DECIDING RESPONSE-UPDATE FEASIBILITY FROM INCOMPLETE COVARIANCE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Updating a model under explicit response budgets requires deciding whether a requested change can be fitted without excessive interference or update size. A compact covariance sketch makes this decision depend on retained information as well as on the request. We develop a three-way feasibility interface for fixed features at one linear projection. A realizable Frequent Directions construction gives identical sketches, shrink values and residual traces with opposite answers to the same request. Given additional covariance-vector actions, classical positive semidefinite (PSD) interpolation of the residual and its complement yields attainable cost endpoints for a radius-and-column completion set, uniformly over adaptively selected updates. An upper-feasible candidate and a full-residual-corrected lower value then give a witness, an exclusion or an unresolved interval for unregularized target error. The same budgets apply jointly to shared responses and to complete cumulative updates. A captured GPT-2 shared-response problem exhibits all three outcomes at one target tolerance; its intermediate request remains unresolved despite a feasible full-information reference. A saved-state analysis of 32 factual requests finds nearly equal first-witness costs for optimistic queries and strong Krylov after the initial sketch pass is counted, despite differences in fitting accuracy. Native factual evaluation and cumulative candidate rejections expose the limits of finite-feature protection. The results connect covariance information to actionable response decisions without turning candidate failure into a claim of infeasibility.

1 INTRODUCTION

An engineer changing a model must decide whether to commit an update, not merely whether an optimization objective has decreased. The requested response must be fitted, changes on protected inputs must stay within a specified tolerance, and the update must respect a size budget. These requirements arise together when an existing model is modified at a linear projection. Knowledge-editing and adapter-composition methods already use covariance, projection and low-rank constructions to balance competing responses (Meng et al., 2022; 2023; Fang et al., 2025; Zhang & Zhou, 2025). An explicit admission decision asks a further question: *can the requested response and both budgets be satisfied by the same update?*

The information available to make that decision can itself be limited. Protected feature rows define a covariance, but retaining a dense covariance requires quadratic storage in the feature dimension. A compact sketch reduces retained state while losing directional information. Revisiting the feature rows can recover covariance-vector actions, at the cost of additional passes. The resulting engineering choice is therefore not just how to fit an update, but how much response-relevant information to acquire before accepting it, revising it or leaving the request unsettled.

This distinction changes the interpretation of failure. A conservative sketch bound can reject a candidate that satisfies the true protection budget. An optimistic fit can violate that budget. Even a genuinely unsafe candidate does not establish that every update meeting the requested response is unsafe. We make this gap concrete with two realizable feature streams that have identical Frequent Directions (FD) sketches, accumulated shrink values and residual traces, yet give opposite feasibility answers to one response request. A binary decision based only on that shared state cannot be correct for both streams.

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.

We study the problem for fixed feature matrices and one linear projection. The retained information consists of a covariance sketch, a simultaneous residual radius and explicitly acquired covariance-vector actions. The theory connects three questions that arise in this setting. First, what can the sketch fail to distinguish? The equal-state construction gives a realizable information boundary. Second, what does an additional action establish? Positive semidefinite (PSD) interpolation of the residual and its complement gives attainable lower and upper operators for a specified completion set, uniformly over adaptively selected updates. Third, how do these operators settle a response request? An upper-feasible update supplies a witness, while a full-residual-corrected lower value can exclude the requested target error after regularization is removed. Otherwise the decision remains unresolved.

The contribution connects retained information to joint response feasibility. The FD envelope, Nyström interpolation, pseudoinverse solutions and weak duality are established tools. Together they distinguish what one candidate demonstrates from what follows about all admissible updates. The decision interface can therefore accompany an existing editor without treating that editor as a new method.

The budgets apply to the update actually used. Simultaneous response demands share one update and one aggregate budget. Repeated edits at a fixed untied projection reuse upstream features but require checking the entire cumulative update, including cross terms. Exact old-logit preservation additionally requires complete protected prefixes, including answer paths.

The GPT-2 diagnostic exhibits all three decisions at one target tolerance; its intermediate request remains unresolved despite a feasible full-information reference. Strong scalar and rank-aware block Krylov controls measure fitting and storage costs, while native factual evaluation and cumulative admission test the boundary between finite-feature response protection and preservation of model behavior.

2 Response decisions from retained covariance information

2.1 From an update budget to a covariance question

At a linear projection, the effect of changing the weight matrix can be evaluated exactly on fixed input features. If the protected rows are $a_i^\top$, changing the projection by D changes their outputs by Da_i. The total squared change is $\sum_i \|Da_i\|_2^2 = \|DA^\top\|_F^2 = \operatorname{tr}(DA^\top AD^\top)$. Thus the covariance is sufficient for this finite response cost even when the feature rows themselves are no longer retained. It need not be sufficient for downstream behavior on other inputs.

Let $A \in \mathbb{R}^{n\times d}$ contain protected feature rows, $K \in \mathbb{R}^{d\times k}$ contain new feature columns, and $R \in \mathbb{R}^{m\times k}$ contain requested changes relative to a fixed reference projection. An update $D \in \mathbb{R}^{m\times d}$ has protected response cost $c_Q(D) = \operatorname{tr}(DQD^\top)$, where $Q = A^\top A$. The request is to find an update satisfying

$$\|DK - R\|_F^2 \le \eta^2, \qquad c_Q(D) \le \epsilon, \qquad \|D\|_F^2 \le \tau^2. \tag{1}$$

Here $\eta, \epsilon, \tau \ge 0$ specify target fit, protection and update size. We use summed costs; for mean costs, replace A, B by $A/\sqrt{n}, B/\sqrt{n}$ and divide covariance bounds and the protection budget by $n > 0$. The target and protection budgets specify the requested behavior; the norm budget limits the magnitude of the modification. They are constraints on one update, rather than weights in a penalty whose achieved tolerances would have to be determined afterward. All statements in this section concern exact arithmetic and these finite features. Their numerical evaluation and native-model scope are separate.

The retained information is a sketch covariance $C = B^\top B$, a radius $\rho \ge 0$, and covariance actions on orthonormal columns of V:

$$0 \preceq E = Q - C \preceq \rho I, \qquad Y = EV = QV - CV, \qquad V^\top V = I. \tag{2}$$

Frequent Directions supplies a simultaneous covariance envelope of this kind (Ghashami et al., 2016). PSD completion and Nyström interpolation supply the operators used below (Chen et al., 2025). We use these tools to distinguish a witness for the requested response, a lower bound excluding it, and an unresolved decision. For a streamed feature matrix, the action $Qv = A^\top(Av)$ can be accumulated without forming Q. This is the additional information purchased by each query: it recovers a full covariance column in a chosen basis, not only the cost along that direction.

2.2 WHAT THE SKETCH LEAVES UNDETERMINED

The covariance order immediately gives

$$\|DB^\top\|_F^2 \le c_Q(D) \le \|DB^\top\|_F^2 + \rho\|D\|_F^2. \tag{3}$$

The upper allowance can be necessary even when the retained sketch is zero.

Proposition 1 (Realizable equal-sketch ambiguity). *Fix $\rho > 0$ and a nonzero rank-r update D, with $r \le \ell$ and $d \ge \ell + r$. Under the ℓ-row FD shrink that subtracts the smallest squared singular value from all ℓ squared singular values, two feature streams produce the same state $B = 0$, accumulated shrink ρ, and residual trace $\ell\rho$, while their costs for D are respectively $\rho\|D\|_F^2$ and zero.*

The streams can therefore agree on the available sketch information while requiring opposite candidate-admission decisions. In Figure 3, the example uses $\ell = 2$, $d = 4$ and $\rho = 1$, with covariances $Q_1 = \mathrm{diag}(1, 1, 0, 0)$ and $Q_2 = \mathrm{diag}(0, 0, 1, 1)$. They give opposite answers to one complete request: take $K = e_1$, $R = 1$, $\eta = 0$, $\epsilon = 1/2$ and $\tau = 1$. For the first stream, $DK = 1$ forces protected cost at least one. For the second, $D = e_1^\top$ satisfies all three budgets with zero protected cost. The trace alone does not remove this example. More generally, for $\rho > 0$, the classical capped-trace support function for $0 \preceq E \preceq \rho I$, $\mathrm{tr}\, E = t$ is

$$\sup_E \mathrm{tr}(DED^\top) = \rho \sum_{j=1}^{h} \sigma_j(D)^2 + (t - h\rho)\sigma_{h+1}(D)^2, \quad h = \lfloor t/\rho \rfloor, \tag{4}$$

for $0 \le t \le d\rho$, padding singular values by zeros and omitting the last term when $h = d$. When $\rho = 0$, feasibility requires $t = 0$ and the support value is zero. For the stated FD convention, $t = \ell\rho$; if $\mathrm{rank}\, D \le \ell$, the expression is $\rho\|D\|_F^2$. This is a specified worst-case information boundary, not a claim that every calibration stream attains it.

2.3 DIRECTIONAL INFORMATION AND ITS COMPLETION SET

Define the radius-and-column completion set

$$\mathcal{E}(V, Y, \rho) = \{M = M^\top : 0 \preceq M \preceq \rho I,\ MV = Y\}. \tag{5}$$

Each member represents residual information still compatible with the radius and acquired actions. The actual residual belongs to this set, but the interface does not require reconstructing it. Assume the set is nonempty. Set $T = \rho V - Y$ and

$$E_- = Y(V^\top Y)^\dagger Y^\top, \qquad E_+ = \rho I - T(\rho I - V^\top Y)^\dagger T^\top. \tag{6}$$

The pseudoinverses retain singular cases; feasibility supplies the necessary range conditions. With no columns, these endpoints are $0, \rho I$; when $\rho = 0$, feasibility forces both endpoints to be zero.

Proposition 2 (Uniform response endpoints). *The matrices E_-, E_+ belong to $\mathcal{E}(V, Y, \rho)$, and every member E satisfies $E_- \preceq E \preceq E_+$. Consequently, for every update D,*

$$c_{C+E_-}(D) \le c_{C+E}(D) \le c_{C+E_+}(D), \tag{7}$$

with both endpoint costs attained in this completion set. The bounds hold simultaneously for updates selected from the retained information. Adding consistent query directions increases E_- and decreases E_+ in PSD order.

