-point updates are checked for a decrease, and independently recomputed final objectives are required to detect numerical drift. The exact-arithmetic proof is not a bound on that drift.

E.3 Optimal constant mean correction for a fixed mask

Proposition 4 (fixed-input least-squares constant). Let $y = Wx$ and $y_M = WD_M x$, where D_M retains M. Among deterministic vectors $c \in \mathbb{R}^d$, the unique minimizer of $\mathbb{E}\|y - y_M - c\|^2$ is

$$c^* = \mathbb{E}[y - y_M] = W(I - D_M)\mathbb{E}[x]. \tag{11}$$

Proof. Put $\delta = y - y_M$. Expanding around $\mathbb{E}[\delta]$ and using its centered mean gives

$$\mathbb{E}\|\delta - c\|^2 = \mathbb{E}\|\delta - \mathbb{E}\delta\|^2 + \|c - \mathbb{E}\delta\|^2. \tag{12}$$

The second term is uniquely minimized at zero. Linearity of expectation gives Eq. 11. □

For a finite calibration set, this proves optimality of its sample mean on that set. It does not prove population optimality of the sample estimate. More importantly, the result fixes the local input distribution and mask: upstream pruning changes those inputs, and nonlinear downstream computation can amplify or reverse the relationship between local squared error and task quality. Thus mean correction can improve this local objective while harming PPL or accuracy. An end-to-end dense-minus-masked offset is a different fitting target from the local fixed-input formula.

E.4 Full-output Procrustes and unconstrained directions

To avoid confusing output rotation with the projection Q, denote the former by $O \in \mathbb{R}^{d \times d}$. Let $Y, Z \in \mathbb{R}^{d \times n}$ contain dense and fixed-mask outputs on the same calibration inputs. Consider $\min_{O^\top O = I} \|Y - OZ\|_F^2$, with no determinant constraint.

Proposition 5 (rotation fit and rank-deficient nonidentifiability). Let $YZ^\top = L\Sigma V^\top$ be a full singular-value decomposition with rank q. An optimum is $O = LV^\top$. More generally, every matrix

$$O_K = L \begin{bmatrix} I_q & 0 \\ 0 & K \end{bmatrix} V^\top, \qquad K \in \mathbb{R}^{(d-q)\times(d-q)}, \quad K^\top K = I, \tag{13}$$

attains the same minimum. If $q < d$, the calibration objective therefore cannot identify a unique rotation.

Proof. Orthogonality makes the objective equal to $\|Y\|_F^2 + \|Z\|_F^2 - 2\,\mathrm{tr}(O^\top Y Z^\top)$. For $G = L^\top O V$, G is orthogonal, and the trace is $\sum_{j=1}^{q} \Sigma_{jj} G_{jj} \le \sum_{j=1}^{q} \Sigma_{jj}$ because $G_{jj} \le 1$. The displayed family attains equality. Distinct choices of K give distinct rotations when $q < d$. □

There is a stronger prediction-level statement when $\mathrm{rank}(Z) < d$. Let $\mathcal{S} = \mathrm{span}(Z)$ and let T be any orthogonal map that fixes $\mathcal{S}$ pointwise. Then $(OT)Z = OZ$ for any fitted rotation O, so the calibration predictions are identical. For a new output $z = s + t$, with $s \in \mathcal{S}$ and $0 \neq t \in \mathcal{S}^\perp$, choose T to reflect the direction t and fix $\mathcal{S}$. The two predictions differ by $2Ot$, whose norm is $2\|t\|$. Hence agreement on calibration outputs alone cannot control behavior in unseen directions. In particular, $n < d$ implies $\mathrm{rank}(Z) < d$. This construction explains a possible failure mechanism, not an identified cause of the measured recovery degradation.

For fixed optimal O and a nonnegative scalar multiplier α, if $\|Z\|_F > 0$, minimizing $\|Y - \alpha O Z\|_F^2$ yields

$$\alpha^* = \frac{\mathrm{tr}(\Sigma)}{\|Z\|_F^2}. \tag{14}$$

This follows by completing the square in α; its numerator is nonnegative. If $Z = 0$, the objective is independent of both scale and rotation and the ratio is undefined. With column-vector activations, rotated retained weights are $OW_{:M}$; with row-oriented batches this produces $Z_{\mathrm{rows}} O^\top$. These are the same operation, not different recovery conventions.

E.5 A small reconstruction error need not preserve decisions

Take deterministic $x = (\varepsilon, 1)^\top$, with $0 < \varepsilon < 1$, and retain exactly one channel. Under unweighted activation deletion error, deleting channel 1 costs ε^2, while deleting channel 2 costs 1; the lower-error mask deletes channel 1. Let the downstream scalar output be $Wx = x_1$, with $W = (1, 0)$, and let the correct class be selected precisely when this output exceeds $\varepsilon/2$. The dense model and the mask deleting channel 2 are correct, whereas the lower-error mask deleting channel 1 is wrong. As $\varepsilon \to 0$, both its activation error and squared output error tend to zero, yet its accuracy loss is always one.

The example varies the decision margin with ε: it does not contradict a stability guarantee under an assumed positive margin and a uniformly smaller output perturbation. It shows why a reconstruction guarantee alone, without such decision assumptions, does not imply preserved task accuracy. It is not a model of the observed benchmark distribution, and it gives no preference between SAE, PCA, or any other estimator.

E.6 Proxy error and decision limits in the executed selector

The reconstructed covariance specifies which interactions the selector retains. Write $U = Q^\top W$, $e = x - \hat{x}$, and let D_a be the diagonal deletion mask. For a fixed layer and input distribution, define $A_a = \mathbb{E}\|U D_a \hat{x}\|^2$ and $R_a = \mathbb{E}\|U D_a e\|^2$. The implemented residual penalty is $V_a = \sum_j a_j \mathbb{E}[e_j^2]\|u_j\|^2$, so $J_a = A_a + V_a$. The actual projected deletion energy $L_a = \mathbb{E}\|U D_a x\|^2$ satisfies

$$L_a - J_a = 2\mathbb{E}\langle U D_a \hat{x}, U D_a e\rangle + \sum_{j \neq k} a_j a_k \mathbb{E}[e_j e_k] u_j^\top u_k. \tag{15}$$

By Cauchy–Schwarz, a deterministic upper bound is

$$|L_a - J_a| \le 2\sqrt{A_a R_a} + \sum_{j \neq k} a_j a_k |u_j^\top u_k| \sqrt{\mathbb{E}[e_j^2]\mathbb{E}[e_k^2]}. \tag{16}$$

Thus accurate coordinate-wise reconstruction alone does not justify dropping residual correlations or reconstruction–residual cross terms. If two candidate proxies differ by less than their combined approximation bounds, the proxy ordering does not certify the true local-energy ordering. We do not claim to have certified the selected experimental masks using this bound. Moreover, even an exact local ordering would concern a projected single-layer error, not a nonlinear end-to-end task loss.

The same distinction applies to recovery. A constant sample-mean correction uniquely minimizes squared local residual error on its fitting inputs, yet it need not optimize prediction quality after upstream and downstream changes. Full rotation can fit a low-rank calibration span while leaving its orthogonal complement undetermined. Appendix E proves these claims, including a counterexample in which vanishing reconstruction error still changes a classification decision as its margin shrinks. These limited guarantees explain which inferences the experiments must test; they do not supply a theorem of SAE superiority.

F Prospective functional feature-to-deletion validation protocol

F.1 The decision problem and its evidence boundary

Let T be a frozen model, $m = (m_1, \ldots, m_L)$ a binary MLP-channel mask, and T_m its masked network. For layer l, channel activations $h_l(x) \in \mathbb{R}^{C_l}$ are inputs to the down projection $W_{d,l}$. Removing channel j deletes the matching rows of the gate and up projections and the matching column of the down projection. With residual width d_l, this removes $3d_l$ parameters. We use the actual whole-model budget

$$P(m) = P_{\text{fixed}} + \sum_{l=1}^{L} 3d_l \|m_l\|_0 \le B, \qquad \mathcal{M}(B, \mathbf{k}) = \{m : P(m) \le B,\ \|m_l\|_0 = k_l\}. \quad (17)$$

This count applies to the bias-free Qwen3 gated MLP before recovery. The first comparisons fix $\mathbf{k}$ to isolate channel identities; varying layer allocation is a separate intervention. Attention, embeddings and other retained weights belong to P_{fixed}; any new recovery parameters must also be counted. A percentage of MLP channels is not the same percentage of total parameters.

