WHEN ROUTING EDITS DO NOT ADD UP: BUDGETED SELECTION IN MIXTURE-OF-EXPERTS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Compressing a sparse mixture-of-experts model changes which experts its router selects. Candidate edits are often scored one at a time, although an upstream edit changes the hidden states on which later edits act. We find that edits that help in isolation can lose reference-answer log probability when combined. Under additive logit responses, a local quadratic expansion expresses pairwise interaction through the softmax Fisher inner product. We propose Ranked Validation, which measures the empty set and every single edit, orders larger candidate sets by summed single-edit scores, measures the leading sets, and returns the best set measured. For bounded interactions, we establish a sharp worst-case regret limit for deterministic scan-only selection; Ranked Validation satisfies the same upper bound once its record contains the additive maximizer. Across four checkpoints and thirteen candidate populations, it improves pooled exact selection over greedy and random validation under matched query budgets. On the half-pruning population, four extra evaluations recover 29.0% of the lost reference-answer log probability. A set selected once and reused across held-out questions adds little beyond a calibration-selected single-edit baseline.

1 INTRODUCTION

Sparse mixture-of-experts (MoE) models route each token to a small set of experts, which keeps the computation per token low as the parameter count grows (Jha et al., 2026). The router picks those experts from the token's hidden state, and that choice determines the hidden state every block above receives, so the decision at one layer shapes the input distribution of all the layers that follow.

Making such a model smaller means changing what the router selects. Practitioners remove experts they judge unimportant (Lu et al., 2024), restrict how many experts each token may use (Yue et al., 2025; Guo et al., 2025), or repair altered routes. Each intervention is an edit applied to one block. Scoring edits one at a time supplies a ranking, but does not determine what they contribute together in the same forward pass. We ask how useful this ranking remains and where a small budget of joint evaluations should go.

We test the single-edit scores directly and find that they can fail. One edit changes the hidden states entering the blocks above it, so a second edit operates on a different input than the one its score was measured on. We decompose the utility of an edit set into its single-edit contributions and an interaction term, and measure both. Consider a one-shot pruning rule (Lasby et al., 2026) that weighs each expert by how strongly the router selects it and by how large its output is, and deletes half of the experts in every layer. Figure 1 shows a repair with eight candidate restorations. The three restorations with the highest single-edit scores together lose reference-answer log probability, and the best set of three comes only fourth in the order of summed single-edit scores. In this example, combining the highest-scoring edits makes the pruned model worse under the measured utility.

To interpret this interaction, we study a local expansion around the intact model. Under additive logit responses, the quadratic pair term is the negative Fisher inner product of the edits' logit shifts. Nearly Fisher-orthogonal shifts contribute little at this order, while higher-order terms and non-additive responses contribute further interactions. The quadratic terms depend on the question being scored, motivating a comparison between selection per question and reuse of a single set across questions.

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.

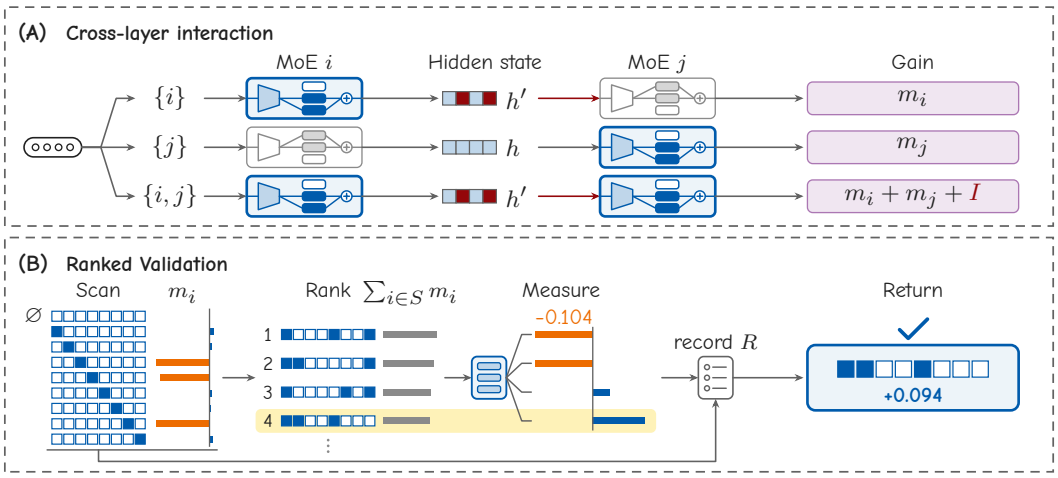

Figure 1: Interacting routing edits. (A) Upstream edit i shifts downstream hidden states, changing edit j's gain from its isolated value. Bold outlines mark edited blocks; m_i and m_j are single-edit gains, I their interaction. (B) Restoring experts after pruning half the experts, with at most three restorations. Filled cells mark edits; nine scan and four joint evaluations share one return record. Queue bars show additive scores; other bars show reference-answer gains over the unedited set (nats), with separate scales per group. The fourth mask is optimal among 93 enumerated sets.

We propose Ranked Validation. It measures the empty set and every single edit, orders the larger candidate sets by their summed single-edit scores, measures the leading sets in that order, and returns the best set it has measured. The scan supplies the ordering, and keeping the empty set in the record prevents the returned utility from falling below the unedited model on the same decision. Each utility evaluation uses the reference answer and runs the model once, so the budget is counted in model runs.

For bounded interactions, we establish a sharp worst-case regret limit for deterministic rules that read only the scan. Ranked Validation satisfies the same upper bound once the additive maximizer has been measured. Under the additional assumption that interactions, conditional on the scan, are independent draws from a common integrable distribution, the additive order maximizes expected returned utility at a fixed budget.

We evaluate the rule on four checkpoints across thirteen candidate populations, covering natural pruning, restoration, and route repair. Ranked Validation improves pooled exact selection over greedy and random validation at matched query budgets, although either comparator can lead at some budgets on individual populations. On the half-pruning population, four extra evaluations recover 29.0% of lost reference-answer log probability. The gain is largely per-question: a set selected on calibration questions adds little beyond the single-edit baseline on held-out questions, and a routing configuration that helps one benchmark can fail on another.

2 Related Work

Expert compression and routing. Expert pruning and budget allocation reduce sparse-inference cost by changing the expert pool (Lasby et al., 2026; Zhang et al., 2026a; Li et al., 2026). Other approaches merge experts, recalibrate routing, or steer behavior through selected routes (Huang et al., 2026; Zhang et al., 2026b; Ye et al., 2026). Our selection problem fixes the candidate edits and asks which combinations retain measured task utility under a shared evaluation budget.

Interaction-aware objectives. Interaction-aware compression uses joint expert effects or coalition values to move beyond isolated importance scores (Tseng et al., 2026; Zhang, 2026). The Möbius decomposition represents these interactions, while sparse transforms estimate them from evaluations (Grabisch et al., 2000; Erginbas et al., 2026). We use the exact decomposition to iden-

tify additive and pairwise selection errors, and relate the pair term to the softmax Fisher metric under local, additive-logit assumptions (Martens, 2020).

Budgeted selection. Greedy optimization, combinatorial Bayesian optimization, and costly inspection address candidate selection with limited observations (Ward & Feldman, 2026; Wan et al., 2021; Fisch et al., 2026). Their guarantees depend on structure in the objective or inspection model. We compare selectors under common evaluation caps against the exhaustive optimum, without assuming submodularity for routing utilities.

Tracing intervention effects. Activation patching and causal mediation trace effects through model states, including MoE experts and routing (Heimersheim & Nanda, 2024; Lu et al., 2026; Zhu, 2026). Our clamping controls test which downstream state reproduces isolated edit effects. On the measured examples, fixing the complete block input recovers those effects, whereas fixing the router input alone does not.

3 Problem Setting

3.1 Routing edits and the set function

A sparse MoE block scores experts from its input, routes each token to its top κ experts, and mixes their gate-weighted outputs. A routing edit changes the selected experts or their gate weights at one block, for the scored token of one decision. An edit set therefore applies once per decision. A decision fixes the model weights, the input, the candidate list $\mathcal{I} = \{1, \dots, n\}$, and the edit budget k, which is the largest number of edits a set may contain. A feasible set is any $S \subseteq \mathcal{I}$ with $|S| \le k$. Deterministic execution assigns every feasible set one log probability of the reference answer (an option label for multiple-choice questions) at the scored next-token position,

$$Y(S) = \log q_S(y), \qquad S^\star \in \operatorname*{arg\,max}_{S \subseteq \mathcal{I},\ |S| \le k} Y(S), \tag{1}$$

where q_S is the scored distribution under the edits in S and y is the reference answer. The regret of a selected set $\widehat{S}$ is $Y(S^\star) - Y(\widehat{S})$. Selection is exact at zero regret, and harmful when $Y(\widehat{S}) < Y(\varnothing)$.

3.2 Additive scores and interaction

Write $m_i = Y(\{i\}) - Y(\varnothing)$ for the single-edit gain. The Möbius coefficient of a nonempty set T, its Harsanyi dividend (Grabisch et al., 2000), is

$$a_T = \sum_{U \subseteq T} (-1)^{|T|-|U|} Y(U). \tag{2}$$

Every set function on $\mathcal{I}$ decomposes as

$$Y(S) - Y(\varnothing) = \sum_{i \in S} m_i + I(S), \qquad I(S) = \sum_{T \subseteq S,\ |T| \ge 2} a_T. \tag{3}$$

The additive score $\widehat{Y}(S) = Y(\varnothing) + \sum_{i \in S} m_i$ errs on S by exactly the interaction $I(S)$, and the shortfall of an order that assembles S telescopes to the same total for every order. A set is harmful precisely when $I(S) < -\sum_{i \in S} m_i$. Because the empty set satisfies the edit budget, the additive argmax has nonnegative summed gain, so its harm requires a negative interaction.

Proposition 1 (Third-order error of pairwise selection)**.** *Let $k = 3$, and let $\widetilde{Y}$ be the pairwise model fitted on the empty set, every single edit, and every pair. Then $\widetilde{Y} = Y$ on every set of size at most two, and $Y(S) = \widetilde{Y}(S) + a_S$ on every triple. Every harmful selection of the pairwise argmax $\widetilde{S}$ is therefore a triple with $a_{\widetilde{S}} < Y(\varnothing) - \widetilde{Y}(\widetilde{S}) \le 0$, and $Y(S^\star) - Y(\widetilde{S}) \le 2\eta_3$ with $\eta_3 = \max_{|S|=3} |a_S|$.*

For distinct edits i, j, ℓ, the third-order interaction measures the change in a pair dividend after applying edit ℓ. The dividend of the triple equals the dividend of the pair under edit ℓ, less the dividend of the pair alone.

Table 1: Exhaustive populations at $k = 3$. n counts candidate edits; median $R^\star$ counts extra evaluations until Ranked Validation reaches an optimum.

Checkpoint	Edit	n	Decisions	Median $R^\star$
Held-out pruning and restoration				
Qwen3.5-35B-A3B-Base	Pruning	5	768	4
Qwen3.5-35B-A3B-Base	Pruning	8	1,024	18
Qwen3.5-35B-A3B-Base	REAP restore	8	1,024	18.5
Qwen3.6-35B-A3B	Pruning	5	640	4
Qwen3.6-35B-A3B	Pruning	8	640	11
Development swap restoration				
Qwen3.6-35B-A3B	Boundary 1	3–5	59	7
Qwen3.6-35B-A3B	Boundary 2	3–5	523	4
Instella-MoE-16B-A3B	Boundary	3–5	41	3
Instella-MoE-16B-A3B	Distance	3–5	28	5.5
Held-out swap restoration				
Qwen3.5-35B-A3B-Base	Boundary	3–5	368	6
Qwen3-30B-A3B-Base	Boundary	3–5	261	2
Instella-MoE-16B-A3B	Boundary	3–5	347	6
Instella-MoE-16B-A3B	Distance	3–5	105	3

3.3 Evaluation budget and populations

An evaluation runs the model once on one edit set and returns its utility. The common scan spends $n + 1$ evaluations on the empty set and the single edits, and budgeted rules add at most t further evaluations. Returned sets satisfy the edit budget, and the permutation Shapley adapter can also measure larger coalitions. A *measured* rule returns the highest-utility feasible set in its record, which includes $\varnothing$, and a *predictor* selects from its fitted utility model. Every measured set is billed, the pairwise fit is billed separately, and each evaluation requires the reference answer.