Why two endpoints are available. Complete V to an orthonormal basis $[V, V_\perp]$. In that basis, the acquired columns fix the first diagonal block and the off-diagonal block of E, leaving only its $V_\perp$ block unknown. Positive semidefiniteness bounds that unknown block below by the generalized Schur complement. Taking this smallest block gives E_-. Applying the same argument to $\rho I - E$ gives the largest admissible block and hence E_+. Feasibility supplies the range conditions when a fixed block is singular. This construction explains both attainment and the simultaneous ordering for every candidate; it is stronger than an estimate for one prespecified quadratic form. Appendix A gives the complete proof.

This sharpness refers to (5). A known residual trace, the full FD history, or a fixed feature-row rank constraint can exclude an endpoint. The inequalities remain valid with such additional information,

but their interval need not then be sharp. In particular, the trace-aware bound in (4) uses information beyond the radius-and-column interface.

For a fixed candidate D, add a basis of the unqueried row space $\mathrm{range}((I - VV^\top)D^\top)$. At most its rank in additional directions makes the two candidate costs equal. The result follows from $(E_+ - E_-)V = 0$ and symmetry. A repaired candidate can have a different row space; this statement does not bound an entire adaptive repair sequence by the rank of one candidate. Each streamed action costs $O(nd)$ arithmetic over the feature rows, and storing V, QV already costs $2d\dim(V)$ scalars. Feature extraction, the initial sketch and solver storage are additional.

2.4 Witnesses, exclusion bounds and unresolved requests

Write $Q_- = C + E_-$ and $Q_+ = C + E_+$. For $\gamma > 0$, consider the regularized objective $F(D) = \|DK - R\|_F^2 + \gamma\|D\|_F^2$ with the same protection and norm constraints. An upper-feasible update for Q_+ is feasible for the actual Q. A dual lower value for Q_- bounds the actual regularized optimum. These are the usual feasible-set order and weak-duality arguments from convex optimization (Boyd & Vandenberghe, 2004).

In particular, choose any $\lambda, \mu \geq 0$ and any matrix Z. Let

$$\begin{aligned} H &= (\gamma + \mu)I + KK^\top + \lambda Q_-, \qquad G = ZH - RK^\top, \\ \mathcal{L}_-(Z) &= F(Z) + \lambda(c_{Q_-}(Z) - \epsilon) + \mu(\|Z\|_F^2 - \tau^2), \\ \ell(Z) &= \mathcal{L}_-(Z) - \frac{\|G\|_F^2}{\gamma + \mu}. \end{aligned} \tag{8}$$

The correction uses the full output-space residual. There is no small- residual or primal-feasibility assumption on Z.

Proposition 3 (Three-way response decision)**.** *Let D_+ satisfy $c_{Q_+}(D_+) \leq \epsilon$ and $\|D_+\|_F \leq \tau$. Then*

$$L = \max\{0, \ell(Z) - \gamma\tau^2\} \leq \min_{c_Q(D)\leq\epsilon,\, \|D\|_F\leq\tau} \|DK - R\|_F^2 \leq U = \|D_+K - R\|_F^2. \tag{9}$$

Thus $U \leq \eta^2$ supplies a witness for (1), whereas $L > \eta^2$ excludes such a witness. Otherwise the request is unresolved by these bounds. The subtraction $\gamma\tau^2$ is required when converting a regularized lower value to unregularized target error.

Proof idea: certify the full space, then remove regularization. For any displacement Δ, the lower Lagrangian expands as

$$\mathcal{L}_-(Z + \Delta) = \mathcal{L}_-(Z) + 2\langle G, \Delta\rangle + \mathrm{tr}(\Delta H \Delta^\top).$$

Since $H \succeq (\gamma+\mu)I$, completing the square bounds its infimum below by $\ell(Z)$. Weak duality makes this a lower bound for the regularized constrained optimum. Every feasible update has regularization at most $\gamma\tau^2$, so subtracting that amount gives a valid target-error lower bound. The upper endpoint uses an actual feasible update and its actual target error. This proof explains why small stationarity inside a search subspace cannot replace the full matrix G and why a regularized objective ratio cannot be used as an unregularized response decision.

An unresolved output can reflect conservative information, the candidate or solver, or the prescribed work budget. It is not by itself an information-theoretic impossibility. In contrast, Proposition 1 constructs an actual ambiguity. Candidate rejection also does not exclude a different feasible update.

What a stopping decision changes. A witness permits committing the returned update under the stated finite budgets. An exclusion establishes that fitting the request requires changing at least one budget or the underlying response problem. An unresolved interval calls for a different kind of choice: improve the candidate, tighten the lower calculation, acquire another action, or stop without a decision. These possibilities have different costs, so a useful interface reports the bounds and the information state along with its status.

3 CUMULATIVE UPDATES AT A FIXED PROJECTION

Fix token sequences, masks, positions and all model parameters except one MLP output projection W. Assume that parameter is used only at this projection, with no tied use upstream or downstream. With this upstream graph fixed in evaluation mode, those features are unchanged by repeated edits to W. For $S_t = \sum_{i \le t} D_i$, the protected response relative to W_0 is therefore $c_Q(S_t)$. The same covariance can be reused, with the declared information-maintenance cost. Per-increment bounds only imply

$$c_Q(S_t)^{1/2} \le \sum_{i \le t} c_Q(D_i)^{1/2}; \tag{10}$$

adding squared budgets would omit cross terms. A cumulative request applies its protection and norm gates to S_t, rather than just to D_t.

Let H_{t-1} contain features at every token position of the old fixed prefixes to be retained. If $D_t H_{t-1} = 0$, all projection outputs on those prefixes are unchanged, and deterministic downstream computation preserves their logits in exact arithmetic. Applying this condition at each stage preserves an old prefix relative to its insertion state. Subject-key-only protection does not imply the same statement, because other positions feed later attention. New prefixes and upstream parameter edits are outside this fixed-feature implication.

The classical complete-prefix control is explicit. Let P be the orthogonal projector onto $\ker H^\top$ and $Z = PK$. For a requested *incremental* response T, the equations $DH = 0$, $DK = T$ are feasible exactly when $T = TZ^\dagger Z$; the minimum-Frobenius-norm increment is $D_* = TZ^\dagger$. For an absolute target R after an existing update S, use $T = R - SK$. This minimum-increment construction does not automatically minimize $\|S + D\|_F$ or satisfy a cumulative protection budget: both must be evaluated on $S + D_*$.

4 CONSTRUCTING AND CHECKING RESPONSE UPDATES

The decision interface separates candidate construction from admission. At a retained state (B, ρ, V, QV), form the two endpoint operators, obtain an upper-feasible candidate and evaluate a corrected lower value. Compare their target-error interval with the requested tolerance. If the interval straddles the threshold and the action budget allows refinement, acquire one new covariance action, update the endpoints and refit. Otherwise return the decision together with the candidate and the bound values. The zero update always provides a protection- and norm-feasible fallback for nonnegative budgets, although its target error may be uninformative.

Solving the constrained response problem. For positive budgets and $\gamma > 0$, the regularized problem is a strictly convex quadratically constrained quadratic program (QCQP). With multipliers $\lambda, \mu \ge 0$, its stationary point is

$$D = RK^\top \big(KK^\top + \lambda Q_+ + (\gamma + \mu) I\big)^{-1}. \tag{11}$$

The inverse notation defines a linear solve. Our implementation uses the sketch/query/target factors and reduced systems, without assuming that the Euclidean radius is inactive. The upper problem supplies the deployed candidate; a solve with Q_- supplies guidance or a lower-value diagnostic. Each returned candidate is checked in the original physical response units. Solver stationarity and complementarity are recorded alongside the actual cost and norm, rather than replacing those feasibility checks.

Which direction to query. The conservative policy uses the leading right singular direction of the upper candidate outside the queried span. The optimistic policy uses the lower candidate to choose that direction, then deploys only an upper-feasible candidate. Double reorthogonalization removes the already queried component. Each new direction acquires the complete vector Qv; it is not just a scalar quadratic-form observation. Candidates are refit after refinement. These are concrete heuristics for the decision interface, not optimal query policies. A numerical resolved-span test is a declared stopping convention; the exact rank statement in Section 2 applies to one fixed candidate.

All output directions share the budget. If $R = U_R T_R$ with $U_R^\top U_R = I$ spanning its complete nonzero range, projection onto that range cannot increase target error, protected cost or norm. The unique regularized optimum therefore has $D = U_R X$, with the same aggregate budgets. For repeated targets $R = \delta \mathbf{1}^\top$, the scalar representation $D = \delta z^\top$ multiplies objective, squared norm, protection and squared residual by $\|\delta\|_2^2$; this factor must also appear in the dual correction. A shared response problem cannot be replaced by its first output direction or by independent per-column budgets. Appendix A.5 gives the elementary decomposition.

The classical control. Let $M = \gamma I + KK^\top$. The strong scalar reference builds a Krylov space using $M^{-1}Q$ and the objective's right-hand side, then solves the original constrained problem in the retained physical subspace. The shared reference seeds a block spanning all target directions, with explicit ordering and deflation and one total covariance-action budget. When $\mu = 0$, the stationarity systems form an affine-shift family after this preconditioning. An active norm multiplier adds μM^{-1}; the restricted QCQP remains a valid control, but that single-operator explanation is then insufficient. We retain the full-space residual lower value even when it is negative. A highly accurate restricted solve need not be close to the full-space optimum.

Information accounting. FD construction consumes a feature pass before directional refinement. Each acquired action makes another pass over the fixed rows. For each method, the declared stored-array totals include sketch, query, endpoint and prepared-solver factors; additional guide arrays are identified separately. The block control's cap-independent chain is reused across its three caps. These counts exclude shared model/features, temporary workspaces and process overhead; they are not process peak memory or isolated runtime measurements. All feature access that generated an already saved proposal also belongs to its construction cost, even if subsequent admission uses fewer new queries.

5 Controlled response studies

The studies ask whether acquired information settles a response request, what fitting and storage costs it entails, and where finite protection stops predicting native behavior. The shared-response and scalar factual studies use the same small GPT-2 snapshot (Radford et al., 2019), with different fixed tasks and projections. They are development studies; the retrospectively added controls and decision diagnostics are identified below.