The functional hypothesis is that a risk estimated through validated SAE features predicts deletion damage beyond matched statistics. Sections F.2–F.4 specify how to test it. The executed selector in Section 3.2 instead uses reconstructed second moments, without validated functional weights. Keeping these objects separate prevents a reconstruction result from being relabeled as semantic evidence.

F.2 Discover a candidate, then validate its behavioral relevance

A frozen encoder produces $z_l = E_l((h_l - \mu_l)/\sigma_l)$ and a decoder reconstructs $\hat{h}_l = \mu_l + \sigma_l \odot D_l z_l$. Discovery data nominate candidate coordinates; a distinct validation split tests whether their activation tracks a prespecified behavior across new examples and input controls. In a natural-input quartet $q = (x_{00}, x_{10}, x_{01}, x_{11})$, two editable factors vary while the underlying item and answer convention are held fixed. Teacher capability is checked on all cells. Length, lexical changes, difficulty and answer-position effects require controls because a feature can track these instead of the intended function.

Activation selectivity is only a screening criterion. A stronger validation intervenes on a nominated coordinate while retaining the original reconstruction residual. If $d_{l,k}$ is decoder column k, the intervention is

$$h'_l = h_l + (\sigma_l \odot d_{l,k})(z'_{l,k} - z_{l,k}). \quad (18)$$

The original forward pass continues from h'_l. Prespecified ablation or a matched donor value is compared with norm-matched random/PCA directions, unrelated-feature controls and sham interventions. Donors and positions must be aligned without using test outcomes. Dose response, targeted behavior, collateral damage and matched-patch restoration are recorded. The patch preserves the reconstruction residual, but re-encoding need not leave other features unchanged. Decoder overlap and off-distribution perturbations prevent interpreting it as proof of a uniquely localized natural mechanism. A failed feature receives no functional importance weight.

Discovery, feature validation, mask calibration, hyperparameter selection and final evaluation have distinct roles. Candidate counts, intervention doses, outcome metrics and multiple-comparison correction are specified before feature confirmation. Confirmation and weight estimation use separate folds: otherwise selecting the largest measured effects and reusing them as weights introduces

winner's-curse bias. Scales and weights are frozen before mask fitting. All representations receive the same examples and behavior labels, and the allowed teacher/intervention queries are counted. A zero-validation-feature outcome is a result, not a reason to search the test set for a more favorable definition.

F.3 Measure joint deletion risk in a many-to-many map

For an actual mask, let $h_{l,m}(x)$ denote the intermediate activation obtained by running T_m up to layer l, and let $\widetilde{h}_{l,m} = m_l \odot h_{l,m}$. Re-encoding with the *same* dense-model normalization gives $z_{l,m} = E_l((\widetilde{h}_{l,m} - \mu_l)/\sigma_l)$. It reflects both the current deletion and upstream changes. Replacing $h_{l,m}$ with cached dense activations defines a cheaper local proxy; it must be labeled and separately checked.

Let $\mathcal{F}_l$ be the independently validated features, $w_{l,k} \geq 0$ their frozen validation weights, and $s_{l,k}$ their validation-set scale. A candidate risk is

$$F(m) = \sum_l \sum_{k \in \mathcal{F}_l} w_{l,k}\, \mathbb{E}_{x \in D_{\text{cal}}} \frac{(z_{l,m,k}(x) - z_{l,T,k}(x))^2}{s_{l,k}^2 + \epsilon^2}. \tag{19}$$

This measures the effect of a *whole mask*; it is not a sum of independent neuron importances. For a currently retained group G, the incremental risk is $F(m \setminus G) - F(m)$, which depends on the other retained groups. Nonlinear encoding and shared feature dependencies can make it non-additive or even negative. A single channel may affect many validated coordinates, while no individual channel need dominate a distributed feature. Decoder energy alone does not capture these interactions.

Feature preservation is itself a proxy. A general prediction-matching term can anchor behavior not represented by $\mathcal{F}_l$:

$$\min_{m \in \mathcal{M}(B,\mathbf{k})} \mathbb{E}_{x \in D_{\text{cal}}} \mathrm{KL}(p_T(\cdot \mid x) \| p_m(\cdot \mid x)) + \eta F(m), \qquad \eta \geq 0. \tag{20}$$

Prediction positions are prespecified, scales come from feature-validation data, and η is chosen using development data. Equal-cost keep/drop swaps preserve $\mathbf{k}$; candidate pools, initialization and evaluation allowances are matched. An advantage over $\eta = 0$ is evidence for the added risk only when controls receive the same calibration information. Dictionary fitting, feature validation, candidate forwards and search time all count toward construction cost. Equation 20 is the functional-risk proposal, not the executed reconstruction objective in Section 3.2.

F.4 The decisive test: rank actual damaging deletions

Before a larger method claim, the risk must pass a decision-level test at one frozen Qwen3-4B budget. Starting from the same mask and per-layer widths, construct equal-cost alternative deletions from a common candidate pool. Freeze both the statistical and feature-risk rankings on calibration data. Select disagreements symmetrically: statistically cheap but feature-risky groups, and statistically risky but feature-cheap groups. Include agreement and random groups to avoid restricting conclusions to selected extremes. Alternatives are evaluated with explicit channel masks and recovery disabled; finalists are physically exported separately.

Independent examples measure the paired damage

$$D_g(m) = \mathbb{E}_{x \in g,\, T(x) \text{ correct}}[\ell(T_m, x) - \ell(T, x)], \tag{21}$$

for each prespecified behavior g, together with unconditional loss, completion rate and PPL. Teacher-correct conditioning isolates loss of demonstrated capability; unconditional results reveal its selection effect. The same item under two masks controls question difficulty. Accuracy, answer likelihood and termination are reported separately when a finite generation budget can censor answers.

The primary diagnostic is whether frozen feature risk better orders the damage of paired equal-cost alternatives on unseen items. Report pairwise ordering accuracy, rank association and paired damage differences with intervals that resample independent items or contexts, not repeated tokens. Then test whether acting on that ranking improves a frozen complete mask. A different mask, higher feature retention, or better calibration fit is insufficient. Failure at the ranking stage redirects the work toward the representation-to-channel map; it cannot be repaired by presenting additional configurations as independent support.

F.5 Attribute a gain to selection rather than recovery or information

The primary comparison changes channel identities while holding all other interventions fixed. For each mask, the no-recovery column measures the selection rule directly. A second column applies the same mean compensation, fitted on the same activations, to every mask. Further columns can use a shared reconstruction or rotation rule. The difference between two masks within a column estimates a selection contrast under that recovery; the difference between columns for one mask measures recovery. Comparing an unrecovered statistical mask with a recovered SAE mask would confound these contributions.

Matched examples alone do not imply matched information. If functional validation uses labels, counterfactual pairs or teacher queries, PCA, random dictionaries and the statistical controls must have access to the same supervision and query allowance. Representation dimension, active-coordinate count, output projection, candidate group size and solver budget are held fixed where applicable. Training a representation introduces extra cost even when dimensions match: report fit time and forward counts separately, and include a total-construction-budget comparison. A representation-free behavioral selector on the same candidate pool checks whether direct calibration answers explain the gain without an SAE.

Table 19: Matched controls and the selection claims they distinguish. Functional variants refer to the proposed risk in Eq. 19; completed reconstruction comparisons are reported separately.

Matched contrast	Alternative explanation addressed
RMS versus functional risk	Does additional functional evidence improve deletion decisions beyond activation magnitude?
SAE energy versus validated risk	Does validation and joint deletion measurement add value beyond dictionary reconstruction?
SAE versus PCA / random	Does the sparse learned representation help beyond another equally informed coordinate system?
Same-data behavioral selector	Is the benefit explained by additional labels or teacher queries rather than the feature map?
Real versus shuffled validation weights	Does the mapping from validated effects to feature weights matter? Shuffling preserves the weight multiset.
Same masks, common recovery	Does the result persist when post-selection correction is matched?