Table 1 lists the candidate populations. Natural pruning candidates remove one selected expert and reroute to the next-ranked one, and restoration candidates return one expert removed by Router-weighted Expert Activation Pruning (REAP) (Lasby et al., 2026). Swap-restoration populations perturb one route per example with an outcome-blind swap, and candidates restore native routes at later layers that the swap changed. Boundary chooses the root swap with the smallest router-score gap, and Distance maximizes its local routed-output displacement. Development populations fixed the rules, and the held-out populations evaluate the fixed rules.

4 Ranked Validation

Algorithm 1 specifies Ranked Validation. After the single-edit scan it enumerates every multi-edit set, sorts them by decreasing additive score, and measures the leading sets up to the budget. A fixed candidate order breaks ties, and no model is fitted.

Algorithm 1 Ranked Validation

Require: candidates $\mathcal{I}$, edit budget k, extra evaluations t, utility oracle Y

$R \leftarrow \{(\varnothing, Y(\varnothing))\} \cup \{(\{i\}, Y(\{i\})) : i \in \mathcal{I}\}$ ▷ scan, $n + 1$ evaluations

$m_i \leftarrow Y(\{i\}) - Y(\varnothing)$ for every i; $\widehat{Y}(S) \leftarrow Y(\varnothing) + \sum_{i \in S} m_i$

$L \leftarrow$ the sets with $2 \le |S| \le k$, sorted by decreasing $\widehat{Y}(S)$

for $s = 1, \ldots, \min(t, |L|)$ **do**

  $R \leftarrow R \cup \{(L_s, Y(L_s))\}$ ▷ one evaluation

end for

return S from $(S, y) \in \arg\max_{(T,z) \in R} z$

4.1 Information limits and evaluation order

The scan identifies the single-edit gains, and it leaves the interaction unrestricted.

Theorem 1 (Information frontier). *Consider utilities whose interaction on any feasible set stays within ε of zero, with more candidates than the edit budget and an edit budget of at least two. Every deterministic selector that reads only the scan has worst-case regret at least twice ε, and the additive argmax attains that bound. Every deterministic measured selector that leaves some size-k set unmeasured beyond the scan also has worst-case regret at least twice ε.*

The adversary assigns single-edit gains ε, measured interactions $-\varepsilon$, and an unmeasured size-k interaction $+\varepsilon$. Every measured rule whose first extra query is the additive argmax, Ranked Validation included, attains this bound, keeps its returned utility at or above the empty-set value, and improves as the budget grows. Ranked Validation is exact precisely when the budget reaches the position of the first optimal set in the sorted list, or, at zero extra cost, when the scan has already measured an optimal set.

Proposition 2 (Additive order under independent interactions). *Conditional on the scan, let the interactions $I(S)$ of the feasible sets with $|S| \geq 2$ be independent draws $Z \sim G$ with $\mathbb{E}|Z| < \infty$, and let each evaluation cost $c > 0$. The Weitzman reservation value of set S is $z_S = \widehat{Y}(S) + w_G(c)$, with one constant $w_G(c)$ for every set, so the optimal inspection order of the Pandora problem (Fisch et al., 2026) is the additive order, and the optimal rule stops once the best measured value reaches the next $\widehat{Y}(S) + w_G(c)$. With a fixed budget $t \leq |L|$ and current best M_0, measuring the t sets of largest $\widehat{Y}(S)$ maximizes $\mathbb{E}[\max\{M_0, Y(S_1), \ldots, Y(S_t)\}]$ over all adaptive policies.*

Independence is a sufficient condition, and sets sharing edits can have correlated interactions. A diagnostic stops the sweep once the best measured utility reaches the next additive score, and the fixed-budget protocol does not use it. Interleaved Validation schedules two greedy proposals followed by one ranked proposal under a shared evaluation record.

4.2 A local Fisher interpretation

Fix a question with intact logits z_*, distribution p_*, and softmax Fisher matrix $F_* = \mathrm{diag}(p_*) - p_* p_*^\top$ (Martens, 2020). Damage shifts the logits by d, and edit i shifts them by v_i. Under additive logit responses, set S has residual logit shift $r_S = d + \sum_{i \in S} v_i$ and output distribution p_S.

Proposition 3 (Fisher form of the interaction). *For the utility $Y(S) = -\mathrm{KL}(p_* \,\|\, p_S)$,*

$$Y(S) = -\tfrac{1}{2}\, r_S^\top F_*\, r_S\, (1 + \theta_S), \qquad |\theta_S| \leq \tfrac{2}{3} \|r_S\|_\infty e^{2\|r_S\|_\infty}, \tag{4}$$

so the quadratic term has pair dividend $a_{ij} = -v_i^\top F_* v_j$ and no dividend of order three or higher.

The expansion comes from the log-sum-exp Bregman divergence, and its third derivative governs the bound. For reference-answer utility $\log p_S(y)$, the logit Hessian at the expansion point is also $-F_*$, which gives the same quadratic pair term, and higher-order remainders and non-additive logit responses contribute additional interactions. Nearly Fisher-orthogonal responses give small quadratic pair terms, and question-dependent terms can cancel across questions, which leaves a shared set only their persistent part.

The sign of a pair term is therefore a statement about the geometry of two edits at one question. Two restorations that move the output in opposed directions cancel, and a set that helps on one question can lose on another when the same two edits move in directions that question does not separate. The ordering therefore has to be evaluated.

5 Experiments

We ask four questions. Does ordered evaluation reach the limit that the theory sets for the scan-limited selector, and does the answer hold across checkpoints? What does the rule recover on a model that has lost half of its experts? Which part of the gain survives when one edit set is chosen once and reused on questions it never saw? When a standard task supplies a routing-mass calibration, does that calibration travel to panels of a different kind?

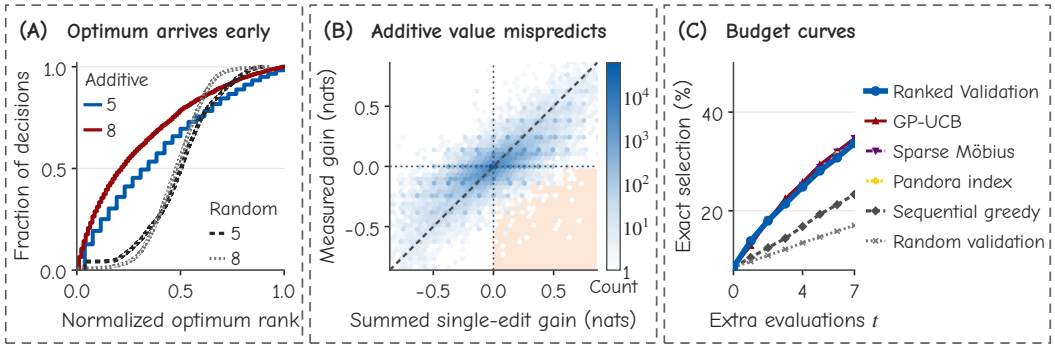

Figure 2: Additive ranking and measured gains. (A) Distribution of the optimum's additive rank divided by feasible-set count, versus cardinality-matched random order. (B) Summed single-edit versus measured gain (nats) for every multi-edit set in five natural populations; shading marks positive summed but negative measured gains. (C) Exact selection versus extra evaluations, pooled over three eight-candidate populations.

5.1 Experimental setup

Evaluation. Every rule replays on exhaustively recorded utilities at $k = 3$, so all rules face the same decisions, the same candidate lists, and one evaluation accounting. Primary metrics are exact selection, the fraction of decisions whose returned set attains the best feasible utility; mean regret in millinats; and harm, the fraction of returned sets scoring below the unedited model. Paired differences use question-cluster resampling, and Appendix D carries the resampling and per-population tables.

Models and populations. The thirteen populations of Table 1 span four checkpoints, Qwen3.6-35B-A3B (Qwen Team, 2026), Qwen3.5-35B-A3B-Base, Qwen3-30B-A3B-Base, and Instella-MoE-16B-A3B (Liu et al., 2026), with utilities on ARC-Challenge, HellaSwag, MMLU, and WikiText-103. In the five natural populations, pruning candidates act at five or eight routing layers and restorations at eight layers of Qwen3.5-35B-A3B-Base after REAP removes half of every layer, and root swaps search the first half of routing layers. Four populations developed the rules and the other nine evaluate them. On the eight-candidate populations $t = 4$ measures 13 of 93 feasible sets, and measuring the scan and every pair takes 37 evaluations, leaving every triple coefficient unknown. That budget anchors the comparisons below.

Baselines. Controls split into predictive rules, which fit a utility model and select from it, and measured rules, which select from what they have already run. The predictive set covers the additive argmax, the pairwise fit, a budgeted second-order proposal, and coalition adapters; the measured set covers greedy, random validation, local search, and four surrogate searches under one cap. Appendix C specifies the adapters, and Appendix F reports component experiments for the published edit families.

5.2 Main results

Ordered evaluation buys exact selection at a cost far below enumeration, and the advantage holds across checkpoints. Ranked Validation exceeds greedy and random validation at every budget from one to seven extra evaluations in the pooled paired analysis, and at four extra evaluations it exceeds greedy on 12 of 13 populations and ties it on the Instella boundary development population (Figure 3A). Greedy or random validation leads at some budgets on five of eight swap-restoration populations. The order itself carries the gain, since a cardinality-matched random order trails Ranked Validation by 11.1 points at seven extra evaluations, and Figure 2A shows the optimum arriving early in the additive order for five- and eight-candidate decisions. Figure 2B gives the reason the scan cannot be trusted alone, with a shaded quadrant holding sets whose summed gain is positive and whose measured gain is a loss.

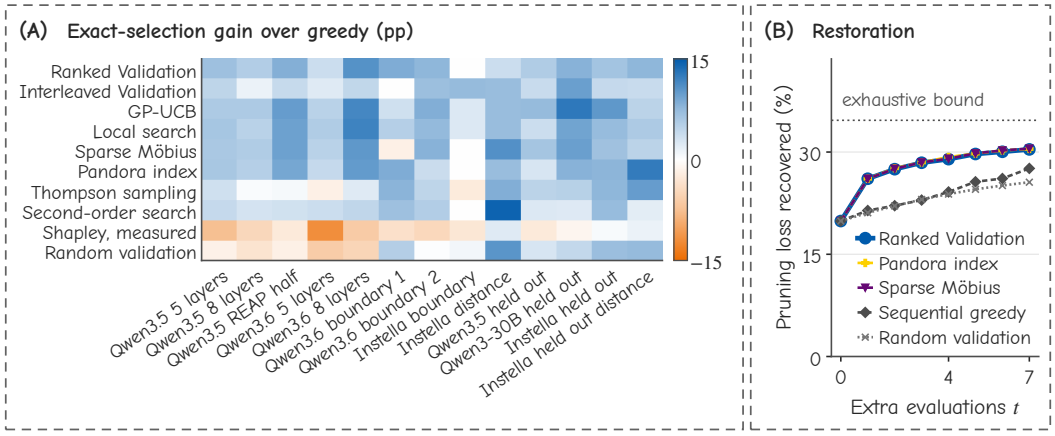

Figure 3: Selection across populations. (A) Exact-selection gain over sequential greedy at $t = 4$ across thirteen populations (percentage points). (B) Fraction of half-pruning loss recovered by each rule's restorations, versus the exhaustive bound.

What a rule returns decides whether it can harm. A budgeted second-order proposal, which prices expert pairs in a gate-weighted scoring function and applies its model argmax, is the one shared-cap rule with nonzero harm, rising to 14.5% of decisions at seven extra evaluations, where every rule that returns its best measured set keeps $\varnothing$ in its record and never falls below the unedited model. Returning the best measured set under the same search adds 9.9 exact-selection points at four extra evaluations, since a measurement replaces the model's estimate, whose error is the omitted interaction, by the true utility. Fitting the interaction adds little beyond the order: the question-cluster bootstrap finds the lead of Hamming-kernel GP-UCB over Ranked Validation significant from four extra evaluations, and local search, adaptive sparse Möbius, and the Pandora index land in the same narrow band. Selector computation stays far below one model forward pass per evaluation, so the budget that matters is the number of model runs.

5.3 Restoration after pruning

The rule reuses the scan to repair a model that has lost half of its experts. Half pruning of Qwen3.5-35B-A3B-Base lowers reference-answer log probability by 0.382 nats, and accuracy on 768 items with option scores falls from 80.9% to 79.0%. No set of at most three of eight restorations undoes this loss: the exhaustive optimum, which bounds what any rule can recover here, reaches 80.1% accuracy. Within that bound, Ranked Validation obtains 29.0% of the lost log probability at four extra evaluations, above greedy at 24.1% and close to the Pandora index and the adaptive sparse Möbius transform, so on this population the ordering and the fitted surrogates reach similar recovery (Figure 3B). Only the additive argmax and the second-order proposal, the two rules that apply a predicted set, return harmful restorations, which repeats the failure of Figure 1 at population scale.