5.1 One shared response: test the three-way decision

The shared problem fits 16 prescribed natural next-token response changes at the final MLP output projection. Protection rows are averaged over 512 positions; new features and target responses are divided by $\sqrt{16}$. The 768×16 target has exact stored-value rank 16. All methods use the same protection caps, norm radius and ridge weight. The added block Krylov control retains all target directions, using one cap-independent chain of 32 total covariance actions and nine QCQPs at $q = 8, 16, 32$.

Figure 1 applies the target-error decision to all 198 saved policy states. The existing retrospective tolerances $\eta/\|R\|_F \in \{0.25, 0.5, 0.75\}$ yield 594 decisions; they were not chosen before the campaign. Independent replay from physical factors checks original-feature protection, norm and the full matrix residual before subtracting $\gamma\tau^2$. It reproduces all 215 exclusions, 119 witnesses and 260 unresolved outputs. These correlated counts describe one problem, not success rates over independent requests. The numerical guard and full-information primal/dual checks appear in Appendix B.

At relative tolerance 0.5, the squared threshold is 1235.766. At $q = 32$, the optimistic policy gives intervals approximately $[1394.888, 3628.743]$, $[1175.352, 1963.695]$ and $[800.896, 802.894]$; the respective full-information target errors are 1507.545, 1200.896 and 801.120. The strict request is excluded and the loose request has a witness. The middle interval straddles the threshold despite a feasible full-information reference, directly separating an unresolved decision from infeasibility.

Fitting quality and retained information. The optimistic policy has a lower regularized objective than the block control in all nine comparisons; the conservative policy does so in eight. At $q = 32$,

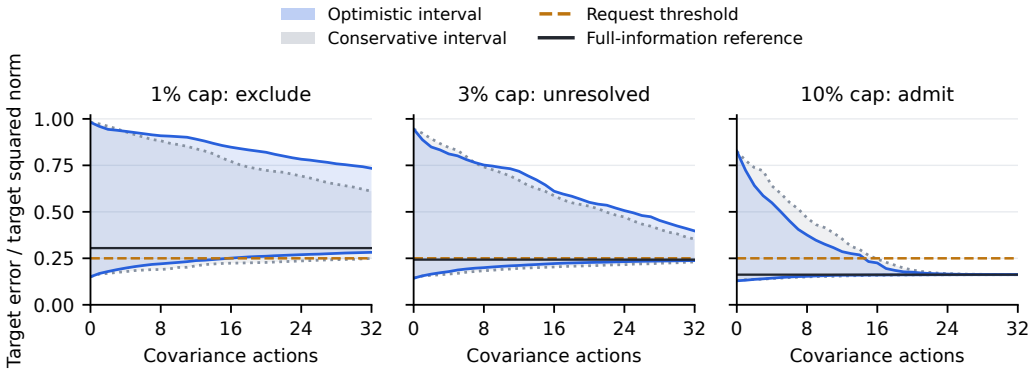


Figure 1: Three decisions for the same target tolerance on the captured shared-response problem. Bands show normalized target-error bounds $[L, U]/\|R\|_F^2$; the dashed threshold is $(\eta/\|R\|_F)^2 = 0.25$. The black full-information unregularized reference is used only for evaluation. At 32 actions, the optimistic policy excludes the strict request, leaves the intermediate request unresolved and supplies a witness for the loose request. All states are retained; these are guarded FP64 diagnostics.

the block's objective/full ratios are $2.65037, 2.09937, 1.01614$. Its declared arrays, including the preconditioner, occupy 2,896,008 bytes, versus 7,196,688 bytes for the policy upper and coexisting guide arrays. The smaller block representation trades fitting accuracy for storage on this problem. Eight of its nine full-space residual lower values are negative despite restricted stationarity at most 4.50×10^{-14}. A precise subspace solve therefore need not settle the full-space decision. Task, features and projection also differ from the scalar study, so the comparison does not isolate target rank as a cause.

5.2 Scalar factual responses and the strong control

The factual pilot retains all 32 initially hash-selected CounterFact development cases (Meng et al., 2022), without outcome filtering. Four rewrite contexts define layer-5 subject keys and one latent target per case; 512 fixed calibration features define protection. The relative RMS caps are $1\%, 3\%, 10\%$. Query policies start from FD8 and acquire at most 32 actions. Appendix B specifies the common radius, ridge, target-fitting and evaluation conventions.

All 96 full-information scalar solutions have inactive norm multipliers, motivating the subsequently added objective-preconditioned Krylov control. At the strictest cap and 16 queries, the optimistic policy puts all 32 cases within 1% of the full objective, compared with 28 for the strong control and one for the conservative policy (Figure 4 in Appendix F). At 32 queries, all three do so for every case. The strong control retains 1,761,848 tensor bytes and uses 32 feature passes; the policy retains 4,585,264 upper-checkpoint bytes and uses 33 passes, with guidance storage additional. This strong scalar result makes a rank-aware block control necessary for interpreting the shared-response comparison.

First witnesses across the 32 factual requests. The fitting comparison does not determine when a requested tolerance is met. We therefore replay all saved policy and strong-control states at the existing relative target tolerances $\eta/\|R\|_F \in \{0.25, 0.5, 0.75\}$, fixed before this scalar decision analysis. The primary comparison uses 0.5. Each decision uses an upper-feasible candidate and the full-residual lower value after subtracting $\gamma\tau^2$. Endpoint factors and retained Krylov actions suffice; no new covariance action enters the decision.

At the $1\%, 3\%, 10\%$ protection caps, median first-witness costs are $11, 6, 2$ feature passes for conservative queries and $10, 5.5, 2$ for both optimistic queries and strong Krylov (Figure 2). At the strictest cap, optimistic queries use one fewer pass in six cases, tie in 25 and use one more in one. All scalar requests have independently checked full-information witnesses, and no exclusions occur. This panel therefore measures the cost of finding a witness: the fitting advantage above translates

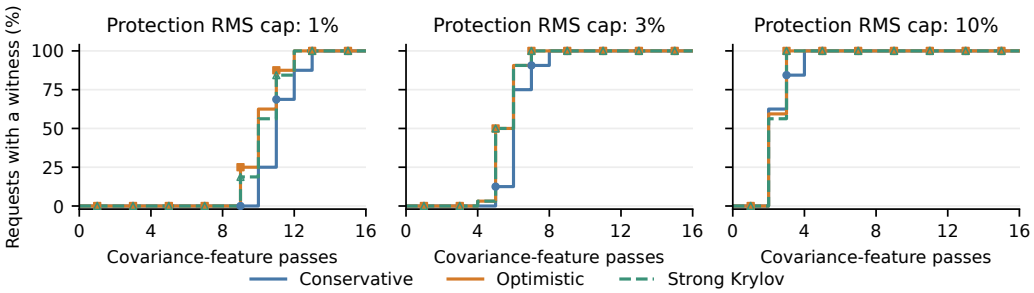


Figure 2: First-witness cost on the 32 factual requests at $\eta/\|R\|_F = 0.5$. A feature pass includes FD construction for the query policies; strong Krylov has no FD pass. Curves retain every earlier valid witness. All cases eventually have witnesses. The optimistic and strong Krylov curves are nearly coincident once this access is counted.

into only a small decision-cost difference from the strong control. Appendix G gives all tolerances, raw bounds, feasibility scaling and access accounting.

Native evaluation measures paired answer-token likelihood preferences, macro-averaged over cases, rather than generated-answer accuracy. At query budget 32, both policies, the strong control and the full-information reference reach 100% rewrite and 98.4375% paraphrase preference at every cap. Neighborhood preference falls to $49.0625\%, 48.125\%, 39.375\%$ as protection is relaxed, versus 70.3125% for the unedited model. The minimum-Euclidean equality control matches rewrite/paraphrase preference but reaches only 3.75% neighborhood preference. Finite response protection moderates this change without preserving native neighborhood behavior. Complete native comparisons and replay scopes remain in the appendices.

6 Cumulative Admission: A Distinct Application Test

The cumulative theorem changes the object being checked from an increment D_t to the entire update S_t. Two frozen groups of 16 factual cases, traversed forward and backward by two methods, give eight trajectories and 128 attempted insertions. One 3% protection cap, one fixed radius and target tolerance $0.5\|R_i\|_F$ apply throughout. The first method tests the saved single-edit proposal added to the committed state. The classical control constructs a complete-prefix nullspace increment and checks the same cumulative budgets. Queries are retained after rejection; failed candidates are not committed.

The complete study yields four nominal admissions and 124 candidate protection rejections, with no unresolved candidate decisions. Each fixed-proposal trajectory accepts its first case only; the prefix control accepts none. All four admitted cases prefer the new answer on their original rewrite prompt, and no committed state exceeds the native protection cap. Zero old-logit drift therefore provides no nontrivial test of multi-edit preservation, and rejecting these proposals does not exclude all possible updates. Appendix D preserves every trajectory, its broad evaluation and its independent 128-stage replay. It also accounts separately for the 256 covariance actions used to construct the 32 original proposals and the 18 new actions used for admission.

7 Related Work

Covariance information. Frequent Directions gives a deterministic simultaneous covariance envelope and streaming construction (Ghashami et al., 2016). The basic response envelope follows by summing its directional inequality over update rows. Classical Nyström interpolation provides PSD lower order and queried-column reproduction; RPCholesky develops an efficient randomized coordinate-column algorithm with trace-based approximation control (Chen et al., 2025). Our access to arbitrary covariance-vector products has a different cost contract. Applying interpolation to the residual complement gives the upper operator used for response admission. Sharpness is restricted

to the stated completion set, and the separate equal-FD-state construction supplies a realizable ambiguity.

Editing and accumulating history. ROME and MEMIT connect model changes to fixed internal key/response constraints (Meng et al., 2022; 2023). AlphaEdit uses protected-subspace projection and a penalty on past edited keys (Fang et al., 2025). EvoEdit maintains an evolving nullspace using incoming keys (Lyu et al., 2026); BetaEdit retains leakage/history penalties and refreshes the approximate protection space (Liu et al., 2026). Nullspace protection, leakage control, history and efficient subspace maintenance are established ingredients. We specify all three budgets and ask what incomplete information can establish about them, allowing an unresolved outcome. The cumulative contract checks the total response and distinguishes finite prefix preservation from behavior on unobserved inputs.