All controls retain the same held-out items and adverse outcomes. If PCA or a random dictionary gives the same gain, the evidence favors a representation-assisted or generic reconstruction decision rule rather than SAE-specific interpretability. If a behavioral selector matches the method, supervision may explain the benefit. If feature validation passes but deletion ranking fails, the representation-to-channel map remains the bottleneck. If ranking improves but the final mask does not, interactions between successive deletions or allocation across layers remain unresolved. These distinctions make each experiment test a separate link instead of counting configurations as interchangeable evidence.

Complete FLAP and RCPU systems answer a complementary question: which complete recipe performs best at a given whole-model budget? Their original recovery and structure allocation are part of that recipe. Results from a restricted MLP-only adaptation can help isolate a component, but cannot substitute for such a system comparison. Both comparison types need PPL and the same downstream tasks; neither a single favorable capability nor a favorable mean permits omitting regressions elsewhere.

F.6 Use factorial responses as a diagnostic, then verify export

For fixed answer-score coordinates v, quartet responses are

$$d_A = (v_{10} - v_{00} + v_{11} - v_{01})/2,$$
$$d_B = (v_{01} - v_{00} + v_{11} - v_{10})/2, \quad d_{AB} = v_{11} - v_{10} - v_{01} + v_{00}, \quad R = [d_A, d_B, d_{AB}]. \quad (22)$$

Their changes diagnose the behavior a deletion distorts. FRP augments same-data KL with $\lambda\|(R_m - R_T)S\|_F^2$, where S uses frozen teacher scales with a positive floor. This is a quadratic weighting of ordinary score errors, not additional supervisory information (Appendix J). KL, unweighted score

MSE, diagonal weighting and removal of d_{AB} therefore use identical candidates and queries. The historical FRP/MSE mask identity in Table 21 provides no evidence for a uniquely factorial benefit; later iterative tests must be assessed separately.

The selected mask is exported by slicing gate/up rows and down columns, saving and reloading the smaller checkpoint, and checking outputs against the masked network. Numerical tolerances are fixed for the evaluation precision. The SAE and teacher are absent at inference. Mask-only quality is measured first, followed by comparisons in which all masks use the same recovery rule and fit data. Full FLAP/RCPU comparisons are an additional system-level test, because their original allocation and module scope can differ. Parameter count, storage, peak memory, prefill/decode throughput and construction cost are separate measurements; physical deletion alone establishes neither exact low-precision equivalence nor speedup.

G Mechanism controls and behavioral diagnostics

G.1 Calibration transfer and fixed-mask recovery

Table 20: Complete standard-suite evaluation of the original frozen masks. All pruned masks remove 35.53% of MLP channels; attention is retained. Scores are percentages; PPL is lower-better. Original arithmetic calibration is distinct from the revised calibration in Table 12.

Method	MLP%	BQ	PIQA	HS	WG	ARC-e	ARC-c	OBQA	Mean	PPL
Dense	0	85.02	74.86	68.46	65.90	78.37	53.75	40.00	66.62	18.42
Original RMS	35.53	74.31	64.58	42.40	55.17	50.00	31.40	27.80	49.38	51.46
Token-matched RMS	35.53	74.59	63.71	42.53	54.85	50.80	31.06	28.20	49.39	51.60
Original SAE	35.53	73.94	63.87	42.71	54.46	50.29	30.20	28.00	49.07	51.63
Original PCA	35.53	74.07	63.60	42.35	55.64	49.83	30.97	28.00	49.21	51.58
Random dictionary	35.53	74.77	64.64	42.75	53.83	50.63	31.57	28.80	49.57	51.18
Conditional KL	35.53	74.59	64.36	42.37	54.54	50.29	31.14	28.00	49.33	51.53
Restricted FRP	35.53	74.37	64.64	42.30	55.25	50.38	31.40	28.00	49.48	51.34

The arithmetic-calibrated original masks in Table 20 have poor language-model transfer. Their PPL cannot be compared with WikiText-calibrated masks as evidence for an SAE mechanism. The earlier factor-association audit does not show a stable directional effect (Appendix H).

Across 36 fixed-mask compensation pairs, mean arithmetic changes are $+1.765$, $+3.009$, -5.700, and -37.211 percentage points at requested 5/10/20/35% removal. A new-item confirmation (Table 24) uses 64 new numeric pairs and 256 questions per template, but reuses the template families. Its direction persists on 246 teacher-correct questions and in a separate conservative analysis counting all 81 unfinished generations as wrong. The archived table uses the original last-recognized-answer metric; eight unfinished answers were originally scored correct. This tests recovery, not channel-selection correctness. Length and vocabulary remain confounded. The development-only reliability-shrinkage candidate is also negative: its box/clean-inventory/distractor accuracies are 81.25/53.13/61.46%, versus 76.04/62.50/80.21% without compensation and 79.17/57.29/65.62% with a 0.25 scalar. Rank-eight recovery reaches 88.54/78.13/78.13% for that mask, but its advantage is not consistent across seeds or calibration distributions. The matched full-task recovery comparison above tests the same general distinction with shared WikiText calibration.

G.2 Executed SAE and FRP component pilots

These uncompensated Qwen3-4B pilots initialize all layers with RMS at requested 35% removal; only layers 12/24 change, without export. Four discovery, eight calibration and sixteen test quartets use disjoint numeric pairs, addition/subtraction $\times$ English/Chinese, and shared answer options. The 64 test questions reuse templates; algorithm seeds do not vary calibration questions.

The initial signed TopK SAE has 128 features, eight active coordinates and 150 training steps. Up to 32 tokens per prompt supply activations; the 80/20 row split does not separate prompts. Diagonal energy scoring omits feature cross terms. PCA/random match dimension and sparsity; matched RMS shares sampled rows. Features have no validated semantic interpretation.

FRP uses conditional answer-label probabilities after forced `Answer:`, choosing among the shared initializer and seven 64-channel swaps. Weight one and scale floor 0.1 are fixed. Unlike the later chain, this is neither full-vocabulary KL nor iterative search. In Table 21, FRP beats KL by one/zero answers but shares MSE's mask; SAE loses to simpler controls despite reconstruction explained energy 0.835–0.887. This pilot does not validate functional risk (Eq. 19).

Table 21: Executed bounded pilot: generated final-answer accuracy (%). Every column uses the same 64 held-out questions. Seeds index algorithm randomness, not independent calibration sets.

Method	Seed 0	Seed 1
Dense	95.31	95.31
RMS (all calibration tokens)	62.50	62.50
Matched-token RMS	70.31	64.06
Signed TopK SAE	60.94	56.25
PCA + TopK	68.75	64.06
Random orthogonal + TopK	73.44	73.44
Conditional KL selection	64.06	62.50
Conditional KL + FRP	65.62	62.50
Diagonal score control	65.62	62.50
Score MSE selection	65.62	62.50

H FACTORIAL AUDIT: SEPARATE MARGINAL COVERAGE FROM JOINT SUPPORT

H.1 A CONTROLLED CALIBRATION FAMILY

Let T be a frozen teacher and T_m the network under a binary structural mask m. A context c defines four natural inputs $q_c = (x_{00}, x_{10}, x_{01}, x_{11})$, where two editable factors $A, B \in \{0, 1\}$ vary independently. The edits must preserve the answer convention and avoid unintended changes in difficulty; independent editability is a property to validate, not one guaranteed by an SAE.

We vary association through

$$P_\rho(0,0) = P_\rho(1,1) = \frac{1+\rho}{4}, \qquad P_\rho(0,1) = P_\rho(1,0) = \frac{1-\rho}{4}, \quad \rho \in [0,1]. \tag{23}$$

Both marginals equal $1/2$ for every ρ. With signed encodings $a = 2A-1$ and $b = 2B-1$, $\mathbb{E}[ab] = \rho$. At $\rho = 0$ calibration is balanced; at $\rho = 1$ it has diagonal support only. For intermediate values, off-diagonal combinations are underrepresented, not unseen. Finite implementations record realized counts, lengths, unique contexts, and total tokens; fixed marginal counts alone do not guarantee matched difficulty or token cost.