The gain is per-question, and it does not transfer for free. Over every set of at most three restorations, the pair-to-single ratio has median 0.55 within a question against 0.15 for the globally averaged utility, so the interaction that the ordering exploits is largely specific to the question being scored. Ranking per question gains 0.0087 nats over the best single edit at four extra evaluations (standard error 0.0009), and its difference from per-question greedy is -0.0020 nats (standard error 0.0009), a small deficit outside the pooled exact-selection advantage. A rule that must commit to one set before seeing the question therefore has less to work with, which the next subsection measures directly.

5.4 Transfer of one set across questions

A shared set keeps little of the interaction it was selected on and adds little beyond the best single edit. We group 8 to 64 questions by benchmark, MMLU subject, HellaSwag activity, or random assignment, select on half of each cluster, and evaluate on the held-out half, where one calibration evaluation averages a set over that half at one forward pass per question. In every cell the persistent

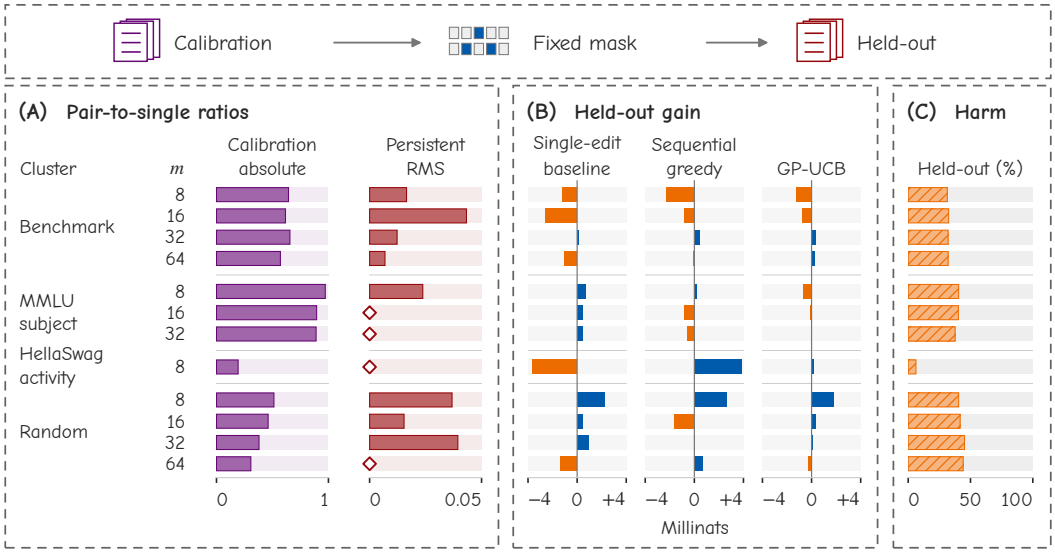

Figure 4: Transfer at $t = 4$; m is cluster size in questions. The set selected on the calibration half is reused unchanged on the held-out half. (A) Calibration and cross-half persistent pair-to-single ratios; hollow diamonds mark negative cross-half energy truncated to zero. (B) Held-out gains (millinats) over calibration-selected baselines; the single-edit baseline includes the empty set. (C) Hatching marks replicate-mean held-out harm relative to the pruned model.

ratio stays far below the calibration ratio (Figure 4A), and the held-out differences between Ranked Validation and the single-edit baseline, greedy, or GP-UCB stay within 0.0039 nats (Figure 4B). The cross-half product credits only interaction that recurs on both halves, so a set chosen once has little interaction to use. On the global pool, the set that Ranked Validation selects with four extra evaluations gains 0.092 nats over the pruned model on 768 held-out questions (standard error 0.009), and its 0.001 nats over the calibration-selected single edit (standard error 0.001) is indistinguishable from zero.

5.5 Calibrating a routing mass and carrying it across panels

A calibration chosen on one panel carries a selection cost, and the direction it fixes does not hold on every panel. Power-mass calibration reweights the retained gate mass by an exponent α, where $\alpha = 0$ reproduces the native four-expert output and $\alpha = 1$ reproduces the reference eight-expert normalization. Maximizing official accuracy over 256 HellaSwag development documents selects $\alpha = 0.25$ at 79.7% against 79.7% for the best arm. On 512 held-out documents of the same task the selected exponent changes accuracy by -0.8 points relative to $\alpha = 0$ (one-sided exact sign test $p = 0.819$), so the development ranking does not repeat on fresh questions at this sample size. Across the full 10,042 documents of the validation split the same exponent gains 0.84 points over the native four-expert output ($p < 0.001$), which is the direction the development stage chose. Raising the reference normalization from four to eight experts gains 0.47 points on that split ($p = 0.045$), and raising it to sixteen changes accuracy by -1.62 points, so more reference mass does not accumulate a gain.

The calibration travels to one independent task and stops at the other, which is the transfer behaviour the per-question results predict. On ARC-Challenge the selected exponent gains 2.90 points over the native four-expert output ($p = 0.004$), and on PIQA the same comparison gives $p = 0.389$ on the length-normalized metric. A fixed routing configuration therefore carries its benefit to some panels and not to others, which is a reason to confirm it on the panel being served.

5.6 Interaction and state analysis

The interaction has a local form and a measured carrier. The mean absolute pair and triple dividends of the globally averaged utility are 0.15 and 0.09 times the mean absolute single-edit effect, so pair terms carry most of the departure from additivity, and linearity of the Möbius transform carries each interaction's signed mean into the shared objective. A first-order transport model of the downstream gain predicts the sign of the measured pair interaction on 49.6% of 14,350 pairs, at the 50.0% chance level for a sign prediction, so the transport direction alone does not determine the sign and interactions have to be measured. Clamping identifies the carrier: patching the router input reproduces 0.0% of 95 isolated downstream effects and patching the complete block input reproduces all of them, and on Instella's two-stream blocks primary-stream clamping recovers 9.1% of 44 effects against 0.0% for the auxiliary stream. The effect travels through the block input, which is the same channel the ordering assumes when it treats later edits as acting on changed states.

A selector should report what it measured, because a measured set and a predicted set differ exactly by the interaction, and that difference reaches a sign change on a small share of decisions. The scan is also worth its cost at every population studied, since it supplies both the ordering that carries the selection gain and the empty-set entry that bounds the returned utility from below.

6 Limitations and Conclusion

Single-edit scores do not determine what a set of routing edits contributes, because one edit changes the hidden states that every later block reads. Ranked Validation turns that observation into an evaluation order: measure the scan, sort the larger sets by their summed single-edit scores, and spend the remaining budget on the leading sets in that order. The order attains the worst-case limit that Theorem 1 sets for any rule reading only the scan, and it improves pooled selection over greedy and random validation.

The scope of that claim is bounded by the evidence. Evaluation covers four checkpoints, deterministic execution, and at most three edits, and the primary utility needs reference answers. The REAP checkpoint's exact revision is absent from retained metadata, the label-free panel runs Prefix Validation rather than Ranked Validation, so Ranked Validation without reference answers is untested, shared-set transfer can lower individual-question utility, and larger pools raise the scan cost.

Reproducibility Statement

Section 3 and Appendix B specify populations, candidate edits, utilities, eligibility rules, and checkpoint and dataset revisions. Algorithm 1 and Appendix C define the selection rules and evaluation budgets. The supplementary `recorded-utility-replay.zip` contains the complete response cubes for Table 1 and the original selector-analysis entry point. Appendix A proves the main results; further proofs accompany the results in Appendices C–E.

AI Use Statement

Language models assisted literature search, method implementation, figure development, mathematical formulation and proof checking, result interpretation, and manuscript editing. The authors verified the sources, derivations, results, and statements and take responsibility for the paper.

References

Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 462–471. PMLR, 2018. URL https://proceedings.mlr.press/v80/baptista18a.html.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012. URL https://jmlr.org/papers/v13/bergstra12a.html.

Endre Boros and Peter L. Hammer. Pseudo-Boolean optimization. *Discrete Applied Mathematics*, 123(1–3):155–225, 2002. doi: 10.1016/S0166-218X(01)00341-9.

Xing Chen and Hengshuai Yao. Training-free halving of activated experts in fine-grained mixture-of-experts models. *arXiv preprint arXiv:2609.04575*, 2026. URL https://arxiv.org/abs/2609.04575.

Yigit Efe Erginbas, Justin Singh Kang, Elizabeth Polito, and Kannan Ramchandran. Adaptive sparse Möbius transforms for learning polynomials. *arXiv preprint arXiv:2602.06246*, 2026. URL https://arxiv.org/abs/2602.06246.

Adam Fisch, Shubhendu Trivedi, Fantine Huot, William W. Cohen, Michael Kaisers, Mirella Lapata, Kate Larson, and Jacob Eisenstein. Pandora's AI model routing box: Efficient allocation with costly value estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. URL https://openreview.net/forum?id=yxwL9oTlWU.

Shi Fu, Youming Qiao, Dacheng Tao, Zongqi Wan, and Qixin Zhang. Poisson exchange beyond submodularity: Effective approximation algorithms for offline and online subset selection over matroids. *arXiv preprint arXiv:2609.24569*, 2026. URL https://arxiv.org/abs/2609.24569.

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000. doi: 10.1287/moor.25.2.157.12225.

Yongxin Guo, Zhenglin Cheng, Xiaoying Tang, Zhaopeng Tu, and Tao Lin. Dynamic mixture of experts: An auto-tuning approach for efficient Transformer models. In *International Conference on Learning Representations (ICLR)*, pp. 79643–79672, 2025. URL https://openreview.net/forum?id=T26f9z2rEe.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

Canbin Huang, Tianyuan Shi, Xiaojun Quan, Jingang Wang, Jianfei Zhang, and Qifan Wang. When model merging breaks routing: Training-free calibration for MoE. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=1ORZJtadQl.

Saurav Jha, Maryam Hashemzadeh, Ali Saheb Pasand, Ali Parviz, Min-Joong Lee, and Boris Knyazev. REAM: Merging improves pruning of experts in LLMs. *arXiv preprint arXiv:2604.04356*, 2026. URL https://arxiv.org/abs/2604.04356.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger (eds.), *Complexity of Computer Computations*, The IBM Research Symposia Series, pp. 85–103. Plenum Press, New York, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. REAP the experts: Why pruning prevails for one-shot MoE compression. In *International Conference on Learning Representations (ICLR)*, pp. 146883–146912, 2026. URL https://openreview.net/forum?id=ukGxWd2aDG.

Lie Li, Wen Li, Junxiao Shen, and Guosheng Hu. MAPLE: MoE adaptive plug-and-play layer-wise expert allocation. *arXiv preprint arXiv:2608.15299*, 2026. URL https://arxiv.org/abs/2608.15299.

Jiang Liu, Sudhanshu Ranjan, Prakamya Mishra, Yonatan Dukler, Gowtham Ramesh, Jialian Wu, Ximeng Sun, Wen Xie, Chaojun Hou, Vikram Appia, Zhenyu Gu, Zicheng Liu, and Emad Barsoum. Instella-MoE technical report. *arXiv preprint arXiv:2609.00791*, 2026. URL https://arxiv.org/abs/2609.00791.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.334. URL https://aclanthology.org/2024.acl-long.334/.

Yuetian Lu, Ali Modarressi, Yihong Liu, and Hinrich Schütze. Expert-aware causal tracing of factual recall in sparse MoE language models. *arXiv preprint arXiv:2606.03780*, 2026. URL https://arxiv.org/abs/2606.03780.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. Qwen blog, 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, pp. 1015–1022, 2010. URL https://icml.cc/Conferences/2010/papers/422.pdf.

Alex M. Tseng, Prannay Kaul, Luca Zancato, Wei Xia, and Stefano Soatto. Higher-order pruning of experts in mixture-of-experts language models. *arXiv preprint arXiv:2609.18916*, 2026. URL https://arxiv.org/abs/2609.18916.

Xingchen Wan, Vu Nguyen, Huong Ha, Binxin Ru, Cong Lu, and Michael A. Osborne. Think global and act local: Bayesian optimisation over high-dimensional categorical and mixed search spaces. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10663–10674. PMLR, 2021. URL https://proceedings.mlr.press/v139/wan21b.html.