Shared updates. OSRM chooses low-response subspaces for LoRA merging and also reconstructs existing learned updates before merging (Zhang & Zhou, 2025). Our shared-response study concerns a joint finite target under covariance uncertainty and common protection/norm budgets. It supplies a decision contract for such response demands; its finite projection experiment does not evaluate the merging of independently trained adapters. Fixed adapter-matrix addition is commutative, whereas admission and reoptimization can make the retained updates depend on order.

8 Discussion

The theory connects a practical admission decision to the information available to justify it. A witness meets all three budgets; an exclusion bounds the best achievable target fit; an unresolved output leaves the request unsettled. The realizable FD construction explains why information loss matters, and the completion endpoints make additional actions usable for both candidate checking and request-level decisions.

Spend computation on the decision that remains open. The shared-response diagnostic separates these outcomes at one fixed target tolerance. Its intermediate request has a feasible full-information reference but remains unresolved after the prescribed actions. In this case, improving an upper-feasible candidate enough to cross the threshold would settle the request, whereas tightening a lower bound below the threshold would not. More generally, a high lower bound on one candidate's protected cost calls for revising that candidate; a corrected target-error lower bound above the tolerance excludes every update under the stated budgets. Reporting both endpoints makes this distinction visible to the engineer choosing the next operation.

Information has a construction and maintenance cost. Structure-matched Krylov is a strong scalar control; shared responses need the complete target range in a block comparison. Improved fitting comes with feature passes, retained actions and solver factors. The array counts expose this tradeoff, rather than process peak memory or deployment speed. The 32-request replay makes the distinction concrete: optimistic queries and strong Krylov have the same median first-witness cost at every tested protection cap for the primary tolerance, despite their different fitting trajectories. The relevant endpoint crossing can therefore occur well before further objective improvement would change the decision. Fixed upstream features justify covariance reuse across cumulative edits, but earlier proposal construction still has a cost.

Finite response guarantees have a defined application boundary. Calibration protection does not prevent neighborhood degradation in the factual study. Exact preservation of specified prefixes needs all of their relevant feature positions, and the cumulative study accepts too few updates to test nontrivial multi-edit preservation. Its 124 candidate rejections constrain these proposals, not the space of all possible cumulative updates. Ordinary floating-point diagnostics also remain distinct from outward-rounded certificates of the exact inequalities. These limits are consequential for interpreting a witness: it certifies the stated finite response problem in exact arithmetic, rather than model behavior on unobserved inputs.

The resulting obligation is concrete: check the shared or cumulative update actually used, and keep the request unresolved when its bounds do not settle the target.

AI USE STATEMENT

Generative AI assisted the development and refinement of the mathematical formulation, claims and proofs. It also assisted the identification and comparison of related work; experimental design; implementation and debugging; interpretation of results; preparation of figures and artifacts; and drafting and editing of the manuscript. Programmatic checks include exact small-instance witnesses, independent saved-array calculations and native-model replay where stated. Separate AI-assisted review passes checked proof assumptions, source attribution and manuscript transcription. Such checks do not constitute external peer review or establish human author review. The submitting human authors are responsible for reviewing the paper and its AI-use declaration before submission and for its final claims and artifacts.

REPRODUCIBILITY STATEMENT

The proof assumptions and complete arguments appear in Appendix A. Appendix B gives normalization, selection, solver, precision and native-evaluation conventions. The existing experiments retain versioned protocols, source and input digests, per-state outputs and separate replay reports. The reported scalar, shared-response and cumulative comparisons have independent checks with the scopes stated in those protocols. Saved-array checks and fresh native execution are distinguished. Reproducing the native outputs requires the pinned model, tokenization, batching and numerical runtime; solver-quality failures and native rounding deviations remain in the retained records.

ETHICS STATEMENT

Factual editing can change both requested and unrelated model responses. We evaluate such changes on a public benchmark and a small existing model; neighborhood degradation and failed preservation decisions are retained. Finite calibration protection is not a guarantee against misinformation, unintended behavior or bias on other inputs. No human-subject study or deployment is reported.

REFERENCES

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 978-0-521-83378-3. URL https://web.stanford.edu/~boyd/cvxbook/.

Yifan Chen, Ethan N. Epperly, Joel A. Tropp, and Robert J. Webber. Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations. *Communications on Pure and Applied Mathematics*, 78(5):995–1041, 2025. doi: 10.1002/cpa.22234. URL https://onlinelibrary.wiley.com/doi/10.1002/cpa.22234.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/29c8c615b3187ee995029284702d3f43-Abstract-Conference.html.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent Directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792, 2016. doi: 10.1137/15M1009718. URL https://epubs.siam.org/doi/10.1137/15M1009718.

Bingqing Liu, Wei Liu, and Yuhua Li. BetaEdit: Null-space constrained sequential model editing. *arXiv preprint arXiv:2605.09285*, 2026. URL https://arxiv.org/abs/2605.09285v1. Version 1, 10 May 2026.

Sicheng Lyu, Yu Gu, Xinyu Wang, Jerry Huang, Sitao Luan, Yufei Cui, Xiao-Wen Chang, and Peng Lu. EvoEdit: Evolving null-space alignment for robust and efficient knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1520–1540. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.75. URL https://aclanthology.org/2026.findings-acl.75/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Haobo Zhang and Jiayu Zhou. Unraveling LoRA interference: Orthogonal subspaces for robust model merging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26459–26472. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1284. URL https://aclanthology.org/2025.acl-long.1284/.

A Proofs of the response-information contract

A.1 The sketch envelope and realizable ambiguity

For each row d_i of D, the PSD order in (2) gives $0 \le d_i E d_i^\top \le \rho\|d_i\|_2^2$. Summing proves (3), simultaneously for all D; no independence between the retained information and D is needed.

To prove Proposition 1, let $\mathcal{R}$ be the row space of D. Choose an ℓ-dimensional subspace $\mathcal{U}$ containing $\mathcal{R}$ and another ℓ-dimensional subspace $\mathcal{W} \subseteq \mathcal{R}^\perp$. Both choices exist under $r \le \ell$ and $d \ge \ell + r$. Supply ℓ orthogonal rows of norm $\sqrt{\rho}$ spanning each subspace. Their covariances are $\rho P_{\mathcal{U}}$ and $\rho P_{\mathcal{W}}$. All ℓ squared singular values equal ρ, so the specified FD shrink leaves the zero sketch, records shrink ρ, and loses trace $\ell\rho$ in either case. Since $DP_{\mathcal{U}} = D$ and $DP_{\mathcal{W}} = 0$, the response costs are $\rho\|D\|_F^2$ and zero. Identical zero rows may be appended if a larger fixed sample count is desired; they do not distinguish the states.

For completeness, diagonalize $D^\top D = U\Lambda U^\top$, with descending eigenvalues $\lambda_j = \sigma_j(D)^2$. For any feasible E, set $e_j = (U^\top E U)_{jj}$. Then $0 \le e_j \le \rho$, $\sum_j e_j = t$, and $\operatorname{tr}(ED^\top D) = \sum_j e_j \lambda_j$. This scalar linear program allocates ρ to the h largest eigenvalues, then the remaining $t - h\rho$ to the next. A diagonal $U^\top E U$ with exactly these entries is feasible, proving (4). At $\rho = 0$, feasibility forces $E = 0$ and the response cost is zero; the displayed ratio t/ρ is not used.

A.2 Directional completion endpoints

For a PSD matrix M and any matrix V, write

$$\mathcal{N}_V(M) = MV(V^\top MV)^\dagger V^\top M = M^{1/2} P_{M^{1/2}V} M^{1/2}. \tag{12}$$

The second equality follows from the orthogonal projector formula for a matrix and its Gram pseudoinverse. It proves $0 \preceq \mathcal{N}_V(M) \preceq M$. Since that projector fixes $M^{1/2}V$, it also proves $\mathcal{N}_V(M)V = MV$, including singular Gram matrices. This is the classical PSD Nyström identity.

Take any feasible completion E. Applying (12) to E gives $\mathcal{N}_V(E) = E_- \preceq E$, using only $EV = Y$. Applying it to $\rho I - E$ gives $\mathcal{N}_V(\rho I - E) = T(\rho I - V^\top Y)^\dagger T^\top$, so $E \preceq E_+$. In particular $0 \preceq E_- \preceq \rho I$ and $0 \preceq E_+ \preceq \rho I$. Interpolation gives $E_- V = E_+ V = Y$. Thus the two endpoints are feasible completions and bound every other completion in PSD order. Summing rowwise quadratic forms proves Proposition 2 and attainability of its costs.

When a larger query subspace is observed, its completion set is contained in the previous one. The new lower endpoint belongs to the old set and therefore lies above the old lower endpoint; similarly, the new upper endpoint lies below the old upper endpoint. The bounds are matrix inequalities for each realized information state, so they remain uniform over candidates selected after observing that state. This argument does not assert sharpness after additional trace, history or row-rank constraints.

Let $M = E_+ - E_-$. Interpolation gives $MV = 0$, and symmetry gives $V^\top M = 0$. If the rows of D lie in the queried span, $D = DVV^\top$, hence $DM = 0$ and the response interval has zero width. Appending a basis of $\mathrm{range}((I - VV^\top)D^\top)$ achieves that condition using at most its dimension in new independent directions. The proof fixes D; a later repair is not constrained to retain its row space.