H.2 WHY CALIBRATION AGREEMENT IS INSUFFICIENT

For $a, b \in \{-1, 1\}$, consider $T(a, b) = (a, b)$ and the fixed rank-one mapping $S(a, b) = ((a + b)/2, (a + b)/2)$. Its squared error is $1 - ab$, hence

$$\mathbb{E}_{P_\rho}\|S - T\|^2 = 1 - \rho, \qquad \mathbb{E}_{P_0}\|S - T\|^2 = 1, \qquad \mathbb{E}[\|S - T\|^2 \mid a \neq b] = 2. \tag{24}$$

Increasing calibration association hides a missing direction without restoring it (Figure 11). This is an elementary support example, not evidence that an actual pruning algorithm selects S or that transformer channels encode unique factors.

H.3 MEASURE EXCESS DAMAGE, NOT RAW CELL DIFFICULTY

For cell group C, define the paired damage $D_C(m) = \mathbb{E}_{x \in C}[\ell(T_m, x) - \ell(T, x)]$. For a mask m_ρ selected under P_ρ, use

$$G(\rho) = D_{\text{off}}(m_\rho) - D_{\text{diag}}(m_\rho), \qquad \Gamma(\rho) = G(\rho) - G(0). \tag{25}$$

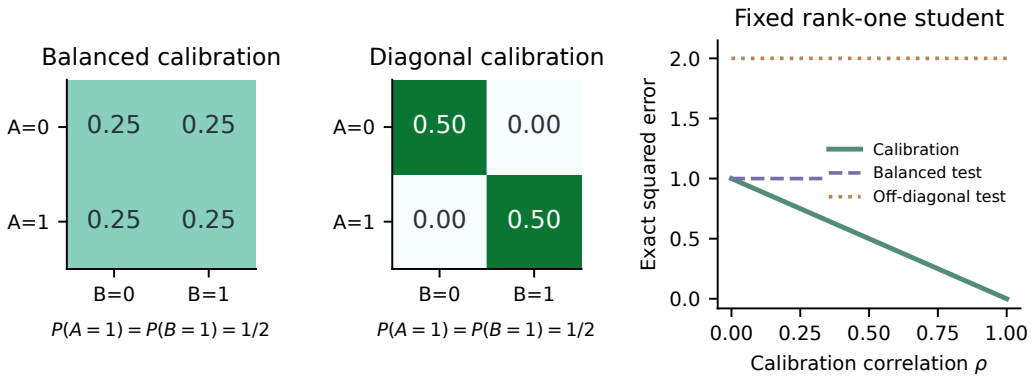

Figure 11: **Coverage of factors is not coverage of combinations.** Left: the two endpoint calibration distributions share balanced marginals. Right: an exact illustration for a fixed rank-one student shows falling calibration error without improved balanced-test fidelity. These are specified distributions and analytical quantities, not neural-model results.

A positive Γ is the predicted direction of additional relative off-diagonal damage. It is not sufficient to identify a neuron-level mechanism. Both component damages and dense-model performance must be reported. A separate diagonal-reversal control is needed to distinguish association from the particular cells emphasized. The statistic applies to error rate or a prespecified loss; changing outcomes after inspecting results yields exploratory, not confirmatory, evidence.

The target evaluation uses disjoint contexts, templates, and entities. The existing pilot has disjoint full prompts but reuses templates and vocabulary; it tests the measurement pipeline rather than this stronger generalization claim.

I Exact support calculation

For $a, b \in \{-1, 1\}$, the score error of the fixed mapping in Eq. 24 is

$$S(a,b) - T(a,b) = \left(\frac{b-a}{2}, \frac{a-b}{2}\right), \qquad \|S - T\|^2 = \frac{(a-b)^2}{2} = 1 - ab. \tag{26}$$

Equation 23 yields $\mathbb{E}[ab] = \rho$, hence calibration error $1 - \rho$. Under balanced evaluation it is 1; conditioned on $a \neq b$, it is 2. The matrix of S is $\frac{1}{2}\left[\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right]$, a rank-one projection onto the diagonal. The teacher has rank two. At $\rho = 1$ no calibration input tests the missing direction; for $\rho < 1$ its evidence is downweighted. The example specifies a candidate mapping and does not solve an optimization problem or characterize the representation learned by a transformer.

J Response loss, invariances, and the exact score control

For $E = [e_{00}, e_{10}, e_{01}, e_{11}] \in \mathbb{R}^{d\times 4}$,

$$C = \begin{bmatrix} -1/2 & -1/2 & 1 \\ 1/2 & -1/2 & -1 \\ -1/2 & 1/2 & -1 \\ 1/2 & 1/2 & 1 \end{bmatrix}, \qquad R_m - R_T = EC. \tag{27}$$

The loss is $\|ECS\|_F^2 = \operatorname{tr}(ECSS^\top C^\top E^\top)$. Each contrast column sums to zero, so adding the same error vector to all four cells leaves the response loss unchanged. The columns are mutually orthogonal with squared norms $1, 1, 4$.

This yields a useful exact control. Let $U = C\,\mathrm{diag}(1, 1, 1/2)$ and $P = I_4 - \frac{1}{4}\mathbf{1}\mathbf{1}^\top$. Then $U^\top U = I_3$ and $UU^\top = P$, so

$$\|EU\|_F^2 = \|EP\|_F^2 = \sum_{i=1}^{4} \|e_i - \bar{e}\|^2, \qquad \bar{e} = \tfrac{1}{4}\sum_{i=1}^{4} e_i. \tag{28}$$

Moreover $\|E\|_F^2 = \|EU\|_F^2 + 4\|\bar{e}\|^2$. Thus equally weighted orthonormal factorial contrasts are exactly centered score-error matching. Teacher normalization in FRP changes these weights but does not create new target information. Comparing against centered score matching and ordinary score matching controls for this interpretation. Equation 28 is an algebraic identity, not an empirical advantage or a new general theorem.

The floor ϵ must be positive and fixed before final testing. Report it together with λ, teacher score coordinates, prediction positions, and the definition of μ_{cal}. Compare normalized and unnormalized losses under the same validation-search budget. A loss-scale change must not be mistaken for a semantic benefit.

K Selection algorithm specification

The following procedure records the general FRP specification. The historical pilot executed a one-round, two-layer adaptation. The completed all-layer, four-round extension is documented separately in Appendix C; its exact candidate, query and export protocol takes precedence over this historical specification.

1. Freeze teacher weights, tokenizer, contrast factors, score coordinates, and split assignment. Cache teacher predictions on the permitted calibration and validation inputs.
2. Choose an exact structural budget and legal block size. Initialize a shared activation mask for ordinary KD, centered score matching, and FRP.
3. Generate a fixed number of seeded keep/drop block pairs. Use layer-local swaps when channel costs differ between layers, or explicitly match total parameter/compute cost.
4. Evaluate every candidate on identical calibration batches. Accept the best strict improvement, with lexicographic tie-breaking by layer/block. Stop at the evaluation budget or when no candidate improves the objective.
5. Select hyperparameters on validation data. Slice the coupled gate/up rows and down columns, preserving biases if present. Keep residual and attention dimensions unchanged.
6. Verify masked/exported outputs in evaluation mode using declared tolerances and precision. Save shapes, masks, model revision, and export metadata.
7. Evaluate the fixed mask on the final test. If recovery is studied, apply a common recovery recipe and report both pre- and post-recovery results.

Count teacher/student calls, validation trials, and search time. Match quartets and evaluations across objectives. Hardware-aligned blocks are prospective; the pilot ranks individual channels.

L Data boundaries and stronger evaluation design

Assign each quartet to one split and record its context/template/entity IDs, labels, prompt, answer, tokenizer revision, and source. Separate factor discovery, edit validation, calibration, hyperparameter selection, and testing. Task-defined audits need no SAE discovery split; the pilot does not instantiate all these future roles.

Teacher competence. Fix a validation capability threshold before pruning and require it in every cell. Do not lower it after observing test failures. Report teacher performance on all test examples. Teacher-correct-only analysis may be supplemental, but must state its denominator; otherwise filtering can hide an unsuitable task.