Justin Ward and Moran Feldman. On maximizing a weakly submodular function over a matroid constraint via the greedy algorithm. *arXiv preprint arXiv:2609.05008*, 2026. URL https://arxiv.org/abs/2609.05008.

Martin L. Weitzman. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, 1979. doi: 10.2307/1910412.

Juntong Wu, Yifei Liu, Junyi Chen, Siqi Fan, Chaoran Feng, Minghao Li, Liujie Zhang, Weihang Chen, and Li Yuan. ExFold: Unified expert folding for training-free MoE prefill-decode acceleration. *arXiv preprint arXiv:2608.24938*, 2026. URL https://arxiv.org/abs/2608.24938.

Zukang Xu, Zhixiong Zhao, Xing Hu, Jiangyong Yu, Houji Wen, Jun Li, Zhe Jiang, and Dawei Yang. ACE: Adaptive calibration-free expert skipping for MoE-based LLMs. *arXiv preprint arXiv:2609.05228*, 2026. URL https://arxiv.org/abs/2609.05228.

Charles Ye, Bo Yuan, and Lee Sharkey. Polysemantic experts, monosemantic paths: Routing as control in MoEs. In *ICLR 2026 Workshop on Representational Alignment (Re-Align)*, 2026. URL https://openreview.net/forum?id=IfK2ygp25S.

Tongtian Yue, Longteng Guo, Jie Cheng, Xuange Gao, Hua Huang, and Jing Liu. Ada-K routing: Boosting the efficiency of MoE-based LLMs. In *International Conference on Learning Representations (ICLR)*, pp. 89619–89635, 2025. URL https://openreview.net/forum?id=9CqkpQExe2.

Yuhao Zhang. SHAPE: Coalition-aware expert pruning for sparse mixture-of-experts LLMs. *arXiv preprint arXiv:2606.09886*, 2026. URL https://arxiv.org/abs/2606.09886.

Zeliang Zhang, Nikhil Ghosh, Jiani Liu, Bin Yu, and Xiaodong Liu. Does a global perspective help prune sparse MoEs elegantly? *arXiv preprint arXiv:2604.06542*, 2026a. URL https://arxiv.org/abs/2604.06542.

Zhibo Zhang, Zhen Ouyang, Ling Shi, and Kailong Wang. RARE: Decoupling representation steering from expert routing in mixture-of-experts language models. In *Actionable Interpretability Workshop at COLM 2026*, 2026b. URL https://arxiv.org/abs/2608.21236.

Tianyang Zhu. From expert reduction to behavioral divergence: Tracing numerical state through sparse MoE inference. *arXiv preprint arXiv:2607.28097*, 2026. URL https://arxiv.org/abs/2607.28097.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proposition 1 (Third-order error of pairwise selection). *Let $k = 3$, and let $\widetilde{Y}$ be the pairwise model fitted on the empty set, every single edit, and every pair. Then $\widetilde{Y} = Y$ on every set of size at most two, and $Y(S) = \widetilde{Y}(S) + a_S$ on every triple. Every harmful selection of the pairwise argmax $\widetilde{S}$ is therefore a triple with $a_{\widetilde{S}} < Y(\varnothing) - \widetilde{Y}(\widetilde{S}) \le 0$, and $Y(S^\star) - Y(\widetilde{S}) \le 2\eta_3$ with $\eta_3 = \max_{|S|=3} |a_S|$.*

Proof. The pairwise model has coefficients $a_\varnothing$, $a_{\{i\}}$, and $a_{\{i,j\}}$ identified from the measured sets, so it reproduces every set of size at most two exactly and misses exactly the dividend a_S of a triple S. If the argmax $\widehat{S}$ of $\widetilde{Y}$ has size at most two, then $Y(\widehat{S}) = \widetilde{Y}(\widehat{S}) \ge \widetilde{Y}(\varnothing) = Y(\varnothing)$, which is not harmful. A harmful choice is therefore a triple with $Y(\varnothing) > Y(\widehat{S}) = \widetilde{Y}(\widehat{S}) + a_{\widehat{S}}$, and $\widetilde{Y}(\widehat{S}) \ge Y(\varnothing)$ gives $a_{\widehat{S}} < Y(\varnothing) - \widetilde{Y}(\widehat{S}) \le 0$. For the regret, $Y(S^\star) - Y(\widehat{S}) = \widetilde{Y}(S^\star) - \widetilde{Y}(\widehat{S}) + a_{S^\star}[|S^\star| = 3] - a_{\widehat{S}}[|\widehat{S}| = 3] \le 2\eta_3$, since $\widehat{S}$ maximizes $\widetilde{Y}$. A budgeted second-order model sets the unmeasured pair dividends to zero, so its error on a set is the sum of those omitted pair dividends and of the third-order dividend. □

A.2 PROOF OF THEOREM 1

Theorem 1 (Information frontier). *Consider utilities whose interaction on any feasible set stays within ε of zero, with more candidates than the edit budget and an edit budget of at least two. Every deterministic selector that reads only the scan has worst-case regret at least twice ε, and the additive argmax attains that bound. Every deterministic measured selector that leaves some size-k set unmeasured beyond the scan also has worst-case regret at least twice ε.*

Proof. Scan-only selectors. Fix a deterministic selector and the transcript $Y(\varnothing) = 0$, $Y(\{i\}) = \varepsilon$ for every i. Let R be the set it returns on this transcript. Choose a set $T \ne R$ with $|T| = k$, which exists since $n > k$. Define Y by $I(T) = \varepsilon$, $I(R) = -\varepsilon$ when $|R| \ge 2$, and $I(S) = 0$ otherwise; this keeps the transcript and satisfies $\varepsilon_k \le \varepsilon$, and the selector returns R. When $|R| \ge 2$, $Y(T) - Y(R) = (k+1)\varepsilon - (|R| - 1)\varepsilon \ge 2\varepsilon$; when $|R| \le 1$, $Y(T) - Y(R) = (k + 1 - |R|)\varepsilon \ge 2\varepsilon$. For the upper bound, let $\widehat{S}$ maximize $\widehat{Y}$ over $\mathcal{F}_k$; by Eq. (3), $Y(S^\star) - Y(\widehat{S}) = \widehat{Y}(S^\star) - \widehat{Y}(\widehat{S}) + I(S^\star) - I(\widehat{S}) \le 2\varepsilon$.

Measured selectors. Keep the transcript $Y(\varnothing) = 0$ and $Y(\{i\}) = \varepsilon$, and answer every multi-edit set the selector measures with interaction $-\varepsilon$. The answers depend only on the queried sets, so the deterministic selector follows one fixed query path; fewer than $\binom{n}{k}$ queries leave a set T of size k unqueried. Set $I(T) = +\varepsilon$ and $I(S) = 0$ on every other unqueried set; the transcript is unchanged and $\varepsilon_k \le \varepsilon$. The returned set R is measured, so $Y(R) = |R|\varepsilon - \varepsilon \le (k-1)\varepsilon$ when $|R| \ge 2$ and $Y(R) \le \varepsilon$ otherwise, and $Y(T) = (k+1)\varepsilon$ gives $Y(T) - Y(R) \ge 2\varepsilon$ in both cases. □

A.3 PROOF OF PROPOSITION 2

Proposition 2 (Additive order under independent interactions). *Conditional on the scan, let the interactions $I(S)$ of the feasible sets with $|S| \ge 2$ be independent draws $Z \sim G$ with $\mathbb{E}|Z| < \infty$, and let each evaluation cost $c > 0$. The Weitzman reservation value of set S is $z_S = \widehat{Y}(S) + w_G(c)$, with one constant $w_G(c)$ for every set, so the optimal inspection order of the Pandora problem (Fisch et al., 2026) is the additive order, and the optimal rule stops once the best measured value reaches the next $\widehat{Y}(S) + w_G(c)$. With a fixed budget $t \le |L|$ and current best M_0, measuring the t sets of largest $\widehat{Y}(S)$ maximizes $\mathbb{E}[\max\{M_0, Y(S_1), \dots, Y(S_t)\}]$ over all adaptive policies.*

Proof. After the scan, set S with $|S| \ge 2$ is a box with value $\widehat{Y}(S) + I(S)$, where $\widehat{Y}(S)$ is known and $I(S) \sim G$ independently across sets, and opening it costs c. Weitzman's reservation value z_S solves $\mathbb{E}[(Y(S) - z_S)^+] = c$. Writing $Y(S) = \widehat{Y}(S) + Z$ with $Z \sim G$ gives $\mathbb{E}_G[(Z - (z_S - \widehat{Y}(S)))^+] = c$. The left side is continuous and strictly decreasing in $z_S - \widehat{Y}(S)$ where it is positive, so the unique solution is $z_S - \widehat{Y}(S) = w_G(c)$, one constant for every set, and $z_S = \widehat{Y}(S) + w_G(c)$. Weitzman's

theorem (Weitzman, 1979) states that the optimal policy opens boxes in decreasing z_S and stops once the best opened value exceeds every unopened z_S; since $z_S - \widehat{Y}(S)$ is constant, decreasing z_S is decreasing $\widehat{Y}(S)$, which is the order of Algorithm 1, and the stopping condition compares the best measured value with the next $\widehat{Y}(S) + w_G(c)$. The scan values are known exactly and enter as opened boxes.

For the fixed budget, write $X_S = \widehat{Y}(S) + Z_S$ and let $V_t(m, A)$ be the largest expected value of $\max\{m, \text{measured values}\}$ with t measurements left among the unmeasured sets A. We show by induction on t that $V_t(m, A) = \mathbb{E}\max\{m, X_{S_1}, \dots, X_{S_t}\}$ for the t sets of A with the largest $\widehat{Y}$. For $t = 1$, $\mathbb{E}\max\{m, \widehat{Y}(S) + Z\}$ is nondecreasing in $\widehat{Y}(S)$. For $t > 1$, measuring S first and then following the induction yields $\mathbb{E}\max\{m, X_S, X_U\}$ for the top $t-1$ sets U of $A \setminus \{S\}$. If S lies among the top t sets this equals the claimed value. Otherwise U is the top $t-1$ of A, and coupling Z_S with $Z_{S'}$ for the t-th largest set S' gives $X_S \le X_{S'}$ pathwise, so measuring S is no better. Adaptivity adds nothing, since the induction already optimizes over every first choice and every continuation. □

A.4 Proof of Proposition 3

Proposition 3 (Fisher form of the interaction)**.** *For the utility* $Y(S) = -\mathrm{KL}(p_* \,\|\, p_S)$*,*

$$Y(S) = -\tfrac{1}{2}\, r_S^\top F_*\, r_S\,(1 + \theta_S), \qquad |\theta_S| \le \tfrac{2}{3}\|r_S\|_\infty e^{2\|r_S\|_\infty},$$

so the quadratic term has pair dividend $a_{ij} = -v_i^\top F_* v_j$ *and no dividend of order three or higher.*

Proof. Write $A(z) = \log \sum_c e^{z_c}$, so $\log p_z = z - A(z)\mathbf{1}$ and $\mathrm{KL}(p_* \,\|\, p_S) = A(z_* + r_S) - A(z_*) - p_*^\top r_S$, the Bregman divergence of A (Martens, 2020). The gradient of A is p_z and its Hessian is $\mathrm{diag}(p_z) - p_z p_z^\top = F_z$. Taylor's theorem with Lagrange remainder gives $\mathrm{KL} = \frac{1}{2} r^\top F_* r + \frac{1}{6} D^3 A(\xi)[r, r, r]$ for some $\xi = z_* + sr$, $s \in [0, 1]$. The third derivative is the third central moment of the coordinate r_c under p_ξ, so $|D^3 A(\xi)[r, r, r]| \le \max_c |r_c - \mathbb{E}_{p_\xi} r| \cdot \mathrm{Var}_{p_\xi}(r) \le 2\|r\|_\infty \mathrm{Var}_{p_\xi}(r)$. Since $p_\xi(c)/p_*(c) \le e^{2s\|r\|_\infty}$, $\mathrm{Var}_{p_\xi}(r) \le \mathbb{E}_{p_\xi}[(r - \mathbb{E}_{p_*} r)^2] \le e^{2\|r\|_\infty} r^\top F_* r$, which gives the bound on θ_S in Eq. (4). For the quadratic part $Q(S) = -\frac{1}{2}(d + \sum_{i \in S} v_i)^\top F_* (d + \sum_{i \in S} v_i)$, the Möbius coefficient of $\{i, j\}$ is $Q(\{i, j\}) - Q(\{i\}) - Q(\{j\}) + Q(\varnothing) = -v_i^\top F_* v_j$, and every coefficient of order three or higher vanishes because Q is a polynomial of degree two in the indicators of S (Grabisch et al., 2000). □

B Populations and Protocol

Checkpoints and data. Recorded checkpoint revisions are `995ad96e` for `Qwen/Qwen3.6-35B-A3B`, `fd6d4912` for `amd/Instella-MoE-16B-A3B-Base`, and `1b75feb7` for `Qwen/Qwen3-30B-A3B-Base`. Dataset revisions begin with `c30699e8356d` for `cais/mmlu`, `210d026faf99` for `allenai/ai2_arc`, `218ec52e09a7` for `Rowan/hellaswag`, and `b08601e04326` for `Salesforce/wikitext`. These revisions are recorded in the model metadata and evaluation records. The Qwen3.5-35B-A3B-Base records identify a saved local checkpoint by model name; its exact checkpoint revision is absent from the retained metadata. Dataset records retain the full revisions and source digests.