A.3 Residual-corrected dual values and target feasibility

Set $a = \gamma + \mu > 0$ and $T_0 = RK^\top$. The lower-operator Lagrangian is

$$\mathcal{L}_-(D) = \mathrm{tr}(DHD^\top) - 2\,\mathrm{tr}(T_0 D^\top) + \|R\|_F^2 - \lambda\epsilon - \mu\tau^2. \tag{13}$$

For an arbitrary Z, its half-gradient is $G = ZH - T_0$. Completing the square about Z shows

$$\inf_D \mathcal{L}_-(D) = \mathcal{L}_-(Z) - \mathrm{tr}(GH^{-1}G^\top) \geq \mathcal{L}_-(Z) - \|G\|_F^2/a, \tag{14}$$

because $H \succeq aI$. Every feasible point for the actual Q is feasible for Q_- and has $\mathcal{L}_-(D) \leq F(D)$. Weak duality therefore gives $\ell(Z) \leq p_\gamma(Q_-) \leq p_\gamma(Q)$, where p_γ is the regularized constrained optimum. This proof requires neither stationarity of Z nor optimal multipliers. If $R = \delta\mathbf{1}^\top$ and $Z = \delta z^\top$, then $G = \delta g^\top$, so the correction is $\|\delta\|_2^2\|g\|_2^2/a$; omitting the output weight changes the bound.

For any actual feasible D, $\|DK - R\|_F^2 = F(D) - \gamma\|D\|_F^2 \geq \ell(Z) - \gamma\tau^2$. The error is also nonnegative. Minimization gives the lower side of (9). Since $Q \preceq Q_+$, the stated D_+ is actual-feasible and supplies its upper side. The two decision implications follow immediately. The weak-duality proof permits zero budgets. A positive- budget finite-multiplier root algorithm, however, must handle zero-budget nullspace restrictions separately; that algorithmic limit is not proved by the existence of the lower bound. If $\tau = 0$, $D = 0$ is the only candidate and the target error is $\|R\|_F^2$.

Ordinary rounded endpoint construction, residual evaluation and native factor application need their own error accounting before these exact inequalities become outward-rounded certificates. Numerical comparisons of saved scalar values alone do not supply that accounting.

A.4 Cumulative response and complete-prefix preservation

Write $X_i = D_i Q^{1/2}$. Then $c_Q(S_t) = \|\sum_i X_i\|_F^2$ and the triangle inequality proves (10). Expanding the square also exposes the missing terms $2\sum_{i<j}\mathrm{tr}(D_i Q D_j^\top)$. Euclidean orthogonality of the updates does not generally set these Q-weighted terms to zero.

For a fixed prefix at one fixed MLP output projection, the graph computing its incoming features contains no use of the edited parameter by assumption. That graph is therefore unchanged. If an increment annihilates the features at every position, every output of that projection is unchanged. The residual inputs and all subsequent deterministic computations consequently coincide. Induction over edits preserves each protected prefix relative to the state at which it was inserted. Subject features alone do not suffice: with two features e_1, e_2 and increment $[0, 4]$, the subject response at e_1 stays zero while the other response changes by four. A later uniform attention average changes one logit from zero to two; a competing fixed logit of one changes the preferred answer. Uniform attention is realized by zero query/key scores. This is an exact construction, not an observed pretrained-model result.

Finally, $DH = 0$ is equivalent to $D = DP$ for the orthogonal projector P onto $\ker H^\top$. With $Z = PK$, the requested increment satisfies $DK = DZ = T$. Necessity of $T = TZ^\dagger Z$ follows from the row-space projector $Z^\dagger Z$. Under that condition, $D_* = TZ^\dagger$ obeys both constraints, because the rows of $Z^\dagger$ belong to $\mathrm{range}\,P$. Any other solution is $D_* + F$ with $F = FP$ and $FZ = 0$; hence $\langle D_*, F\rangle_F = 0$ and $\|D_* + F\|_F^2 = \|D_*\|_F^2 + \|F\|_F^2$. This proves minimum increment norm and its uniqueness, including $Z = 0$, where feasibility requires $T = 0$. For an increment-norm cap, the

minimum norm therefore supplies an exact feasibility test. For a cumulative cap, the cross term with the existing S changes the optimization problem; the adopted control evaluates $S + D_*$ explicitly and makes no claim that rejecting this candidate excludes every other cumulative update.

A.5 Complete output-range reduction

Let $P_R = U_R U_R^\top$ project orthogonally onto the complete range of R. Decompose $D = D_\parallel + D_\perp$ with $D_\parallel = P_R D$ and $D_\perp = (I - P_R)D$. Since $P_R R = R$ and $P_R(I - P_R) = 0$,

$$\|DK - R\|_F^2 = \|D_\parallel K - R\|_F^2 + \|D_\perp K\|_F^2, \tag{15}$$

$$\|D\|_F^2 = \|D_\parallel\|_F^2 + \|D_\perp\|_F^2, \tag{16}$$

$$c_Q(D) = c_Q(D_\parallel) + c_Q(D_\perp). \tag{17}$$

The last summand is nonnegative for PSD Q. Dropping $D_\perp$ therefore preserves feasibility and cannot increase target error; $\gamma > 0$ strictly decreases the objective unless $D_\perp = 0$. This proves the range statement in Section 4. Writing $D = U_R X$ then preserves all three quantities with $R = U_R T_R$ by orthonormality. This argument keeps the entire target range. A numerical rank truncation is a separate approximation. For $R = 0$, the zero update solves the nonnegative-budget request and no division by target norm is used.

B Protocols, units and measurement scope

B.1 Development design and shared model

All studies use the pinned `openai-community/gpt2` snapshot at revision `607a30d783dfa663caf39e06633721c8d4cfcd7e`. The model has 12 blocks; the MLP output projection has mathematical update shape 768×3072. Experiments use deterministic evaluation on CPU with FP32 native activations, eager attention, two Torch/BLAS threads, no past-key/value cache and unchanged model weights. An additive factor branch changes the selected projection's output before the residual addition. The mathematical solvers use FP64 features/factors as recorded. Dense weight merging has a different operation order and is not asserted bitwise equivalent to this branch.

The factual case split and model/layer choice preceded factual editing outcomes. The two query policies were frozen after the classical single-edit controls were inspected. The objective- preconditioned Krylov follow-up was motivated by the observed inactive full-solution radius multipliers, after weaker controls had been completed. The rank-16 block comparison and cumulative study were specified after the existing shared/scalar results. These follow-ups retain fixed matrices and all prescribed outcomes; they are not described as preregistered before the earlier evidence or as held-out confirmations. Repeated comparisons on the same cases are paired development evidence, not independent datasets.

B.2 Factual selection, targets and protection

CounterFact contains 21,919 records in the pinned source. We sort IDs by the SHA-256 digest of `counterfact-v1:20260911:`*case-id*, breaking ties by numeric ID. The first 512 form a development allocation, and the next 1,024 are reserved. This paper uses the first 32 development IDs only; the allocation of a reserved set does not imply that it was evaluated. The two sequential groups are the first and second consecutive groups of 16 in that same order. Subject/relation repetitions and literal subject overlaps remain present and are flagged in the retained data.

Layer 5 was fixed before outcomes. Four rewriting templates are `{}`, `The following fact is known: {}`, `It is reported that {}`, and `One observation is that {}`. One auxiliary `{subject} is a` context supplies the target-fitting Kullback–Leibler (KL) term. The subject position is the last token overlapping its inserted character span, checked against the pinned released helper. For every answer pair, tokenization of the complete string must agree with prefix tokens followed by the separately tokenized leading-space answer. Failed boundaries are recorded rather than shifted or dropped.

For each case, a single latent vector is fitted at the four subject positions using exactly 20 Adam updates, learning rate 0.5, betas $(0.9, 0.999)$ and optimizer epsilon 10^{-8}. The objective is mean target negative log-likelihood (NLL) plus $0.0625\,\mathrm{KL}(p_{\mathrm{edit}} \| p_{\mathrm{base}})$ at the auxiliary subject position and the recorded norm penalty with coefficient 0.5. The latent norm is clamped to four times the clean projection-output norm. All 21 states and the final endpoint are retained, with no best-iterate or success filtering. These settings are inspired by the pinned ROME target procedure, but the fixed contexts, small-model layer and update count differ; we do not claim a released ROME reproduction.

The fixed protected features are the 512 positions in eight 64-token calibration contexts, captured at layer 5. Divide these features by $\sqrt{512}$. If s_0 is the mean squared unedited calibration projection output, set $\epsilon = \alpha^2 s_0$ for $\alpha \in \{0.01, 0.03, 0.10\}$. The new-response objective uses the four unscaled subject keys and $R = \delta \mathbf{1}_4^\top$, so it sums their four errors. The norm radius is twice the minimum-Euclidean equality-update norm; $\gamma = 10^{-4}$ times their mean squared key norm. All methods solve the same physical problem, with scalar reduction and budgets converted together.

B.3 Query and baseline implementation

The policies begin with one FD8 construction. They acquire full covariance actions in 64-row blocks and record every $q = 0, \ldots, 32$ state, including unchanged records after an early numerical span stop. Their next direction comes from the physical upper or lower update outside the retained span, with double reorthogonalization and relative span tolerance 10^{-10}. The deployable update always uses the upper covariance. The conservative lower solve is a diagnostic; the optimistic lower solve is needed for guidance at states that acquire a subsequent action. Terminal diagnostics are not charged as required query guidance.

Matched FD references use the largest FD buffer fitting the declared upper checkpoint tensor payload at each evaluated policy setting. Their lower information access and different retained objects are explicit; matched payload does not imply matched passes, runtime or total process memory. Unconstrained equality, ridge/covariance and exact calibration-nullspace references have their own objectives. The constrained full-covariance reference is the same QCQP as the policies and is the objective denominator.

For the strong scalar control, $M = \gamma I + KK^\top$ and the stable seed is $M^{-1}K\mathbf{1} = K(\gamma I + K^\top K)^{-1}\mathbf{1}$. The four-dimensional factorization and retained seed are counted with the control. A double- orthogonalized physical basis is built from the queue of $M^{-1}Qv$ columns. All four target keys and both aggregate constraints remain in the restricted solve. The $q = 8, 32$ checkpoints match action counts, whereas $q = 9, 33$ provide the additional total-pass comparisons with FD-plus-query policies. All $32 \times 3 \times 4 = 384$ specified native configurations were evaluated.

Restricted projected covariance is symmetrized. Negative eigenvalues below $-10^{-10}\max(1, \|\cdot\|_F)$ cause rejection; accepted negative roundoff modes are clamped for its square-root factor with the count and change recorded. Actual protection is recomputed from the original saved actions. This nominal PSD convention and the original stopping tolerances do not constitute directed-rounding interval arithmetic.