Distribution and budget controls. Publish cell counts, unique contexts, duplicate rates, total token counts, and length distributions for each calibration set. Hold the context pool and sample budget fixed, and add token/length-matched replication. Check whether answer frequencies or difficulty shift with association. A reversed-diagonal replication is necessary before claiming the effect is independent of the chosen pairing. More informative tasks must be developed on new development data and evaluated on a freshly locked test, because the present microtest has already been inspected.

Contrast validity. Audit grammaticality, answer consistency, unintended cues, and whether each edit changes the intended factor. Automatic checks are not human validation. Hidden-state interventions may supplement natural-input tests but cannot establish that a latent edit has a natural-language counterpart. Transfer to unseen templates, to new factor families, and to another model are distinct evaluations.

M Full comparison and rejection criteria

The planned comparison separates data access, selection loss, recovery, and representation:

1. **Phenomenon:** a non-ceiling teacher-qualified task, multiple ρ values and budgets, matched counts/lengths, and diagonal reversal. Report component damages and Γ, not only aggregate accuracy.
2. **Data:** compare original calibration with quartet calibration under the same pruner and token budget. Include faithful published baselines such as FLAP/FANG and a compatible COLA curation comparison in a full study.
3. **Objective:** same-quartet KD, score matching, centered score matching, and FRP under the same initialization/search. Ablate main and interaction terms separately while controlling scale.
4. **Recovery:** evaluate every mask first without recovery, then with identical tokens, optimizer, trainable parameters, and steps. Report final deployed quality rather than only a favorable pre-recovery gap.
5. **Discovery:** task labels, random matched candidates, raw-activation clusters, PCA, and SAE candidates share discovery data and a teacher-query budget. Report representation-training cost separately.

If balanced data explain the benefit, narrow the claim to calibration. If same-data score matching or KD match FRP, drop objective-superiority claims. If cheaper discovery matches SAE, drop SAE necessity. If parameters shrink but latency does not improve, report parameter compression alone. Negative outcomes cannot all be relabeled as confirmation of the same mechanism.

N Executed pilot and overlap correction

The four runs cross two association endpoints with two calibration seeds. Raw records reside under `results_v2/`; the executed script is `experiments/archive/micro_prune_v2_executed.py`. The frozen model snapshot is:

`1cfa9a7208912126459214e8b04321603b3df60c`

Qwen3-4B has 36 layers, hidden width 2,560, and intermediate width 9,728. The recorded environment uses PyTorch 2.6.0+cu124 and Transformers 4.55.2 on shared A800 GPUs.

For calibration seed s, generate 32 contexts with randomly chosen red/blue and round/square attributes. Each context supplies one complementary pair of cells. Sixteen contexts supply diagonal pairs at $\rho = 0$ and all 32 do so at $\rho = 1$; the remaining contexts supply off-diagonal pairs. This yields exact language/task marginals. Test generation uses seed 314159 and an object-ID offset of 10,000. English/Chinese prompts use the same small answer vocabulary and non-thinking chat template. All four words are checked to be single tokens before scoring.

The original v1 pilot had 12/96 and 16/96 exact calibration–test prompt overlaps for seeds 0 and 1. That result was retained, audited, and corrected by the test-ID offset; the four v2 runs have zero exact overlaps. This correction does not create new templates or semantic factors. Current analyses exclude v1 entirely. The corrected runs are an audit-triggered exploratory repetition, not an independent confirmation.

The historical script uses four PyTorch intra-op threads; its archived reanalysis did not launch GPU jobs. The subsequent experiments are documented separately in Appendix Q. Whole-run duration includes evaluation and pruning and is not an inference-speed benchmark.

O Reanalysis, uncertainty, and artifact provenance

The script `paper/reanalyse_pilot.py` reads only corrected raw JSON records. It verifies matching dense/pruned IDs, prompts and answers; 64 calibration records; the four expected test cells for each of 24 contexts; zero exact overlaps; 36 unique retained-channel lists of width 7,782; and identical dense records across all four runs. Tables are generated directly from these records. The saved analysis ledger contains input and script hashes, per-context losses, run summaries, mask overlap, and contrast intervals.

For each seed s and context c, compute

$$g_{\rho,s,c} = \tfrac{1}{2}(\Delta\ell_{01} + \Delta\ell_{10} - \Delta\ell_{00} - \Delta\ell_{11}), \quad \gamma_{s,c} = g_{1,s,c} - g_{0,s,c}. \tag{29}$$

Average $\gamma_{s,c}$ over contexts to estimate $\Gamma_s(1)$. The bootstrap uses 10,000 draws of 24 context indices with replacement, RNG seed 20260919, and identical sampled contexts across cells and masks. The 2.5th/97.5th percentiles form a conditional context interval. They quantify empirical context resampling for fixed masks, not uncertainty over new calibration seeds, templates, or tasks. With two seeds, no population-level significance claim is made. Accuracy is exactly at ceiling in the saved sample, so a degenerate empirical interval would not establish population equivalence.

Mask replacement uses retained channels as denominator; all 36 layers are shown.

P Deployment and reporting requirements

Verify export equivalence against the masked network before deployment tests. Record revisions, software, precision, configuration, data hashes, masks, seeds, and tensor counts. Peak memory during a pruning job is not steady-state exported-model memory; attention and embeddings also remain in the total-parameter denominator.

Benchmark prefill and decode on uncontended devices with matched kernels, inputs, warm-up, and synchronization. Separate search costs; the pilot establishes no speedup.

Q Expanded exploratory evidence, September 20

Evidence status. The comparisons below distinguish historical pilots, matched development diagnostics, a frozen unseen-numeric-item confirmation, and an unsuccessful candidate method. Confirmation results support template-dependent compensation effects in the specified setting, not FRP or SAE effectiveness. This appendix supplements the historical four-run pilot with 193 subsequently completed configurations: 30 WikiText (including three reconstruction runs), nine arithmetic reconstruction runs, 108 arithmetic, one relational capability probe, four compensation-dose probes, four matched-template probes, twelve cross-mask probes, three shrinkage candidates, four scalar controls, and eighteen frozen-item confirmation runs. Another 107 relational configurations were skipped after the dense teacher failed the prespecified capability gate. These are completed configurations, not independent datasets. No FRP or SAE method is validated.

Expanded factorial audit. The arithmetic grid crosses three calibration seeds, three associations $(-1, 0, +1)$, four requested MLP removal fractions (5,10,20,35%), and three methods: activation RMS, an adaptation of the official FLAP statistic with compensation, and the same statistic without compensation. Evaluation uses 384 shared questions and greedy generation with a 512-token limit.

Final-option parsing replaces the historical early-token proxy. The dense teacher answers 370/384 correctly. Difficulty levels two and three pass an 80% per-cell teacher gate and account for two thirds of the examples; all levels remain in the full report. Numeric pairs are disjoint across splits, but the modulo partition also shifts arithmetic residues and is not an IID split.

No consistent association-reversal effect emerges across methods, budgets, and seeds. At 35% removal, uncompensated FLAP has mean reversal contrasts of $+11.806$ and -2.083 percentage points over all levels. However, the teacher-qualified subset reverses direction in one of three seeds. These exploratory means do not establish the proposed mechanism. Raw outputs and negative results are retained.

Public-data sanity benchmark. WikiText evaluation uses the first 128 non-overlapping 512-token test blocks, totaling 65,408 predicted tokens. Calibration uses 64 training windows and three seeds. Dense perplexity is 17.456. At requested 10/20/35% removal, mean RMS perplexities are 17.955/20.132/25.094; FLAP-statistic adaptation with compensation gives 18.557/20.635/25.858, versus 19.274/22.413/29.499 without compensation. This MLP-only, uniform-width Qwen adaptation is not a reproduction of published full-FLAP results. Compensation adds 92,160 bias parameters, so equal channel budgets do not imply equal parameter counts.

Fixed-mask compensation and template dependence. All 36 matched arithmetic comparisons between compensated and uncompensated FLAP retain identical channel indices. At requested 35% removal, compensation reduces mean arithmetic accuracy by 37.211 percentage points across nine configurations, despite helping WikiText perplexity. This observation motivated a separate development diagnosis, not a new-method claim.