Utility and eligibility. For the multiple-choice benchmarks the utility is the log probability of the correct option label at the scored position; for WikiText-103 it is the log probability of the next token. Per-decision edits act at the final prompt token, whose logits score the reference continuation. Candidate discovery completes before any utility is read. Forced supports recompute gate weights from the current router scores, applying each checkpoint's native normalization and scaling. A swap-restoration example is eligible when its perturbation changes at least three later routing events, and at most five of those events are kept as candidates, in depth order; Table 1 reports eligible decisions, and rates use them as the denominator. The natural and REAP populations have a fixed candidate count and keep every example.

Swap-restoration populations. At each eligible layer, the root exchange replaces the selected expert with the lowest router score by the highest-scoring inactive expert. Boundary selects the exchange with the smallest absolute score gap in the first half of routing layers. Distance selects the exchange with the largest Euclidean displacement of the gate-weighted expert output at the captured input, including weight renormalization; score gap and layer order break ties. Candidate discovery then records the first five later layers whose expert supports change, requiring at least three. Both root rules use model states and expert outputs without reference-answer utility. Every measured set keeps the root swap. Registered layers use the native support when selected and the corrupted discovery-pass support otherwise; unregistered layers route from their current inputs.

The Qwen3.6 Boundary 1 population evaluates the first sixteen questions from each of the four frozen benchmark panels. Boundary 2 reevaluates that panel and adds 128 questions per benchmark from the expansion panel. Both apply the same root and candidate rules, and each retains its eligible decisions as listed in Table 1. These development populations include repeated question inputs. The other two development populations apply Boundary and Distance to Instella on the first panel.

Natural candidates. The expert-pruning candidates sit at the routing events with the smallest gap between the κ-th and next-ranked router scores; the selected layers are ordered by depth and alternate lowest-gate and highest-gate removals, beginning with lowest-gate removal. Each replacement uses the highest-ranked inactive expert. The REAP candidates start from the checkpoint with half of the experts of every layer removed by REAP saliency, the mean over routed calibration tokens of the gate weight times the norm of the expert output (Lasby et al., 2026). At each layer the candidate restores the removed expert with the highest router score for the scored token in place of the lowest-weight selected expert, and the eight layers where the restored expert's score exceeds the replaced one's by the most are kept. Natural and REAP sets force supports at their selected candidate layers; other layers route from their current inputs under the active pruning mask. Option accuracy on the REAP population is read from the argmax option under the pruned, restored, and unpruned models.

Replay. For each decision in Table 1, the recorded response cube contains all 2^n subsets. Selection and optimum calculations use the feasible slice $|S| \leq 3$; permutation Shapley can also read larger recorded coalitions. Rules replay on these records; random validation averages fixed seeds per decision and quadratic Thompson sampling uses fixed seeds, derived from the example identifier. The four development populations fixed the rule set and the Interleaved schedule, and the held-out populations were registered afterwards.

Normalized optimum position. For the rank curves, let $N_{\text{all}} = \sum_{j=0}^{k} \binom{n}{j}$ and let $S_1, \ldots, S_{N_{\text{all}}}$ list every feasible set in decreasing additive score, with ties broken by the binary mask. The normalized position is

$$\frac{\min\{j : Y(S_j) \geq Y(S^\star) - 10^{-12}\}}{N_{\text{all}}}.$$

This position includes the empty set and singletons, including decisions whose scan contains an optimum. The query rank $R^\star$ in Table 1 is zero for those decisions and otherwise counts the first optimal multi-edit set in the post-scan list of length $\sum_{j=2}^{k} \binom{n}{j}$. The random-order control permutes sets within each cardinality and preserves the additive order's cardinality sequence.

Interaction ratios. Let $\bar{a}_T$ denote the Möbius coefficient of utility averaged over the global question pool. The reported global pair and triple ratios are

$$\frac{\binom{n}{r}^{-1} \sum_{|T|=r} |\bar{a}_T|}{n^{-1} \sum_i |\bar{a}_{\{i\}}|}, \qquad r \in \{2, 3\}.$$

The per-question ratio applies this mean-absolute aggregation to each question's coefficients and reports its median across questions. The selected-triple and optimum-interaction diagnostics use $\max_i |m_i|$ as their denominator, retaining decisions where it exceeds 10^{-12}.

Deployment protocol. Shared-set transfer uses the global pool of 10 fixed layer–expert pairs from Section 5. The pruning set and candidate pool are frozen using calibration questions from all four benchmark panels before held-out readout. Ranked Validation selects one set on 256 calibration

questions; the fixed set is evaluated on 768 disjoint questions. Each candidate restores its expert at every token. The selected set restores layers 30, 36, and 38; on 576 held-out questions with option scores, accuracy moves from 76.0% to 76.7%, with seven corrected and three newly incorrect answers. The supplementary response cubes support the per-decision study; the shared-set results use this separate protocol.

C Baseline Adapters

Every budgeted adapter starts from the scan of the empty set and every single edit, is billed for every distinct set it evaluates up to the cap $n + 1 + t$, and returns a set of size at most three. Queries also obey this cardinality constraint for every adapter except permutation Shapley. Fixed-cost controls report their evaluation counts in Table 3. Each paragraph states the implemented objective and its relation to the published method.

Proposition 4 (additive value-oracle floor)**.** *For $n \geq 1$ and $1 \leq k \leq n$, no selector with at most n adaptive evaluations is exact on every additive utility $Y(S) = b + \sum_{i \in S} w_i$, and measuring the empty set and every single edit, $n + 1$ evaluations, recovers every parameter.*

Proof. Follow the all-zero transcript for at most n evaluations. The incidence matrix of the evaluated sets has $n+1$ columns and fewer rows, so it admits a nonzero null vector (b, w), and $w \neq 0$ because a pure baseline cannot vanish on a nonempty row. With zero evaluations take any $w \neq 0$. Normalize $\max_i |w_i| = 1$. The utilities Y and $-Y$ share the transcript and the returned set, and their regrets sum to the range of Y over $\mathcal{F}_k$, which is at least one; one regret is at least one half. □

Additive argmax. Applies the set of at most k edits with the largest positive gains. REAP ranks experts by router-weighted activation saliency and removes the lowest (Lasby et al., 2026); the adapter applies the same rank-and-apply logic to measured edit gains.

Proposition 5 (pairwise bound on the additive error)**.** *Let $\Delta_{ij} Y(A) = Y(A \cup \{i,j\}) - Y(A \cup \{i\}) - Y(A \cup \{j\}) + Y(A)$ and suppose $|\Delta_{ij} Y(A)| \leq c_{ij}$ for every pair and every context $A \subseteq \mathcal{I} \setminus \{i,j\}$ with $|A| \leq k-2$. Then $|I(S)| \leq \sum_{\{i,j\} \subseteq S} c_{ij}$ for every $S \in \mathcal{F}_k$, and the additive argmax has regret at most $2\varepsilon_k \leq k(k-1) \max_{i \neq j} c_{ij}$.*

Proof. Along any order of S with prefixes A_t, $I(S) = \sum_t [Y(A_t) - Y(A_{t-1}) - m_{\pi_t}]$. The bracket at step t telescopes into $t-1$ second differences, one for each earlier edit, each taken at a context of size at most $|S| - 2$. Summing gives the bound, and $\sum_{\{i,j\} \subseteq S} c_{ij} \leq \binom{k}{2} \max c_{ij}$. The regret bound is the upper bound of Theorem 1. □

Proposition 6 (additive selection with noisy single-edit gains)**.** *Suppose $\max_i |\widehat{m}_i - m_i| \leq \delta$, and let $\widehat{S}_\delta$ maximize $Y(\varnothing) + \sum_{i \in S} \widehat{m}_i$ over $\mathcal{F}_k$. Then $Y(S^\star) - Y(\widehat{S}_\delta) \leq 2\varepsilon_k + 2k\delta$.*

Proof. Write $f(S) = \sum_{i \in S} m_i$ and $\widehat{f}(S) = \sum_{i \in S} \widehat{m}_i$. The gain error gives $|\widehat{f}(S) - f(S)| \leq k\delta$ on $\mathcal{F}_k$. Since $\widehat{S}_\delta$ maximizes $\widehat{f}$, $f(S^\star) - f(\widehat{S}_\delta) \leq 2k\delta$. The upper-bound argument of Theorem 1 then gives $Y(S^\star) - Y(\widehat{S}_\delta) \leq 2\varepsilon_k + 2k\delta$. □

Pairwise fit. Measures every pair, fits the second-order Möbius model exactly, and applies its argmax (Boros & Hammer, 2002). The fit costs $1 + n + \binom{n}{2}$ queries. The empty-floor and measured-best variants also measure the proposal when it is a triple, adding one query. The empty-floor variant returns the better of the proposal and the empty set. The measured-best variant returns the best set among the scan, the pairs, and the proposal.

Proposition 7 (quadratic search contains CLIQUE)**.** *When k is part of the input, exact maximization of an explicitly given quadratic set function under $|S| \leq k$ contains CLIQUE.*

Proof. For $2 \leq k \leq n$, use zero linear weights and unit pair weights on the edges of a graph. A set of size k reaches $\binom{k}{2}$ exactly when it is a k-clique (Karp, 1972). A smaller set has fewer than $\binom{k}{2}$ pairs, so the optimum over all feasible sets reaches this threshold exactly when the graph contains a k-clique. □

A pairwise surrogate therefore carries a computational barrier in addition to the third-order terms it omits.

Second-order proposal and search in the spirit of HOPE. After the scan, both variants measure $\min\{t, \binom{n}{2}\}$ pairs in decreasing additive score and add their second-order coefficients to the additive model. The proposal variant returns the model argmax. The search variant measures this argmax when it is unmeasured and a query remains after all pairs, then returns the best measured set. At $t \leq \binom{n}{2}$, its record consists of the scan and the queried pairs. HOPE optimizes a constrained quadratic objective built from gate-weighted co-contributions (Tseng et al., 2026). The adapter uses cross-layer edits as variables, measured utility interactions as coefficients, and a query budget that bills every measured pair.

Task-utility Shapley adapter in the spirit of SHAPE. Estimates the Shapley value of each edit from sampled permutations over the full subset lattice. Prefix queries can exceed the edit budget and are billed within the query cap. It forms the coalition of the k largest positive values, measures it once, and returns the better of the coalition and the empty set; the search variant returns the best measured feasible set. SHAPE scores expert coalitions with a cooperative game over routing traces (Zhang, 2026); the adapter keeps the coalition objective and uses the task utility as the game.

Harsanyi Shapley, measured. After the scan, the adapter measures feasible sets in additive order and reserves one query for a proposal. It identifies every Harsanyi dividend whose subsets are all measured, divides each dividend equally among its edits, and sums these shares into Shapley scores. The proposal contains at most k edits with positive scores; it is measured when absent from the record, and the rule returns the best measured set, including the empty set.

Sequential greedy and random validation. Greedy examines the remaining edits in ascending candidate-index order, reuses recorded measurements, and adds the edit with the largest measured positive marginal gain (Nemhauser et al., 1978). A cap reached within a round restricts that round's choice to the candidates already measured. The rule stops when no positive gain is available or the cap prevents another measurement; gain ties use the lowest candidate index. Random validation samples uniformly without replacement from unmeasured sets of two or three edits (Bergstra & Bengio, 2012). The cardinality-matched random order separately shuffles within each set size, preserving the additive order's cardinality sequence.

Single-swap local search. Starts at the additive argmax and measures add, drop, and swap neighbours of the current best set in additive order, moving to the measured best after each evaluation. It follows exchange search over a cardinality constraint (Fu et al., 2026).