B.4 Native factual metrics and complete comparison

Every original rewrite, paraphrase, neighborhood and attribute prompt is retained. For a prompt/answer pair, the released-style score averages answer-token negative log probabilities using model-dtype log-softmax and sequential FP32 accumulation. Distinct FP64 diagnostics are also stored. Rewrite/paraphrase success requires strictly smaller NLL for the new answer; neighborhood success uses the true answer. Ties are failures. Prompt outcomes are averaged within case, then macro-averaged over the 32 cases. The harmonic score uses those three aggregate rates. Attribute behavior is recorded separately. These likelihood preferences are not exact-match generation, and exponentiated mean NLL is not a whole-answer probability.

Table 1: Complete 43-group native comparison. Rates are percentages; Ret. is the number of initially true neighborhood preferences retained out of 225. Fitted FD denotes the parent policy-payload control, not a matched Krylov allocation. q is the retained checkpoint; actual query/pass ranges remain in the source tables.

Cap	Method	Rewrite	Para.	Neighbor	Ret.
–	Unedited	21.8750	17.1875	70.3125	225
–	Euclidean equality	100.0000	98.4375	3.7500	12
–	Covariance ridge	100.0000	98.4375	50.0000	160
–	Calibration nullspace	100.0000	98.4375	50.0000	160
1%	Full covariance	100.0000	98.4375	49.0625	157
1%	Conservative $q = 8$	68.7500	57.8125	64.6875	207
1%	Fitted FD (C, $q = 8$)	25.0000	17.1875	69.6875	223
1%	Conservative $q = 32$	100.0000	98.4375	49.0625	157
1%	Fitted FD (C, $q = 32$)	25.0000	17.1875	69.6875	223
1%	Optimistic $q = 8$	81.2500	65.6250	63.1250	202
1%	Fitted FD (O, $q = 8$)	25.0000	17.1875	69.6875	223
1%	Optimistic $q = 32$	100.0000	98.4375	49.0625	157
1%	Fitted FD (O, $q = 32$)	25.0000	17.1875	69.6875	223
3%	Full covariance	100.0000	98.4375	48.1250	154
3%	Conservative $q = 8$	100.0000	98.4375	50.0000	160
3%	Fitted FD (C, $q = 8$)	25.0000	17.1875	69.0625	221
3%	Conservative $q = 32$	100.0000	98.4375	48.1250	154
3%	Fitted FD (C, $q = 32$)	25.0000	20.3125	68.4375	219
3%	Optimistic $q = 8$	100.0000	98.4375	48.7500	156
3%	Fitted FD (O, $q = 8$)	25.0000	17.1875	69.0625	221
3%	Optimistic $q = 32$	100.0000	98.4375	48.1250	154
3%	Fitted FD (O, $q = 32$)	25.0000	20.3125	68.4375	219
10%	Full covariance	100.0000	98.4375	39.3750	126
10%	Conservative $q = 8$	100.0000	98.4375	39.3750	126
10%	Fitted FD (C, $q = 8$)	31.2500	21.8750	66.8750	214
10%	Conservative $q = 32$	100.0000	98.4375	39.3750	126
10%	Fitted FD (C, $q = 32$)	46.8750	26.5625	64.6875	207
10%	Optimistic $q = 8$	100.0000	98.4375	39.3750	126
10%	Fitted FD (O, $q = 8$)	31.2500	21.8750	66.8750	214
10%	Optimistic $q = 32$	100.0000	98.4375	39.3750	126
10%	Fitted FD (O, $q = 32$)	40.6250	26.5625	65.0000	208
1%	Objective Krylov $q = 8$	68.7500	53.1250	64.3750	206
1%	Objective Krylov $q = 9$	78.1250	65.6250	63.4375	203
1%	Objective Krylov $q = 32$	100.0000	98.4375	49.0625	157
1%	Objective Krylov $q = 33$	100.0000	98.4375	49.0625	157
3%	Objective Krylov $q = 8$	100.0000	98.4375	49.6875	159
3%	Objective Krylov $q = 9$	100.0000	98.4375	48.7500	156
3%	Objective Krylov $q = 32$	100.0000	98.4375	48.1250	154
3%	Objective Krylov $q = 33$	100.0000	98.4375	48.1250	154
10%	Objective Krylov $q = 8$	100.0000	98.4375	39.3750	126
10%	Objective Krylov $q = 9$	100.0000	98.4375	39.3750	126
10%	Objective Krylov $q = 32$	100.0000	98.4375	39.3750	126
10%	Objective Krylov $q = 33$	100.0000	98.4375	39.3750	126

Baseline-conditioned retention counts the 225 neighborhood comparisons that the original model preferred correctly. It differs from the overall neighborhood macro rate. Full comparison files also retain 144 descriptive paired pointwise percentile intervals designed during the native follow-up. They have no multiple-comparison correction and are not an independent training-seed or population guarantee. The manuscript's descriptive point values do not add an inferential claim from those intervals.

The actual FP32 branch factors are saved independently of the FP64 solver factors. Replays check selected-projection output, native logits/NLL and weight invariance. Original-model, rounded-factor linear and actual native response costs remain distinct. The original dense-merge parity attempt's pointwise failures are retained; later branch/replacement verification did not relax that failed dense-merge assertion.

B.5 Unregularized target-decision replay

The decision diagnostic reuses every state of the retained shared-response query campaign: two policies, three protection caps and stages 0 through 32. The three relative tolerances $0.25, 0.5, 0.75$ were introduced in its existing post-run interpretation; they were not preregistered before the underlying campaign. The present replay preserves all 594 decisions and evaluates the full physical residual from the saved left/right factors, original features, sketch and endpoint factors. It imports no candidate-construction or optimization routine. The largest residual-square correction is 5.67×10^{-23}; the separate regularizer subtraction is still required.

The numerical decision guard is $2 \times 10^{-6} \max(1, \|R\|_F^2, \tau^2, \epsilon)$, approximately 0.0098861 here. Exclusion requires $L > \eta^2 +$ guard; admission requires $U < \eta^2 -$ guard and the checked original-feature protection and norm constraints within that nominal tolerance. Other cases remain unresolved. Full-covariance references have active norm constraints; their stored primal target error matches the dual value after subtracting $\gamma\tau^2$, with a nonnegative shifted norm multiplier. These checks validate the numerical interpretation, not outward-rounded feasibility. Figure 1 shows all states at relative tolerance 0.5; the replay retains the other two tolerances, packet hashes, residual corrections and every margin. This is one shared-response problem, so stages are not independent experimental units.

B.6 Rank-16 shared response and its block control

The WikiText-2 natural-text problem (Merity et al., 2017) uses eight 64-token calibration spans, 16 disjoint 32-token target spans from the training text, and eight nonoverlapping 128-token validation contexts. Selection seed is 20260911; spans are not filtered by prediction correctness or loss. The deployment projection is the final block's MLP output. At each target's last position, a zero-initialized response vector is fitted for 150 Adam steps (learning rate 0.1, betas $(0.9, 0.999)$, optimizer epsilon 10^{-8}) using mean next-token cross entropy and a relative squared-norm penalty of 0.001. Each response is projected to 20% of the baseline final residual-stream norm. All final targets, including initially correct ones, remain in the aggregate matrix.

Protected rows are divided by $\sqrt{512}$, and new keys and responses by $\sqrt{16}$, so both costs have their declared mean interpretation. The update radius is 0.05 times the unedited projection's Frobenius norm. The ridge weight is 0.01 times mean squared protected-feature norm divided by 3072. Relative RMS protection caps remain $1\%, 3\%, 10\%$. The target has all 16 independent directions; the exact stored-value rank check keeps a nonzero dyadic minor rather than relying only on a numerical SVD threshold.

The added block control uses unconstrained-ridge-energy seeds, physical double reorthogonalization and the fixed $M^{-1}Q$ queue. Relative seed-rank and span thresholds are 10^{-10}. The query chain is common to the three caps and performs exactly 32 physical actions. At each of $q = 8, 16, 32$, it solves the complete target with one protection multiplier and one norm multiplier. The full-output gradient enters the residual lower value; a restricted dual cannot replace it. Independent replay checks every original-feature action, queue step, state and declared array total.

The new comparison has no additional native inference; its claims are objective, access and saved-array results. Earlier natural-text native studies and all their adverse outcomes remain preserved, but their values are not attributed to the new block method. No learned-adapter merger is constructed by this experiment.

B.7 Cumulative application protocol

The adopted finite protocol uses the two groups of 16 factual cases, both orders and both admission arms. Its full record distinguishes a proposed increment C from a complete candidate $T = S + C$.

It uses $\epsilon_{\text{seq}} = 0.4048113110335834$, the existing middle cap, and $\tau_{\text{seq}} = 23.334111656671244$, the largest existing per-case radius over the 32 cases. This is a fixed order-independent rule, not a scale chosen from sequence outcomes. Current target error is checked against $\eta_i^2 = 0.25\|R_i\|_F^2$ relative to the original model.

Each trajectory begins with an empty queried span and FD8 and has eight new actions in total. It cannot import the parent proposals' directions for free. Candidate norm and target fit precede the directional protection gate; near-boundary values within the declared 10^{-8} scaled margins remain unresolved. Rejected candidates do not change the committed update, while queries about the same covariance are retained. The protocol does not use parent lower certificates whose original per-case radii differ.

The complete-prefix arm protects only accepted earlier cases and all valid positions in their four existing contexts with both true and new answers. Its orthogonal projector and matrix-target pseudoinverse use the declared relative rank threshold 10^{-10} and equality diagnostic 10^{-7} with recorded scaling. No future cases, paraphrases or neighborhood features enter that projector. The increment is rounded once to FP32 factors; the cumulative admission gates operate on those actual factors interpreted in FP64. Native execution rounding is measured after decisions are fixed.

Canonical mathematical subject keys retain the original capture's batching. Complete-prefix features and native anchors use a fixed eight-row batch per case. Token, mask, position and causal-prefix identities are checked, while cross-batch feature-rounding differences are recorded. Canonical keys are not silently replaced by this later capture. The exact fixed-feature theorem does not promise bitwise identity across different FP32 batching layouts.