For one fixed 35% mask (balanced calibration, seed zero), we evaluate 96 matched development questions. Numeric operands, answer positions, options, and masks are identical across templates. The zero-compensation endpoint retains a zero-bias linear layer; it is not asserted to be bitwise identical to the historical bias-free export. Table 22 reports all endpoints. Intermediate compensation multipliers 0.25 and 0.5 each answer 76/96 box-template questions correctly, versus 73/96 at zero and 75/96 at one; this small development difference is not evidence for a tuned method.

Table 22: Matched-template development diagnosis: correct answers out of 96. One calibration seed and one mask; no cross-seed confirmation.

Template	Dense	Compensation 0	Compensation 1
Box	95	73	75
Inventory, clean	92	60	35
Inventory, irrelevant number	96	77	36

The inventory wording changes the compensation effect even without an irrelevant number. Consequently, neither numeric sample changes nor distractors alone explain this particular contrast. Word choice and length remain entangled. Compensated outputs are all parseable, with completion rates at least 98.958%. Independent items, calibration seeds, common teacher-correct subsets, and implementation controls are needed before generalization. The official fluctuation accumulator also aliases its old-mean tensor during an in-place update; the adaptation preserves this implementation rather than silently substituting a corrected estimator. Physical-export BF16 discrepancies are recorded, not claimed to vanish. No isolated speedup measurement is available.

Subsequent cross-mask check on the same development items. Twelve additional runs extend the three-template, two-compensation comparison to calibration seeds one and two. Each seed uses its own saved mask, held fixed across templates and compensation endpoints. Table 23 reports the compensation effect. All three masks show a more negative effect for inventory wording than for box wording. Restricting evaluation to the 91 items the dense teacher answers correctly in all three templates preserves this direction for every seed. This addresses a teacher-capability confound but does not provide independent-item confirmation: all masks are evaluated on the same development questions. In particular, compensation is not consistently beneficial on the box template.

Table 23: Compensation-one minus compensation-zero accuracy (percentage points) on the same 96 development questions. Individual calibration seeds are shown rather than treated as independent test datasets.

Calibration seed	Box	Inventory clean	Inventory distractor
0	2.083	-26.042	-42.708
1	-9.375	-29.167	-39.583
2	-8.333	-23.958	-23.958

A negative candidate-method result. We tested calibration-only channelwise shrinkage of the compensation mean. For each deleted channel, the multiplier is $r = \mu^2/(\mu^2+v+10^{-8})$, where μ and v are independently accumulated calibration moments; the original mask and official compensation mean are retained. This generic shrinkage adaptation has no established novelty. On the seed-zero development comparison, accuracies for box, clean inventory, and inventory with a distractor are respectively 81.25%, 53.125%, and 61.458%. A scalar multiplier of 0.25 gives 79.167%, 57.292%, and 65.625%; zero compensation gives 76.042%, 62.5%, and 80.208%. Thus the candidate does not improve upon either of these simple controls on inventory prompts. We retain this negative result rather than label the candidate a superior method. Its WikiText performance has not been measured.

Scoring definition and completion sensitivity. The final-option parser uses the last recognized answer marker in the generated text. Unparseable answers count as incorrect; truncation alone does not automatically count as incorrect in the original metric. Across the 4,608 pruned confirmation records, 81 generations do not terminate normally, and eight of these receive a correct score. A separate conservative analysis counts every non-terminated generation as incorrect without changing the saved original results. Compensation-one minus compensation-zero effects for seeds zero, one, and two are respectively $(-5.078, -8.984, -9.375)$ percentage points on box wording, $(-32.031, -32.812, -24.609)$ on clean inventory, and $(-50.391, -37.109, -33.203)$ with an inventory distractor. Inventory effects remain more negative for every mask. This does not replace semantic adjudication of ambiguous or self-contradictory text.

Frozen unseen-item confirmation protocol. Before confirmation inference, we froze 64 previously unused numeric pairs, producing 256 questions per template across difficulty levels two and three. Numeric pairs do not overlap collected earlier calibration or evaluation pairs. The same three templates and historical calibration masks are retained; this is new-item confirmation within the template family, not cross-domain evaluation. Eighteen runs cross the three masks, three templates, and compensation endpoints zero and one. Dense accuracy is 100%, 96.094%, and 100% for box, clean inventory, and inventory with a distractor; the weakest difficulty/factor cell is 81.25%.

Table 24: Frozen-item confirmation accuracy (%), mean $\pm$ sample standard deviation over three calibration masks. All masks share the same 256 evaluation questions per template.

Template	Zero compensation	Full compensation
Box	92.32 ± 5.56	84.51 ± 3.18
Inventory, clean	72.66 ± 6.66	42.84 ± 4.30
Inventory, irrelevant number	80.99 ± 3.75	40.49 ± 8.12

Compensation has a more negative effect on both inventory templates than on box wording for each mask. This direction persists on the 246 questions answered correctly by the dense model in every template. Thus the development interaction survives new numeric items under the frozen protocol. This does not identify a unique semantic mechanism, establish behavior at other compression levels, or demonstrate a better pruning method. The earlier shrinkage candidate is not evaluated on this confirmation set and is not credited with these results.

R ADDITIONAL RESPONSE-OBJECTIVE DERIVATIONS

R.1 WHAT THE RESPONSE PENALTY CHANGES

The purpose of factorial matching is to prioritize particular directions of prediction error under a fixed structural budget. It does not supply additional teacher information when the ordinary matching control observes the same four inputs. This distinction matters because a comparison against a baseline calibrated only on selected cells would mix two interventions: expanding calibration coverage and changing the objective. Our same-data control therefore receives all cells, and our bounded experiment evaluates every objective on exactly the same candidate masks. Any difference between the selected masks must arise from their ordering under the objectives, rather than from a different candidate-generation budget.

Write the score error for quartet cell c as $e_c = v_m(x_c) - v_T(x_c)$, and collect these columns in $E \in \mathbb{R}^{d\times 4}$. With cells ordered as $(00, 10, 01, 11)$, define

$$C = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} & 1 \\ \frac{1}{2} & -\frac{1}{2} & -1 \\ -\frac{1}{2} & \frac{1}{2} & -1 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}, \qquad W = CSS^\top C^\top, \qquad \|ECS\|_F^2 = \sum_{c,c'} W_{cc'}\langle e_c, e_{c'}\rangle. \tag{30}$$

The diagonal terms weight each cell's error magnitude. The off-diagonal terms couple errors across different inputs; they distinguish, for example, a common error from errors that change with a factor. Because $C^\top \mathbf{1} = 0$, adding the same vector to all four error columns leaves this penalty unchanged. Ordinary prediction matching constrains behavior that such contrasts do not observe. This invariance is a property of the chosen score representation, not evidence that the ignored error is harmless for generation.

A useful control removes the off-diagonal terms while preserving their associated within-cell weights:

$$\mathcal{L}_{\mathrm{diag}}(E) = \sum_c W_{cc}\|e_c\|_2^2, \qquad \mathcal{L}_{\mathrm{FRP}}(E) - \mathcal{L}_{\mathrm{diag}}(E) = 2\sum_{c<c'} W_{cc'}\langle e_c, e_{c'}\rangle. \tag{31}$$

Thus an advantage over prediction matching alone would not isolate the cross-input terms: it could reflect the additional score penalty or its scale. A matched diagonal control addresses this alternative. Ordinary and cell-centered score errors provide further comparisons, although they are not algebraically equivalent to the scale-weighted response objective. In our implementation, the exact quadratic expression is checked against the contrast computation numerically; these are two calculations of one loss, not two methods.