Hamming-kernel GP-UCB. Fits a Gaussian process with an exponentiated Hamming kernel to the standardized record, selects the lengthscale by marginal likelihood, and measures the unmeasured set maximizing the posterior mean plus twice the posterior standard deviation (Srinivas et al., 2010). It uses the categorical kernel of CASMOPOLITAN (Wan et al., 2021) and omits its trust region, which here would contain every feasible set.

Quadratic Thompson-sampling adapter in the spirit of BOCS. Thompson-samples a Bayesian second-order regression under a normal inverse-gamma prior and measures the unmeasured argmax of the draw. BOCS uses a sparse horseshoe prior with simulated annealing (Baptista & Poloczek, 2018); the adapter keeps the second-order Thompson step and enumerates the feasible sets exactly.

Adaptive sparse Möbius. Identifies every Möbius coefficient whose subsets are all measured, scores each unmeasured set by its identified coefficients, and measures the best. The published method learns sparse polynomials from adaptive queries (Erginbas et al., 2026); the adapter is its greedy query rule on the budget-three lattice.

Pandora index. Assigns each unmeasured set a Weitzman reservation value from the additive prior plus an empirical residual law fitted on the other decisions of the population, and measures in decreasing reservation value. It follows the Pandora-box allocation of Fisch et al. (2026), with the inspection cost set so that the allowance matches the budget.

The Pandora baseline of Table 3 fits one residual law per set size from the other decisions of the same population, leaving the current decision out. These auxiliary recorded utilities lie outside the billed per-decision search budget. Ranked Validation uses no fitted residual law.

Router-gap order. Measures sets in increasing summed router-score gap of their edits, with the same scan and floor. It isolates the value of the additive order, and it runs on the five natural populations, which record router scores.

C.1 Interleaved Validation

Proposition 8 (shared-record bound)**.** *Let M_a^{g} and M_b^{r} be the records that sequential greedy and Ranked Validation build alone with a and b extra evaluations, and let M_c be the record of Interleaved Validation after c extra evaluations under the schedule greedy, greedy, ranked. Then $M^{\mathrm{g}}_{\lfloor 2c/3 \rfloor} \cup M^{\mathrm{r}}_{\lfloor c/3 \rfloor} \subseteq M_c$, and the returned utility is at least $\max\{Y(\widehat{S}^{\mathrm{g}}_{\lfloor 2c/3 \rfloor}), Y(\widehat{S}^{\mathrm{r}}_{\lfloor c/3 \rfloor}), Y(\varnothing)\}$.*

Proof. Each search reconstructs its standalone deterministic query trajectory from the shared cache and proposes the first missing set. Greedy completes every candidate comparison in a round before advancing its logical prefix; ranked search follows its fixed additive order. Cached values satisfy the corresponding queries at no charge. Thus each scheduled measurement advances its search by at least one query on that trajectory, and a step whose search has stopped passes to the other search. For each search, either it has stopped and its full standalone record is already contained in M_c, or every one of its scheduled slots advances its trajectory. The schedule supplies at least $\lfloor 2c/3 \rfloor$ greedy slots and $\lfloor c/3 \rfloor$ ranked slots among the first c measurements, which gives the inclusion in either case. The maximum over a superset dominates the maximum over each record. □

D Additional Results

Table 2: Exact-selection difference (Diff.; percentage points), each rule minus Ranked Validation over all decisions. Paired 95% bootstrap intervals use 2,000 resamples of decisions (Decisions) or 1,280 question clusters (Questions).

	$t = 4$			$t = 7$		
Rule	Diff.	Decisions	Questions	Diff.	Decisions	Questions
GP-UCB	+1.0	[−0.0, +2.0]	[+0.1, +2.0]	+1.4	[+0.3, +2.5]	[+0.3, +2.4]
Local search	+0.3	[−0.4, +1.1]	[−0.4, +1.1]	+1.0	[+0.2, +1.9]	[+0.2, +1.9]
Sparse Möbius	+0.3	[−0.3, +0.9]	[−0.3, +0.9]	+0.3	[−0.5, +1.1]	[−0.5, +1.1]
Pandora index	−0.4	[−1.1, +0.3]	[−1.1, +0.3]	−0.8	[−1.6, +0.0]	[−1.5, −0.0]
Interleaved	−2.9	[−3.9, −2.0]	[−3.9, −1.9]	−1.9	[−2.8, −0.9]	[−2.9, −1.0]
Second-order search	−3.0	[−4.1, −1.8]	[−4.1, −1.9]	−7.0	[−8.3, −5.7]	[−8.4, −5.7]
Thompson sampling	−4.5	[−5.5, −3.5]	[−5.5, −3.5]	−4.7	[−5.7, −3.6]	[−5.7, −3.6]
Shapley, best measured	−11.7	[−12.7, −10.7]	[−12.8, −10.7]	−16.6	[−17.7, −15.5]	[−17.8, −15.5]
Greedy	−6.8	[−8.0, −5.6]	[−8.0, −5.5]	−11.3	[−12.4, −9.9]	[−12.6, −10.0]
Random	−8.1	[−9.0, −7.1]	[−9.1, −7.0]	−11.3	[−12.4, −10.2]	[−12.5, −10.2]
Matched random order	−8.0	[−8.9, −7.1]	[−9.0, −7.0]	−11.1	[−12.2, −10.1]	[−12.2, −10.0]
Shapley, Harsanyi game	−3.9	[−4.4, −3.4]	[−4.4, −3.4]	−3.2	[−3.7, −2.7]	[−3.7, −2.8]
Shapley, truncated game	−9.3	[−10.3, −8.3]	[−10.3, −8.2]	−11.3	[−12.4, −10.2]	[−12.4, −10.1]
Second-order proposal	−12.9	[−14.0, −11.8]	[−14.0, −11.7]	−21.6	[−22.9, −20.3]	[−23.0, −20.2]
Shapley adapter, proposal	−28.7	[−29.9, −27.5]	[−30.0, −27.5]	−39.0	[−40.2, −37.7]	[−40.3, −37.8]

Table 3 reports every rule at each budget under the shared accounting, and Table 2 reports decision-level and question-cluster resampling of each paired difference against Ranked Validation.

Table 4 reports the restoration budgets of Section 5. Table 5 reports exact selection per population. On five of the eight swap-restoration populations, greedy or random validation exceeds Ranked Validation in exact selection at some budgets, 17 population-and-budget cells in all, and Table 3 pools all decisions including these cells. Figure 6 traces exact selection and regret for the displayed measured rules. Figure 7 evaluates the interleaved schedules across all populations; the development populations fixed the selected schedule.

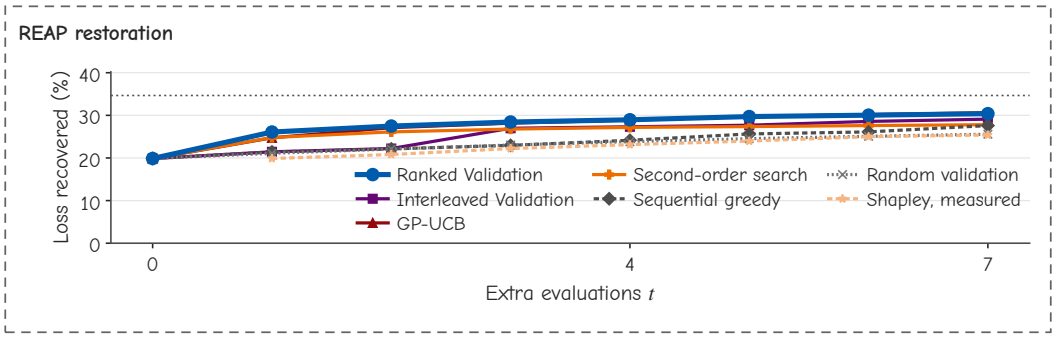

Figure 5: REAP half-pruning restoration on Qwen3.5-35B-A3B-Base: fraction of pruning log-probability loss recovered by each rule versus extra evaluations t. The dotted line marks the exhaustive optimum.

Table 3: Selection on 5,828 decisions ($k = 3$). Exact selection and harm: percentages; regret: millinats. Bold and underline mark best and second-best exact selection and regret within (b) at each t; displayed ties share rank.

		$t = 4$			$t = 7$	
Rule	Queries	Exact ↑	Regret ↓	Harm ↓	Exact ↑	Regret ↓
(a) Fixed-cost references; mean queries						
Additive argmax	7.3	10.8	55.5	17.4	10.8	55.5
Pairwise fit	25.3	20.0	50.5	29.9	20.0	50.5
Pairwise fit, measured return	26.0	56.0	9.3	0.0	56.0	9.3
(b) Shared query caps						
Second-order proposal	$n+1+t$	25.5	37.3	9.5	27.9	36.1
Shapley proposal	$n+1+t$	9.7	60.3	0.0	10.5	52.7
Ranked Validation	$n+1+t$	38.3	16.1	0.0	49.5	11.5
Hamming-kernel GP-UCB	$n+1+t$	**39.3**	**15.7**	0.0	**50.9**	**10.6**
Single-swap local search	$n+1+t$	<u>38.7</u>	16.5	0.0	<u>50.5</u>	11.5
Adaptive sparse Möbius	$n+1+t$	38.6	16.2	0.0	49.8	<u>11.2</u>
Pandora index	$n+1+t$	38.0	<u>16.0</u>	0.0	48.7	11.3
Interleaved Validation	$n+1+t$	35.4	19.0	0.0	47.6	13.3
Second-order search	$n+1+t$	35.4	19.1	0.0	42.5	15.6
Quadratic Thompson sampling	$n+1+t$	33.9	19.5	0.0	44.8	13.2
Shapley, measured return	$n+1+t$	26.7	28.6	0.0	32.9	22.5
Harsanyi Shapley, measured	$n+1+t$	34.5	18.6	0.0	46.3	12.6
Sequential greedy	$n+1+t$	31.5	24.5	0.0	38.2	18.5
Random validation	$n+1+t$	30.3	25.2	0.0	38.2	19.8

Cluster-calibrated selection. Table 6 lists every cell of the cluster study of Section 5. Each of 200 replicates draws m questions from one group of the cluster type and splits them into a calibration half and a held-out half; a rule selects on the calibration mean and is read on the held-out mean. For replicate b, let $a^{\mathrm{c}}_{b,T}$ and $a^{\mathrm{h}}_{b,T}$ be the dividends of its calibration and held-out mean utilities. The persistent ratio of order r is the mean cross-half product of dividends, normalized by the summed singleton cross-half energy,

$$C_{b,r} = \binom{n}{r}^{-1} \sum_{|T|=r} a^{\mathrm{c}}_{b,T} a^{\mathrm{h}}_{b,T}, \qquad E_2 = \frac{\sum_b C_{b,2}}{\sum_b C_{b,1}}, \qquad \rho_2 = \sqrt{\max(0, E_2)}. \tag{5}$$

A zero root-mean-square ratio can reflect a negative finite-sample cross-half energy. All reported cells and their bootstrap resamples have positive summed singleton energy. Energy intervals resample the recorded splits conditional on the fixed question pool. The question bootstrap for gain standard errors preserves each split's selected set. Negative cross-half energy estimates are truncated to zero before taking the square root. The calibration pair ratio averages absolute pair dividends rela-

Table 4: REAP half-pruning restoration: 1,024 Qwen3.5-35B-A3B-Base decisions, $n = 8$, $k = 3$. Metrics are percentages; recovery is restored log probability divided by pruning loss. Bold and underline mark best and second-best recovery and exact selection among shared-cap rows at each t; displayed ties share rank.

	$t = 4$			$t = 7$	
Rule	Recovery ↑	Exact ↑	Harm ↓	Recovery ↑	Exact ↑
Scan reference, 9 queries					
Additive argmax	21.08	6.8	17.4	21.08	6.8
Shared caps, 13 queries at $t = 4$, 16 at $t = 7$					
Second-order proposal	23.65	11.7	8.7	24.47	13.6
Sequential greedy	24.15	13.1	0.0	27.60	18.7
Random validation	23.83	11.4	0.0	25.58	14.7
Quadratic Thompson sampling	26.73	13.7	0.0	29.04	20.5
Interleaved Validation	27.32	17.0	0.0	29.11	24.4
Pandora index	**29.27**	22.1	0.0	30.46	29.9
Adaptive sparse Möbius	<u>29.03</u>	<u>22.6</u>	0.0	<u>30.51</u>	<u>31.2</u>
Single-swap local search	29.00	<u>22.6</u>	0.0	30.31	30.6
Hamming-kernel GP-UCB	28.89	**22.9**	0.0	**30.62**	**32.0**
Ranked Validation	28.96	21.3	0.0	30.39	30.4
Exhaustive optimum (93 queries)	34.66	100.0	0.0	34.66	100.0

tive to absolute singleton dividends within each replicate, then averages across replicates. The harm column counts held-out questions whose utility under the selected set falls below the pruned model.