Every stage scores all previously attempted cases, including rejections, and retains original and insertion-state references separately. Final broad evaluation uses the same existing prompts. Repeated identical inputs in an identical committed state may reuse the exact saved result, with logical and executed counts both recorded. Original rewrite success uses the first, untemplated prompt and a strict new-answer preference; the per-case four-context diagnostic is a different quantity. Ties count neither as success nor as a retained strict insertion preference. Old-case retention excludes the insertion stage. All four accepted cases strictly prefer the new answer on all four templates at insertion, and each preference is retained through 15 later attempted stages. Since none of those later stages commits an update, these 240 template observations do not test preservation under another edit.

Table 2: Final broad likelihood preferences for all eight trajectories, macro-averaged over all 16 attempted cases, in percent. Rewrite uses the original untemplated prompt; paraphrase and attribute use new-answer preference, neighborhood uses true-answer preference. Ties are unsuccessful and remain in the denominators.

Group/order	Arm	Rewrite	Para.	Neighbor	Attr.
0/→	Fixed	25.0000	28.1250	61.8750	64.3750
0/→	Prefix	31.2500	21.8750	61.2500	65.0000
0/←	Fixed	25.0000	21.8750	60.6250	64.3750
0/←	Prefix	31.2500	21.8750	61.2500	65.0000
1/→	Fixed	12.5000	12.5000	79.3750	63.1250
1/→	Prefix	12.5000	12.5000	79.3750	63.1250
1/←	Fixed	18.7500	18.7500	80.0000	66.2500
1/←	Prefix	12.5000	12.5000	79.3750	63.1250

The native workload contains 2,920 logical batches and 14,848 logical answer continuations: 256 shared original-prefix continuations, 8,704 stage continuations and 5,888 final broad continuations. Exact state/input cache reuse yields 2,208 batch hits and 712 executed batches with 5,184 answer continuations. Calibration batches contain no answer continuations. The unchanged-model broad rates appear in both prefix-arm rows for each group because that arm commits no update. All 32 original-prefix captures have bitwise-equal canonical subject features across the two declared batch layouts, and original calibration outputs equal the retained parent capture. Native replay independently reproduces all 712 executed batches exactly.

The shared original-prefix feature cache occupies 38,068,224 array bytes; the full-calibration oracle rows occupy 12,582,912 bytes. These are separate from per-state proposal/query factors and

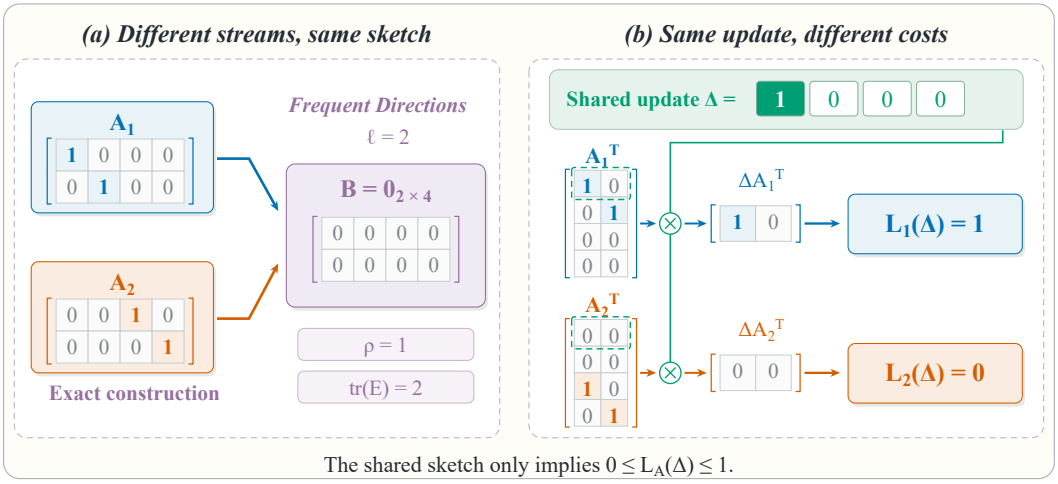


Figure 3: Identical retained information can hide different interference. Ordinary two-row FD maps both displayed feature streams to $B = 0$, accumulated shrink $\rho = 1$ and residual trace 2. The same rank-one update has protected cost one or zero. The losses are calculated from the original streams, not recovered from the shared summary. This exact example motivates acquiring response-relevant covariance information; it is not a measured model failure frequency. Proposition 1 gives the general construction.

the 497,759,232-byte model parameter payload. Lossless native packets occupy 1,826,436,379 disk bytes. These declared payloads are neither peak process memory nor a deployment-speed comparison. Source and input digests, decisions saved before post-decision evaluation, all rejected candidates and their original target fits remain available in the retained records.

C Information-boundary illustration and retained diagnostics

Figure 3 displays the exact two-stream construction behind Proposition 1; Appendix A gives the general proof.

D Complete cumulative admission study

The two frozen groups of 16 factual cases are traversed forward and in reverse by two admission methods, giving eight trajectories and 128 attempted insertions. Every attempted stage remains in the evaluation, including rejections. One cumulative 3% calibration-response cap and one order-independent existing radius apply to the total update relative to the original projection. The requested current response is also anchored to the original model, with error tolerance $0.5\|R_i\|_F$.

The fixed-proposal method adds the saved conservative 3%, $q = 8$ single-edit proposal to the committed update and checks that sum. The classical method constructs the complete-prefix nullspace increment from Section 3, using the residual target $R_i - SK_i$, then checks the same cumulative budgets. Each trajectory starts with FD8 and at most eight additional covariance actions in total. Queries are retained after rejection because they concern the same calibration covariance; failed candidates are not committed. The protocol and all outcome-affecting sources were frozen before the sequence run.

Table 3 retains all eight outcomes. Each fixed-proposal trajectory accepts only its first case; the prefix arm accepts none. The complete matrix has four nominal admissions and 124 candidate protection rejections, with no unresolved decisions. All four accepted cases prefer the new answer on their original rewrite prompt. Nevertheless, the final all-case rewrite counts are only $4, 4, 2, 3$ in the fixed arm, compared with $5, 5, 2, 2$ in the unchanged-model arm. Admission of one fitted response does not imply a net improvement over every attempted factual request. No committed state exceeds the actual native protection cap.

Table 3: All eight fixed cumulative trajectories. Admit counts committed proposals out of 16; New/16 counts final strict new-answer preferences on all 16 original rewrite prompts, including rejected cases. Qv counts only new admission actions. Native cost uses actual FP32 projection-output differences; its cap is 0.404811. Fixed proposals have the separately charged construction cost described in the text.

Group/order	Arm	Admit	New/16	Qv	Native cost
0/→	Fixed	1	4	3	0.381726
0/→	Prefix	0	5	0	0.000000
0/←	Fixed	1	4	7	0.341355
0/←	Prefix	0	5	0	0.000000
1/→	Fixed	1	2	4	0.368412
1/→	Prefix	0	2	0	0.000000
1/←	Fixed	1	3	4	0.376368
1/←	Prefix	0	2	0	0.000000

The result is a limit of these constructions at the prescribed budgets, not a proof that other cumulative updates are infeasible. In particular, the prefix arm rejects every minimum-norm proposal while its protected set is still empty. It therefore provides no empirical test of preservation with populated old prefixes. In the fixed arm no later update is accepted; the observed zero insertion-state logit drift is consequently not evidence of nontrivial multi-edit preservation. The exact implication in Section 3 has its mathematical scope, while this finite study establishes the narrower candidate-admission outcome.

The saved single-edit proposals consumed eight actions each before this study: 256 over the 32 unique proposals, or 128 for an individually constructed 16-case trajectory. This cost is separate from the budget of eight new admission actions per trajectory and is not treated as a free shared query cache. The classical arm requires features of all positions in accepted old rewrite/answer prefixes; the shared original-prefix capture is charged separately. Its intended information access is not matched to the fixed-proposal arm. The completed run uses 18 new admission actions in total. Independent replay checks all 128 stages and 712 unique native batches, with exact saved-output agreement. Appendix B.7 gives the complete broad evaluation, cache accounting and precision conventions. No cap, query budget, proposal or case was changed in response to these outcomes.

E Additional scalar numerical diagnostics

Actual FP32 projection execution has 128 positive cap differences in the query/reference campaign and 202 in the strong-control campaign; their maxima are 4.1493×10^{-7} and 6.2847×10^{-7} in the recorded squared-response units. There are no positive rounded-factor norm excesses in those campaigns. No candidate was silently rescaled to remove an excess. These are nominal implementation results, not outward-rounded feasibility certificates.

E.1 Lower values are not solver-quality declarations

A separate replay evaluates 6,336 existing scalar policy states and 76,032 relative regularized-value decisions, including multiplier transfer at the same saved candidates. It retains the complete physical residual correction. The original 2×10^{-7} relative solver-residual quality threshold fails in 29 states; all 116 associated transferred-bound decisions are false. Every other retained replay assertion passes, and no accepted decision exceeds its requested ratio against the saved full reference. These nominal diagnostics do not show a stopping advantage over the strongest lower-bound alternatives, nor can their regularized ratios replace the target-error decision in Proposition 3. Solver quality, nominal decision validity and machine-certified feasibility remain distinct.