R.2 FROM SPARSE RECONSTRUCTION TO A CHANNEL SCORE

The sparse-feature component answers a different question: whether a learned representation gives a useful ranking of original MLP channels. Let a_j denote an intermediate activation, with calibration mean μ_j and positive normalization scale s_j. A dictionary approximates the standardized activation by $\hat{a}_j = \mu_j + s_j \sum_f z_f D_{fj}$. Expanding its second moment gives

$$\begin{aligned}\mathbb{E}[\hat{a}_j^2] = \mu_j^2 + 2\mu_j s_j \sum_f \mathbb{E}[z_f] D_{fj} + s_j^2 \sum_f \mathbb{E}[z_f^2] D_{fj}^2 \\ + 2s_j^2 \sum_{f<g} \mathbb{E}[z_f z_g] D_{fj} D_{gj}.\end{aligned} \tag{32}$$

Our pilot uses only the squared mean and diagonal feature-energy terms, multiplied by the squared norm of the corresponding down-projection column. It omits both the latent-mean cross term and cross-feature second moments. Signed TopK activations do not guarantee zero latent means or mutually uncorrelated features, so these omissions are approximations rather than identities. Reconstruction residuals also separate the reconstructed second moment from the original activation energy. Moreover, ranking channels separately ignores interactions among the channels removed together. The resulting score is therefore an energy heuristic, not a causal attribution or an exact pruning-error estimate.

R.3 Optimization scope and what a pilot can establish

The general proposal permits iterative, equal-cost swaps from a common initializer. The historical pilot response experiment instead fixes eight candidates: the initializer and seven independently proposed neighbors. Each neighbor exchanges equal numbers of retained and removed channels within one permitted layer. Only two layers may change, while the remaining layers retain the shared RMS decisions. Every objective reranks this identical pool once; no selected neighbor seeds another search round. This design makes the immediate objective comparison inexpensive and controlled, but sharply limits the reachable structures. A negative result cannot exclude stronger optimization, and a positive result would establish only a preference within this pool. Likewise, the separately tested dictionary score does not instantiate SAE-discovered contrasts followed by response optimization. That combination was not executed in the historical pilot. The later feature-screened, task-defined chain is now completed (Appendix C), but it does not validate autonomous semantic discovery or the independent functional-risk hypothesis.

S Component audit protocol and remaining tests

Table 25: Matched masked-model component audit at 35.53% MLP removal: 2,048 features, 512 windows, seed zero. Accuracy (%, ↑) covers 20,465 questions; WikiText-2 PPL (↓) covers 298,493 predicted tokens. Selected contrasts; the complete control table, including the untrained random dictionary, is Table 26. Bold marks the best displayed values.

	No recovery		Common mean	
Variant	Accuracy	PPL	Accuracy	PPL
RMS / no swaps	49.36	26.00	50.75	25.46
SAE diagonal (closed form)	49.04	25.00	50.44	25.36
SAE diagonal (64 swaps)	49.41	25.62	50.58	25.50
SAE without residual	**49.77**	25.21	50.81	25.43
Empirical joint	49.58	25.19	50.71	25.57
TopK PCA joint	49.57	25.49	50.89	25.68
Linear PCA joint	49.42	25.19	**50.96**	25.57
Dense AE joint	49.39	25.90	50.62	25.71
SAE joint (ours)	49.53	**24.77**	50.58	**25.32**

Table 25 reports the completed matched PPL and seven-task component audit. Items 1–5 below have complete PPL and seven-task measurements for all ten variants under both recovery rules. Item 6 and independent-risk, repeated-seed and new-mask export tests remain unexecuted in this campaign. Hold the checkpoint, calibration split, per-layer width, output projection, initialization and evaluation protocol fixed within each contrast; retain adverse outcomes.

1. Replace SAE reconstruction with empirical second moments under identical joint swaps: isolates the need for learned representation.
2. Replace SAE with TopK PCA coordinates, dense linear PCA, a dense autoencoder and a fixed random TopK dictionary on the same samples. The dense autoencoder removes TopK at matched training steps; the random dictionary uses zero training steps and tests learning, not equal training compute.
3. Set the swap count to zero: isolates refinement of the shared RMS initialization. Also compare diagonal scoring against joint scoring. Closed-form diagonal selection changes the solver as well as the objective. A further control uses the same RMS start and 64-swap cap on the diagonal objective to remove this solver confound.
4. Remove the diagonal residual penalty while holding reconstructed factors fixed: isolates residual modeling.
5. Cross every frozen mask with no recovery and the same mean recovery: isolates correction from mask selection.
6. Vary dictionary width, active count, data volume and optimization steps one at a time, with matched seeds and a fixed mask-scoring dataset. Widths below the 9,728-dimensional input remain undercomplete, including 4,096 and 8,192.

Each variant requires all-layer masks, retained-count checks, held-out reconstruction and deletion-risk diagnostics, followed by common PPL and task evaluation. Training completion or decreasing proxy loss alone is not an end-to-end result. The large-data launch that varied both seed and sample count is exploratory and cannot isolate either factor; its failures must be retained. Final candidates require repeated seeds and verified physical export. Selection uses development data only.

Table 26: Matched masked-model component audit at 35.53% MLP removal: 2,048 features, 512 windows, seed zero. Accuracy (%, ↑) covers 20,465 questions; WikiText-2 PPL (↓) covers 298,493 predicted tokens. All predefined controls are included. Random TopK is untrained. Best values are bold within each column.

	No recovery		Common mean	
Variant	Accuracy	PPL	Accuracy	PPL
RMS / no swaps	49.36	26.00	50.75	25.46
SAE diagonal (closed form)	49.04	25.00	50.44	25.36
SAE diagonal (64 swaps)	49.41	25.62	50.58	25.50
SAE without residual	49.77	25.21	50.81	25.43
Empirical joint	49.58	25.19	50.71	25.57
TopK PCA joint	49.57	25.49	50.89	25.68
Linear PCA joint	49.42	25.19	50.96	25.57
Dense AE joint	49.39	25.90	50.62	25.71
Random TopK joint	**50.59**	25.23	**51.90**	25.68
SAE joint (ours)	49.53	**24.77**	50.58	**25.32**

S.1 Matched component audit: task-level outcomes

The component campaign uses one calibration seed and the larger training setting of Table 25. Every task uses its complete labeled split. BoolQ and WinoGrande use raw likelihood accuracy; other tasks use the length-normalized accuracy specified in the main protocol. The macro mean weights tasks equally. These publicly inspected tests are exploratory, not independent confirmation.

Table 27: No recovery: matched component accuracy (%, ↑). Every row covers the same 20,465 questions. Best values are bold within this panel; the complete SAE selector is last.

Method	BoolQ	PIQA	Hella	Wino	ARC-e	ARC-c	OBQA	Mean
RMS / no swaps	62.23	64.36	49.45	58.17	48.53	30.80	32.00	49.36
Diagonal (closed form)	62.20	64.15	49.30	58.17	46.13	30.12	33.20	49.04
SAE diagonal (64 swaps)	62.17	64.25	48.89	58.17	47.77	31.40	33.20	49.41
SAE no residual	62.42	**65.56**	49.91	57.14	48.82	32.17	32.40	49.77
Empirical joint	62.66	65.23	50.08	57.30	47.77	32.00	32.00	49.58
TopK PCA joint	62.60	64.53	49.99	57.38	46.84	32.25	33.40	49.57
Linear PCA joint	62.26	64.42	50.02	57.54	47.52	32.17	32.00	49.42
Dense AE joint	62.57	64.31	49.32	58.72	46.93	31.48	32.40	49.39
Random TopK	63.33	64.47	**50.25**	**60.06**	**49.33**	**32.51**	**34.20**	**50.59**
SAE joint (ours)	**63.58**	64.91	50.05	57.06	46.76	31.57	32.80	49.53

Table 28: Common mean recovery: matched component accuracy (%, ↑). Every row covers the same 20,465 questions. Best values are bold within this panel; the complete SAE selector is last.