Interaction diagnostics. The pairwise argmax selects the empty set on 49 decisions, a single edit on 663, a pair on 612, and a triple on 4,504. The third-order dividend of a selected triple has median magnitude 1.51 times the largest absolute single-edit gain and exceeds the sum of its three pair dividends in magnitude on 19.7% of them. Of the 104,743 pair dividends, 50.2% are negative, and the interaction of the optimum divided by the largest absolute single-edit gain has first and ninth deciles -0.29 and 2.14. The additive score upper-bounds every set only on the 4.3% of decisions whose interactions are all nonpositive, and a zero-cost stop at the first measured value above the next additive score fires within seven extra evaluations on 67.9% of decisions and is exact on 32.7% of them, so the interaction has to be measured. Restricting the Shapley game to feasible sets, by truncation or by Harsanyi dividends, and returning the best measured set leaves the adapter 9.3 and 3.9 points below Ranked Validation at four extra evaluations, both significantly (Table 2).

Proposition 9 (record)**.** *The utility returned by Ranked Validation is nondecreasing in t, is at least $Y(\varnothing)$, and equals $Y(S^\star)$ precisely when $R^\star \le t$.*

Proof. After t extra evaluations the record is the scan together with the first $\min(t, |L|)$ sets of L. Records are nested in t, so their maxima are nondecreasing, and every record contains $\varnothing$. The returned value equals $Y(S^\star)$ precisely when the record contains a set of optimal value: when the scan holds one, $R^\star = 0 \le t$, and otherwise the first optimal set of L enters the record precisely when $R^\star \le t$. □

Theorem 2 (crowding bound)**.** *If $|I(S)| \le \varepsilon$ for every feasible S, every set ahead of the first optimal set in L is within 2ε of the optimum, so*

$$R^\star \;\le\; \big|\{S :\; 2 \le |S| \le k,\; Y(S) \ge Y(S^\star) - 2\varepsilon\}\big|. \tag{6}$$

Proof. If the scan already contains an optimum, $R^\star = 0$ and Eq. (6) is immediate. Otherwise, let S precede the first optimal set $S^\star$ in L. Then $\widehat{Y}(S) \ge \widehat{Y}(S^\star)$, and $|\widehat{Y} - Y| = |I| \le \varepsilon$ gives $Y(S) \ge \widehat{Y}(S) - \varepsilon \ge \widehat{Y}(S^\star) - \varepsilon \ge Y(S^\star) - 2\varepsilon$. The $R^\star - 1$ sets ahead of $S^\star$ and $S^\star$ itself all lie in the set counted by Eq. (6). □

Table 5: Exact selection (%) per population at $n + 1 + t$ evaluations for Ranked Validation (RV), Interleaved Validation (IV), and sequential greedy. Blocks hold the natural, development, and held-out swap populations; bold marks the best rule.

		$t = 4$			$t = 7$		
Population	N	RV	IV	Greedy	RV	IV	Greedy
Qwen3.5-35B-A3B-Base expert pruning, five actions	768	**50.4**	48.2	43.6	63.2	**63.3**	51.6
Qwen3.5-35B-A3B-Base expert pruning, eight actions	1,024	**22.7**	18.8	17.4	**29.9**	27.4	23.5
Qwen3.5-35B-A3B-Base REAP half pruning, eight restorations	1,024	**21.3**	17.0	13.1	**30.4**	24.4	18.7
Qwen3.6-35B-A3B expert pruning, five actions	640	**54.1**	52.7	50.5	69.5	**70.3**	59.5
Qwen3.6-35B-A3B expert pruning, eight actions	640	**30.6**	24.2	20.0	**40.8**	33.8	27.8
Qwen3.6-35B-A3B boundary, first execution	59	**40.7**	32.2	32.2	50.8	**59.3**	45.8
Qwen3.6-35B-A3B boundary, re-execution	523	**51.2**	50.5	43.6	64.2	**65.0**	50.7
Instella-MoE-16B-A3B boundary, first execution	41	56.1	**63.4**	56.1	70.7	**75.6**	58.5
Instella-MoE-16B-A3B expert distance, first execution	28	35.7	**39.3**	32.1	**57.1**	**57.1**	32.1
Qwen3.5-35B-A3B-Base boundary, held-out	368	**42.9**	41.3	37.5	**56.2**	**56.2**	42.7
Qwen3-30B-A3B-Base boundary, held-out	261	60.5	**62.1**	52.5	70.9	**72.4**	59.8
Instella-MoE-16B-A3B boundary, held-out re-execution	347	**44.4**	42.1	38.0	**59.1**	57.9	42.7
Instella-MoE-16B-A3B expert distance, held-out re-execution	105	**58.1**	54.3	50.5	65.7	**68.6**	53.3

Table 6: Global REAP cluster selection at $t = 4$. Calib. pair uses mean absolute dividends; Persistent pair is ρ_2 in Eq. (5). The 95% bootstrap interval for E_2 resamples replicates. Held-out gains and SE are in millinats. Scan best selects the best empty-or-singleton set on the calibration half.

Cluster	m	Calib. pair	Persistent pair	E_2 interval (10^{-3})	RV − scan best (SE)	RV − greedy	RV − GP-UCB	Harm (%)
Benchmark	8	0.64	0.016	$[-4.56, +4.72]$	-1.2 (1.2)	-2.3	-1.2	31.4
Benchmark	16	0.62	0.043	$[-0.28, +4.25]$	-2.6 (1.0)	-0.9	-0.7	32.4
Benchmark	32	0.66	0.012	$[-0.72, +0.97]$	$+0.1$ (0.7)	$+0.4$	$+0.4$	32.0
Benchmark	64	0.57	0.007	$[-0.50, +0.53]$	-1.0 (0.7)	-0.1	$+0.2$	32.0
MMLU subject	8	0.97	0.024	$[-61.95, +61.36]$	$+0.7$ (0.8)	$+0.2$	-0.7	40.5
MMLU subject	16	0.89	0.000	$[-38.37, +29.11]$	$+0.4$ (0.8)	-0.8	-0.1	40.3
MMLU subject	32	0.89	0.000	$[-24.57, +7.01]$	$+0.4$ (0.7)	-0.6	-0.1	37.6
HellaSwag activity	8	0.19	0.000	$[-4.72, -1.22]$	-3.6 (3.4)	$+3.9$	$+0.2$	6.2
Random	8	0.51	0.037	$[-18.28, +20.05]$	$+2.3$ (1.7)	$+2.7$	$+1.8$	40.5
Random	16	0.46	0.015	$[-6.81, +5.99]$	$+0.5$ (1.5)	-1.7	$+0.4$	41.6
Random	32	0.38	0.039	$[-1.55, +4.44]$	$+1.0$ (1.3)	-0.1	$+0.1$	45.2
Random	64	0.31	0.000	$[-1.51, +1.40]$	-1.4 (1.1)	$+0.7$	-0.3	44.1

The median of this bound equals the full list $|L|$ on every population, so the measured distribution of $R^\star$ carries the evidence; the optimum's interaction is positive on 65.5% of decisions, with a median of 0.45 times the largest absolute single-edit gain.

Further rules explored. Three further rules ran on the same records. A Gaussian-process search uses the additive score as its prior mean and queries the additive argmax first when it is absent from

the scan. Its comparator uses a zero prior mean and starts GP proposals immediately after the scan. Both use the same kernel and acquisition coefficient β, the multiplier of posterior standard deviation in $\mu + \beta\sigma$. The tested additive-prior warm-start variant produced no measured gain. With $\beta = 1$ the paired difference is -1.06 points at four and -0.60 at seven extra evaluations (separated and not separated by the test), and with $\beta = 2$ it is -0.91 and -0.79 (not separated and not separated). The acquisition coefficient selected from $\{0, 0.5, 1, 2\}$ by mean development exact selection across budgets one through seven is 0. Prefix Validation measures the nested prefixes of the single-edit ranking, spends 8.6 evaluations on average, and is exact on 25.3%. The pairwise fit with the empty-set floor measures its proposal and returns the better of the proposal and the unedited model, spends 26.0 evaluations on average, and is exact on 22.4%; returning the best set among all measured pairs and its proposal raises this to 56.0% at 26.0 evaluations, a budget far above the $n + 1 + t$ of Table 3.

Proposition 10 (floor, monotonicity, and the 2ε bound)**.** *Let a deterministic measured selector make the additive argmax its first extra query and extend its record one set at a time. For $t \geq 1$ its returned utility is at least $\max\{Y(\varnothing), Y(\widehat{S}_{\mathrm{add}})\}$, is nondecreasing in t, and has regret at most 2ε on the class of Theorem 1.*

Proof. The record after t extra evaluations contains the record after $t - 1$, so its maximum is nondecreasing in t. For $t \geq 1$ it contains $\varnothing$ and the additive argmax $\widehat{S}_{\mathrm{add}}$, which gives the floor, and the returned utility is at least $Y(\widehat{S}_{\mathrm{add}})$, whose regret is at most 2ε by Theorem 1. The residual-GP version fixes its first extra query to $\widehat{S}_{\mathrm{add}}$ and returns its measured best. □

Proposition 11 (nested prefixes)**.** *Let $P_0 \subset P_1 \subset \cdots \subset P_p$ be the nested prefixes of the single-edit ranking with $p \leq k$. For unrestricted utilities, returning the best prefix requires evaluating the $\max(0, p - 1)$ prefixes of size at least two, and evaluating them suffices.*

Proof. If some P_j with $j \geq 2$ is unevaluated, its value can be set above every evaluated value or below $Y(\varnothing)$ without changing the transcript. Returning an evaluated prefix fails in the first case and returning P_j fails in the second. □

Table 7: Selector CPU time per call per decision, excluding utility evaluations. Shared-machine load average exceeded 120; only relative times are meaningful.

Rule	CPU time per call (μs)
Ranked Validation	188
Sequential greedy	7
Single-swap local search	216
Adaptive sparse Möbius	129
Pandora index	1,030
Quadratic Thompson sampling	330
Hamming-kernel GP-UCB	634

Table 7 reports the selector's own computation. Every selector costs far less than one forward pass of the model, so the utility evaluations of Table 3 dominate the cost.

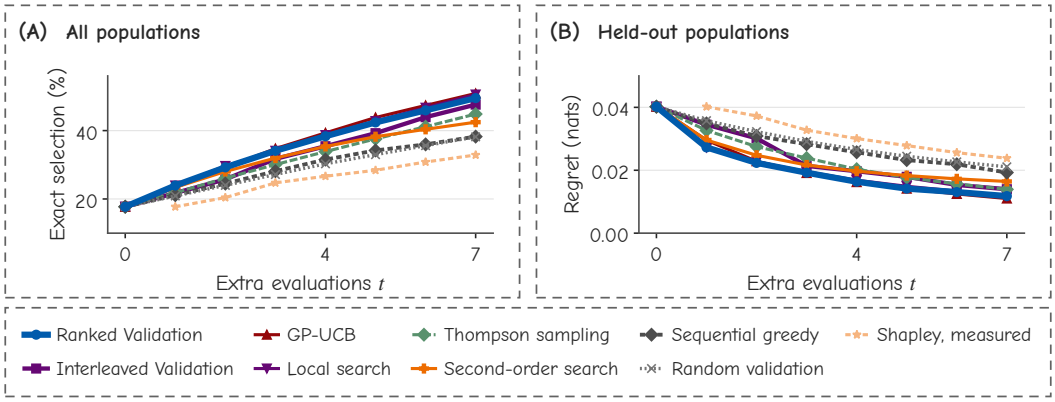

Figure 6: Displayed measured rules versus extra evaluations t: (A) pooled exact selection (%) over all populations; (B) pooled mean regret (nats) over held-out populations.