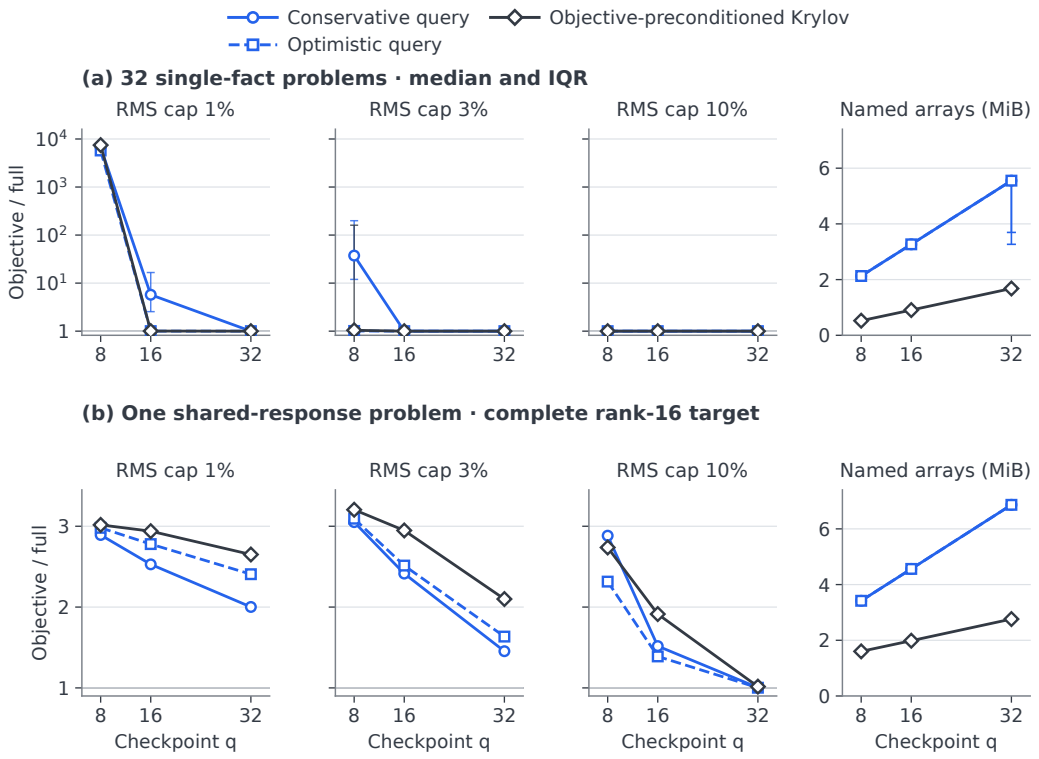


Figure 4: Fitting quality and declared retained arrays against objective-preconditioned Krylov. (a) Thirty-two scalar factual problems; objective panels show medians and interquartile ranges, not confidence intervals. (b) One shared rank-16 problem. Ratios use the matching full-information objective. Cost panels include upper and guide arrays for the policies and the preconditioner for Krylov; points/bars show medians/full ranges (MiB $= 2^{20}$ bytes). Policies also require an initial FD pass. Appendix F details early stops, shared chains and excluded storage. The comparison measures fitting and retained arrays, not peak memory or runtime.

F Additional fitting and storage accounting

Shared-response reference. The 768×16 target's exact stored-value rank is independently checked by an integer minor and rational elimination. The block control's 32 actions are a single aggregate budget, not 32 actions per target column. At the widest cap and $q = 32$, its nominal regularized enclosure is $785.513598 \leq 801.303997 \leq 814.237219$ for the full-space lower value, full-information objective and block objective, respectively. This state's norm multiplier is active; the other eight block states have inactive norm multipliers. Eight of the nine lower values are negative even though restricted stationarity is at most 4.50×10^{-14}. The target-error replay's fixed guard is approximately 0.00989 in squared-response units; the precise rule and independent reference check are given in Appendix B.

Scalar comparison and native replay. The original FD, plain Krylov and FD-preconditioned controls preceded the strong objective-preconditioned control. That follow-up was motivated by the inactive norm multipliers in all 96 full-information scalar solutions. At 32 queries, the strong control's strict-cap median/worst objective ratios are $1.00000150/1.00001881$. Separate native replays cover all 703 query/reference configurations and all 384 strong-control configurations. The main-text preference rates use their declared answer-token metrics, not generated-answer accuracy.

Reading Figure 4. The upper objective row is logarithmic and the lower row linear. At scalar $q = 32$, policy storage including guidance is up to $5{,}822{,}032$ bytes, of which $1{,}236{,}768$ bytes are guide arrays. This total differs from the $4{,}585{,}264$ upper-checkpoint count quoted in the text.

Policies begin with an eight-row FD sketch and one feature pass, followed by actual Qv actions. Krylov has no sketch pass, and its key/target preconditioner uses no Qv. At the 10% cap, early scalar stops use 19–30 actions for the conservative policy and 16–30 for the optimistic policy at the nominal $q = 32$ checkpoint; other displayed checkpoints use q actions. Krylov chains are shared across caps, whereas policy chains are separate. Shared caches and transient construction workspaces are excluded. These development comparisons establish neither process peak memory or runtime, nor native-model effects of the added block control or a causal effect of target rank.

G SCALAR TARGET DECISIONS AND FIRST-WITNESS ACCESS

Fixed questions and inputs. This analysis reuses all 32 factual requests, their three original protection caps, every conservative and optimistic state at stages 0–32, and all strong Krylov states at stages 1–33. It evaluates 9,504 stored states; it makes no new model calls. The relative target-error grid $\eta/\|R\|_F \in \{0.25, 0.5, 0.75\}$ is the grid already used by the shared-response diagnostic and was fixed before this scalar replay, with 0.5 designated as the primary comparison. Neither cases nor thresholds were selected from the new decision outcomes. These remain development requests, with repeated budget conditions on each case.

Physical bounds and their information access. For the repeated target $R = \delta \mathbf{1}_4^\top$, write $D = \delta r^\top$. Thus $\|R\|_F^2 = 4\|\delta\|_2^2$, the fit is $\|\delta\|_2^2 \|K^\top r - \mathbf{1}\|_2^2$, and the squared update norm is $\|\delta\|_2^2 \|r\|_2^2$. Each query policy reconstructs its lower and upper endpoint factors from FD8 and its stored query prefix. The saved lower candidate and nonnegative multipliers enter the full-coordinate correction in Equation (8); the candidate need not be feasible. For strong Krylov, $r = Va$ and the already acquired actions give $Qr = (QV)a$. The same correction uses these full coordinates, not the restricted dual. In both cases we subtract $\gamma\tau^2$ and take the maximum with zero to obtain the unregularized lower value. No additional covariance-vector action is required for either calculation. In this scalar panel every resulting lower value is zero; it does not demonstrate an exclusion advantage.

A declared inner-feasibility check. For this replay only, each saved deployable candidate is multiplied by

$$s = (1 - 10^{-10}) \min \left\{ 1, \sqrt{\epsilon / c_{\mathrm{known}}(D)}, \tau / \|D\|_F \right\},$$

omitting ratios with zero denominators. Here c_{known} is the policy's upper-endpoint cost or the Krylov cost computed from retained actions. The target error is recomputed after scaling. The largest change in scale is 2.50×10^{-10}, and independent original-feature checks place every scaled candidate inside its protection cap. This explicitly recorded check is not attributed to the earlier native evaluations. The decision guard is $2 \times 10^{-6} \max\{1, \|R\|_F^2, \tau^2, \epsilon\}$, as in the shared-response replay: a witness requires $U < \eta^2 - \mathrm{guard}$ and checked feasibility; exclusion requires $L > \eta^2 + \mathrm{guard}$. The calculations are FP64 diagnostics of the stated inequalities, not outward-rounded machine certificates.

First decisions and cost. All methods start with the zero update and lower value zero at zero passes. Thereafter each policy costs one FD construction pass plus its actual covariance actions; strong Krylov costs its actual actions. Unchanged records after a resolved-span stop add no access. At every pass budget we retain the smallest feasible upper value and largest lower value obtained so far. The first threshold crossing determines the first decision. The comparison extends through 33 feature passes for every method. Original-feature products performed to verify this analysis are offline diagnostics; they do not become free inputs to a decision. Their counts are recorded separately in the reproduction material. All costs describe one request at a time. Reusing construction or a query chain across distinct requests would be a different cost comparison.

At the primary tolerance and the 1% cap, optimistic and strong Krylov have median cost ten passes. The paired comparison is six one-pass wins, 25 ties and one one-pass loss for optimistic queries. At 3%, all 32 first-witness costs tie; at 10%, optimistic queries use one fewer pass in one case and tie in 31. Thus the small difference in witness cost persists after the initial sketch pass is included. Conservative queries need more passes at the tighter caps, while every method finds a witness for every case at each of the three declared tolerances.

Table 4: Feature passes to the first witness: median [minimum, maximum] over 32 fixed factual requests. The policy counts include one FD pass; strong Krylov counts actual covariance actions without an FD pass. All 32 requests in each cell receive a witness; there are no exclusions or censored first-witness costs.

$\eta/\|R\|_F$	Protection cap	Conservative	Optimistic	Strong Krylov
0.25	1%	13 [12,15]	12 [11,14]	12 [11,14]
0.25	3%	8 [7,10]	7 [6,9]	7 [6,9]
0.25	10%	4 [3,5]	3 [3,4]	3 [3,4]
0.5	1%	11 [10,13]	10 [9,12]	10 [9,12]
0.5	3%	6 [5,8]	5.5 [4,7]	5.5 [4,7]
0.5	10%	2 [2,4]	2 [2,3]	2 [2,3]
0.75	1%	8 [7,10]	7 [6,9]	7 [6,9]
0.75	3%	4 [3,5]	3 [3,5]	3 [3,5]
0.75	10%	2 [2,2]	2 [2,2]	2 [2,2]

Checks and the role of the full-information reference. The independent saved-array calculation verifies input hashes, physical costs, full-coordinate residuals, query actions and the FD8 envelope. It recovers the original strict-cap counts within 1% of the regularized full-information objective at $q = 16$: one conservative case, 32 optimistic cases and 28 strong Krylov cases. At $q = 32$, all methods recover all 32. The full-information candidate is separately checked and scaled by the same rule. It supplies a feasible diagnostic witness for all 96 case/cap combinations at every declared target tolerance. Its regularized objective is not treated as an unregularized optimum and is never supplied to a method's decision. Consequently, this scalar study measures witness cost rather than exclusion power; the shared-response problem in Figure 1 supplies the three-outcome diagnostic.

The anonymous reproduction package retains the normalized protected rows and all 32 requests with every saved candidate, multiplier and query/action chain. Per-request common orthonormal coordinates store the retained vectors in FP64; the original protected rows and sketch remain available for the global envelope check. Full residuals and certificate values are independently checked against the original feature-coordinate records. The scripts regenerate the bounds, all first decisions, the complete tolerance table and Figure 2, without model inference or optimizer reruns.