Method	BoolQ	PIQA	Hella	Wino	ARC-e	ARC-c	OBQA	Mean
RMS / no swaps	62.20	65.78	49.95	58.88	51.14	33.53	33.80	50.75
Diagonal (closed form)	62.26	64.91	49.80	61.01	50.08	31.83	33.20	50.44
SAE diagonal (64 swaps)	62.20	65.40	49.84	59.83	51.05	32.76	33.00	50.58
SAE no residual	62.26	66.05	50.08	59.04	51.64	33.19	33.40	50.81
Empirical joint	62.87	65.56	50.38	59.51	51.01	32.85	32.80	50.71
TopK PCA joint	62.63	65.29	50.28	59.75	49.96	33.11	35.20	50.89
Linear PCA joint	62.23	**66.10**	50.29	59.75	51.01	33.53	33.80	50.96
Dense AE joint	62.26	65.51	49.85	59.59	50.21	32.94	34.00	50.62
Random TopK	**63.36**	65.51	**50.43**	**61.64**	**53.11**	**33.87**	**35.40**	**51.90**
SAE joint (ours)	63.21	65.67	50.30	58.56	49.16	33.19	34.00	50.58

S.2 MATCHED COMPONENT CAMPAIGN

The follow-up component protocol uses 512 WikiText training windows, 409 fitting and 103 diagnostic windows, with 32 sampled activations per window: 13,088 fitting rows per layer. The reference SAE has 2,048 features, TopK 32 and 1,000 training steps at seed zero. All variants retain 6,272 of 9,728 channels in each of 36 layers, share rank-64 output projections and use a 64-swap cap where applicable. These settings differ from the 512-feature, 64-window main comparison; cross-protocol differences do not isolate a component.

The component contrasts are: full SAE versus zero swaps (refinement from RMS); versus diagonal objectives (both closed-form selection and matched 64-swap search); versus zero residual penalty (residual correction); versus exact empirical second moments (learned representation versus measured moments); versus TopK PCA, dense PCA, dense AE and a fixed random dictionary (representation, sparsity and learning controls). Cross every mask with no recovery and a shared mean-fill rule. In this campaign, recovery fills deleted BF16 activations with fitted means before down projection; historical tables fold the mean into an output bias. These are algebraically equivalent but need not be numerically identical in BF16.

The completed PPL audit verifies all 36 retained-index sets, model shards, fitting-window identities, output projections and recovery means. Each run scores 298,493 tokens in 585 nonoverlapping blocks from the complete WikiText-2 test split; the last partial block is included. Per-block losses reproduce the reported aggregate. Every predefined variant is retained, irrespective of test performance. All 20 conditions also complete the seven official labeled splits (20,465 questions per condition). Per-item scores, predictions, metric choices and mask identities were audited against the PPL records. These are single-seed masked-model measurements, not new physical-export measurements.

Numerical and uncertainty checks. The two evaluation hosts use PyTorch 2.6.0 and 2.5.1 with Transformers 4.55.2 and identical model shards. Repeating the same unrecovered full mask on the second host gives PPL 24.76768 versus 24.76982, an absolute difference of 0.00214. The recovered A100 witness gives PPL 25.32104 versus 25.31856 and seven-task accuracy 50.5513 versus 50.5846 percent on the original host. These repeat evaluations quantify a numerical difference for one mask, not calibration-seed uncertainty; the mixture comparison uses the second-host A100 endpoints consistently. Recovery-mean tensor hashes agree across methods. Paired-block bootstrap intervals condition on this calibration seed and test corpus: the full-minus-diagonal mean NLL interval is $[-0.01395, -0.00427]$ without recovery but $[-0.00587, 0.00268]$ with common mean recovery (5,000 resamples). Thus the smallest recovered PPL difference is a point-estimate difference not resolved by this conditional interval. These intervals do not measure calibration-seed variability.

T EXPLORATORY SAE–RANDOM MIXTURE: PROTOCOL AND RESULTS

This follow-up is motivated by the observed difference between PPL and accuracy rankings; it is exploratory rather than an independent confirmation. Let $J_S(M)$ and $J_R(M)$ be the joint deletion objectives from the trained SAE and fixed random TopK dictionary, respectively, and let M_0 be their shared RMS initialization. For positive $c_S = J_S(M_0)$ and $c_R = J_R(M_0)$, the proposed mixture is

$$J_\lambda(M) = (1-\lambda)\frac{J_S(M)}{c_S} + \lambda\frac{J_R(M)}{c_R}. \tag{33}$$

The normalization uses fitting activations only. The weight mixes objectives, not a fraction of randomly deleted channels. At the primary weight 0.5, the resulting mask replaces an average of 38.44 retained indices per layer relative to the pure SAE mask (range 16–47); each layer still retains exactly 6,272 channels. Stacking $\sqrt{(1-\lambda)/c_S}\,B_S$ and $\sqrt{\lambda/c_R}\,B_R$ and applying the same weights to the residual penalties implements this objective under the shared output projection. It preserves the structural budget but does not guarantee improved PPL or accuracy.

The protocol fixes $\lambda = 0.5$ as the primary mixture and evaluates 0.25 and 0.75 as sensitivities; the two pure endpoints remain controls. All weights are reported, rather than selecting a test-best value. The new masks use the same initialization, 64-swap cap, calibration rows and recovery rules as the component audit. Both dictionaries add offline construction cost, while the deployed

retained channel count is unchanged. All six mixture conditions completed full evaluation. The random endpoint is retained in the complete results even when its standalone row is omitted from the selected main-table contrasts.

Table 29: Exploratory SAE–random mixtures on the same A100 evaluation host. Each mask is evaluated on all seven tasks (accuracy in percent, ↑) and full WikiText-2 (PPL, ↓). The random weight $\lambda = 0.5$ was fixed before mixture evaluation; all tested weights and both endpoints are retained. Bold marks the best within each column.

	No recovery		Common mean	
Variant	Accuracy	PPL	Accuracy	PPL
Pure SAE	49.65	**24.77**	50.55	**25.32**
Pure random TopK	**50.59**	25.23	**51.90**	25.68
Mixture $\lambda = 0.25$	49.27	24.93	50.70	25.33
Mixture $\lambda = 0.75$	50.26	24.98	51.14	25.55
Mixture $\lambda = 0.5$ (ours, primary)	49.46	25.03	50.51	25.44

Completed mixture outcomes. At the prespecified weight 0.5, accuracy is 49.462/50.512 percent without/with recovery, versus the same-host SAE endpoint at 49.648/50.551 percent. PPL increases from 24.768/25.321 to 25.029/25.444. Thus the primary mixture does not improve either point estimate over SAE. The accuracy differences have conditional paired-item 95 percent intervals of $[-0.651, 0.285]$ and $[-0.439, 0.365]$ percentage points (5,000 task-stratified paired resamples); both include zero. These intervals describe this single seed and repeatedly inspected test set, not independent confirmation or seed variability.

The 0.75 sensitivity improves accuracy over SAE but worsens PPL, while remaining less accurate than the pure random endpoint. It therefore illustrates a tradeoff, not dominance over both endpoints. We retain the prespecified primary weight rather than promoting this sensitivity after observing the test scores. A simple convex mixture does not establish a distinct SAE functional advantage.

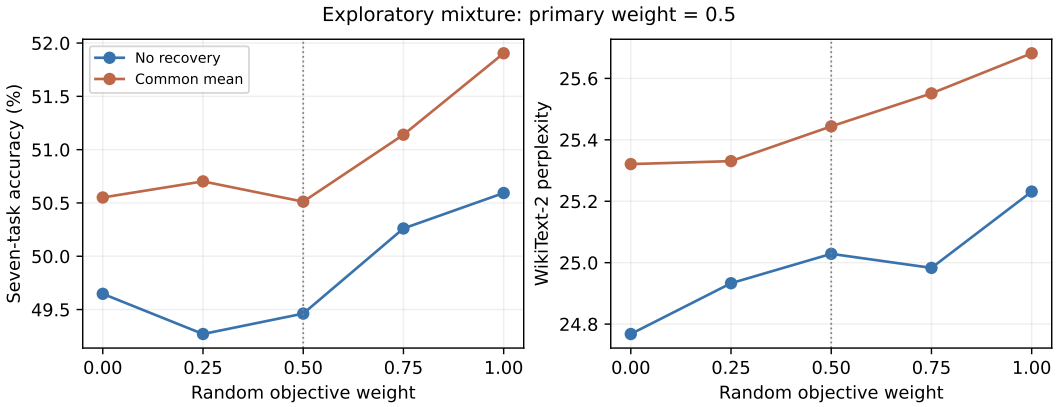

Figure 12: All prespecified mixture weights and both endpoints. The dotted line marks the primary weight 0.5. Points are one calibration seed; lines connect evaluated settings and do not represent uncertainty.