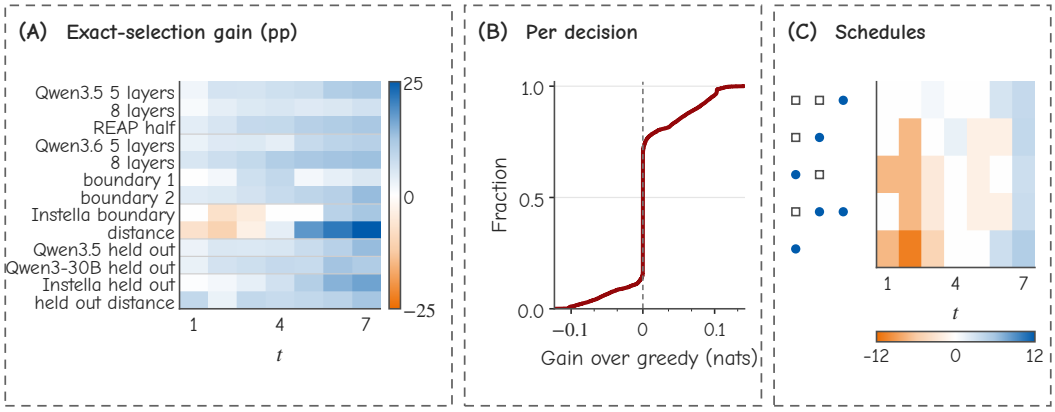

Figure 7: Relative to sequential greedy: (A) Ranked Validation's exact-selection gains by population and budget t, grouped by checkpoint; (B) its per-decision utility-gain CDF at $t = 4$ (nats); (C) minimum exact-selection gain across populations for repeated schedules. Hollow squares denote greedy proposals; filled circles, ranked proposals. Both heatmaps use percentage points.

E MECHANISM

Table 8: Interaction-sign prediction and isolated-effect recovery. N counts edit pairs for transport and isolated effects for clamping; the component-stream rows use Instella. Clamping error is in nats; n/a: not applicable.

Diagnostic	N	Reproduces (%) ↑	Clamping error ↓
Pair-interaction sign agreement			
First-order transport	14,350	49.6	n/a
Isolated later-edit effects reproduced by clamping one scope			
Complete block input	95	100.0	0.000
Router input	95	0.0	0.122
Primary stream	44	9.1	0.191
Auxiliary stream	44	0.0	0.238

Table 8 reports both mechanism diagnostics. The clamping rows hold one scope of the later block at the value it takes without upstream edits. For each measured step j with nonempty upstream context, let m_j be the isolated edit gain, d_j its contextual gain, and d_j^c its gain after clamping. Recovery requires $|d_j^c - m_j| \leq 10^{-8}$ nats at every such step, and the error column averages these

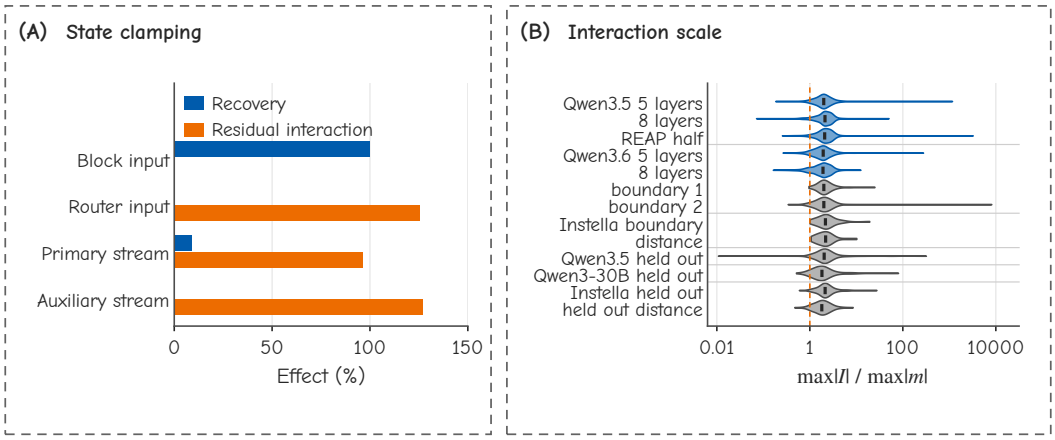

Figure 8: (A) Isolated-effect recovery and residual interaction (%) when clamping the complete block input, router input, primary stream, or auxiliary stream. (B) Per-decision maximum absolute pair/triple interaction divided by maximum absolute single-edit gain (denominator floor 10^{-12}; log axis); rows are grouped by checkpoint.

absolute differences within each example and then across examples. The complete block input reproduces every isolated later-edit effect with zero error, and the router input reproduces none. On Instella, each individual stream leaves unrecovered effects. Figure 8B uses a 10^{-12} floor for the largest absolute single-edit gain.

Proposition 12 (unique decomposition)**.** *The coefficients* $a_T = \sum_{U \subseteq T} (-1)^{|T|-|U|} Y(U)$ *are the unique family with* $Y(S) = \sum_{T \subseteq S} a_T$ *for every* S*, and Eq. (3) holds with* $a_{\{i\}} = m_i$ *and* $a_\varnothing = Y(\varnothing)$*.*

Proof. Substituting the definition of a_T gives $\sum_{T \subseteq S} a_T = \sum_{U \subseteq S} Y(U) \sum_{U \subseteq T \subseteq S} (-1)^{|T|-|U|} = \sum_{U \subseteq S} Y(U)(1-1)^{|S|-|U|}$, which equals $Y(S)$. The map from (a_T) to $(Y(S))$ is triangular with unit diagonal in cardinality order, hence invertible, which gives uniqueness. Grouping the terms of order zero, one, and at least two gives Eq. (3). □

Proposition 13 (chain identity)**.** *Order* S *as* $\pi_1, \ldots, \pi_\ell$ *with prefixes* $P_j = \{\pi_1, \ldots, \pi_j\}$ *and increments* $d_j = Y(P_j) - Y(P_{j-1})$*. Then* $d_j = m_{\pi_j} + \sum_{T \subseteq P_j,\ \pi_j \in T,\ |T| \geq 2} a_T$ *and* $\sum_j (d_j - m_{\pi_j}) = I(S)$ *for every order. The additive argmax* $\widehat{S}$ *satisfies* $Y(\widehat{S}) < Y(\varnothing)$ *exactly when* $\sum_j (m_{\pi_j} - d_j) > \sum_j m_{\pi_j}$ *along any order of* $\widehat{S}$*.*

Proof. The increment collects the coefficients of the sets that contain π_j and lie in P_j. Each $T \subseteq S$ with $|T| \geq 2$ enters exactly one increment, the one at its last element in the order, which gives the sum. The harm criterion rewrites $Y(\widehat{S}) - Y(\varnothing) = \sum_j d_j < 0$. □

Proposition 14 (complete block input as the carrier)**.** *Order the edits of a set by depth, write* $h_j(A)$ *for the block input at edit* j *under the upstream edits* A*, and write* $\tau_j(h)$ *for the effect of edit* j *measured from state* h *with the later expert plan and the readout fixed. With* Q_{j-1} *the upstream edits,* $m_j = \tau_j(h_j(\varnothing))$ *and*

$$\sum_j [\tau_j(h_j(Q_{j-1})) - \tau_j(h_j(\varnothing))] = I(S).$$

Under deterministic feed-forward execution with the same later expert plan and readout, clamping the complete block input to $h_j(\varnothing)$ *reproduces the isolated effect.*

Proof. The depth-ordered increment is $\tau_j(h_j(Q_{j-1}))$, so the identity follows from Proposition 13. Deterministic feed-forward execution maps the complete block input and the current edit to the

downstream utility through one function. Clamping that input to $h_j(\varnothing)$ therefore reproduces the isolated effect. □

The measurements above test which part of the block input suffices.

For an upstream edit i and a downstream edit j, let h be the block input of j without edits, δ_i its displacement under i, and $D(x)$ the gain of j from block input x. The second difference is $D(h+\delta_i) - D(h)$. The transport estimate $\langle \nabla D(h), \delta_i \rangle$ reuses single-edit forward passes and requires backward differentiation and gradient storage. Proposition 15 bounds its remainder when the gradient is Lipschitz along the intervening segment. Instella gradient records use the primary hidden stream. Across the five swap-restoration populations with gradient records, the estimate correlates with the measured interaction at -0.020 and agrees in sign on 49.6% of 14,350 pairs. This local estimator has limited predictive accuracy on the recorded finite edits. Finite displacements, routing discontinuities, omitted state components, and numerical estimation can contribute to its error.

Proposition 15 (first-order transport remainder)**.** *For an upstream edit i and a downstream edit j, let h be the block input of j without edits, δ_i its displacement under i, and $D(x)$ the gain of j from block input x. If D is differentiable and its gradient is L-Lipschitz along the segment $h + s\delta_i$, $0 \le s \le 1$, then*

$$|D(h+\delta_i) - D(h) - \langle \nabla D(h), \delta_i \rangle| \le \frac{L}{2}\|\delta_i\|^2.$$

Proof. The fundamental theorem of calculus gives

$$D(h+\delta_i) - D(h) - \langle \nabla D(h), \delta_i \rangle = \int_0^1 \langle \nabla D(h + s\delta_i) - \nabla D(h), \delta_i \rangle \, \mathrm{d}s.$$

By the Lipschitz condition and Cauchy–Schwarz, the absolute value is at most $\int_0^1 Ls\|\delta_i\|^2 \, \mathrm{d}s = (L/2)\|\delta_i\|^2$. □

F Component Studies and Label-Free Evaluation

Table 9: Accuracy under paired component changes with fixed expert identity: 128 ARC-Challenge and HellaSwag examples (0.78 percentage points per answer). Halving varies the gate denominator; ACE, the retained gate mass; ExFold, the route source. Relative work is completed routed expert–token evaluations divided by native-routing count, using the same checkpoint and examples in each paired experiment.

Model	Method	Setting	Accuracy (%)	Relative work
Qwen3.6	Halving	top-4 gate denominator	60.9	0.50
Qwen3.6	Halving	top-8 gate denominator	62.5	0.50
Qwen3.6	Halving	top-16 gate denominator	64.8	0.50
Instella	ACE	one-sixth target, unit mass	41.4	0.84
Instella	ACE	one-sixth target, native mass	48.4	0.84
Instella	ExFold	3 experts, replayed routes	47.7	0.50
Instella	ExFold	3 experts, current routes	44.5	0.50
Instella	ExFold	4 experts, replayed routes	46.1	0.67
Instella	ExFold	4 experts, current routes	49.2	0.67

Table 9 varies one component inside three published training-free methods at fixed expert identity: the gate denominator of activated-expert halving (Chen & Yao, 2026), the retained gate mass of ACE (Xu et al., 2026), and the route source of ExFold (Wu et al., 2026). Each component moves accuracy by a few points at unchanged expert compute, which marks gate mass and route source as edit families of their own. These families enter the same set-selection problem when combined, and they are the next candidates for exhaustive measurement.

Table 10 evaluates single-edit selection, Prefix Validation, and greedy on the same 1,280 HellaSwag validation documents for each of two checkpoints. For each document, the native model chooses the option with the highest character-normalized continuation log likelihood. The selector maximizes

Table 10: Paired HellaSwag outcomes versus root-empty (upstream swap, no restoration). Baseline accuracy/length-normalised accuracy: Qwen3.6-35B-A3B, 57.6%/75.9%; Qwen3-30B-A3B-Base, 53.4%/70.3%. Gained counts documents wrong under root-empty and correct under the selector; Lost counts the reverse. p is the exact two-sided McNemar p-value.

		Accuracy			Length-normalised		
Checkpoint	Selector	Gained	Lost	p	Gained	Lost	p
Qwen3.6-35B-A3B 1,280 documents	Best singleton (empty allowed)	5	1	0.22	6	4	0.75
	Prefix Validation	7	1	0.07	5	4	1.00
	Sequential greedy	5	0	0.06	5	4	1.00
Qwen3-30B-A3B-Base 1,280 documents	Best singleton (empty allowed)	5	2	0.45	3	3	1.00
	Prefix Validation	5	2	0.45	3	3	1.00
	Sequential greedy	5	1	0.22	3	5	0.73

that option's character-normalized log likelihood under an edit set. Gold answers are used for post-selection correctness scoring; the selector receives the native prediction and model scores. Accuracy selects the option with the largest summed continuation log likelihood. Length-normalised accuracy divides that sum by the original option's character count before selection; the count uses Unicode code points and excludes the context delimiter.

A Boundary root swap is forced in the first half of routing layers. The downstream registry contains layers whose expert supports change. The root-empty baseline keeps the root swap and forces each registered layer to its support from the root-corrupted discovery pass. A selected mask restores native supports at its marked layers and retains the corrupted supports at the others. Prefix Validation orders single edits by measured utility, evaluates nested prefixes of positive-gain edits up to the edit and query budgets, and returns the best measured set. Ranked Validation in Algorithm 1 ranks all feasible multi-edit sets. The panel supplies downstream readouts for the three stated selectors. Its gained and lost counts are both taken against root-empty. Separate native readouts score the unedited checkpoint.