Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Alignment: The Retrieval Value of Offline Teacher Information

Xu Cao

x.cao@astrmira.com

Astrmira Tech

Wenxuan Xiao

Astrmira Tech

Research Article

Keywords: Information Retrieval, Knowledge Distillation, Query Compression, Offline Teacher Information, Dense Retrieval, Retrieval Efficiency

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11536138/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Beyond Alignment: The Retrieval Value of Offline Teacher Information

Xu Cao[1], Wenxuan Xiao[2]
Astrmira Tech.
[1] x.cao@astrmira.com, [2] w.xiao@astrmira.com

Abstract

A compressed retriever can gain substantial retrieval quality from teacher information that its alignment examples never constrain. We isolate this resource by changing only teacher-token responses outside the support of 2,048 fitting queries. Source features, cross-validation and fitted coefficients remain identical, while retrieval improves. Across two encoder settings and three tasks, the full teacher table exceeds the champion of a separate 30-configuration source-only affine, kernel and neighbor search in each encoder setting by .05165 ArguAna, .01334 NFCorpus and .08793 Android nDCG@10. The selected nonlinear completions improve source-coordinate reconstruction yet lose target relevance relative to the original affine rule. Revealing more exact teacher rows can likewise lower retrieval quality along a fixed acquisition order. These interventions separate information accuracy from retrieval value: a response matters through the query–document decisions it changes after pooling and readout. Retained priors provide a complementary source-null witness, and matched-readout comparisons show substantial remaining value in deep full-query computation above the offline table. The resulting account of query compression connects inherited information and offline teacher acquisition to the relevance earned by online computation. It motivates selecting teacher information for its effect on retrieval decisions.

1. Introduction

Query compression learns a cheaper query representation against a fixed document index. Its usual alignment objective measures agreement with a teacher on fitting queries. The deployed retriever also uses pretrained token coordinates, inherited readout directions and teacher responses cached elsewhere. We show that this information can improve retrieval while leaving the observed alignment fit unchanged. Accounting for it changes which cheap endpoint provides a meaningful baseline and which teacher responses are worth acquiring.

Our central intervention isolates information outside the fitting support. A teacher-token table stores a hidden-state response for each vocabulary ID and serves a query by averaging its rows. We preserve every row visited by the 2,048 fitting queries and replace only unvisited rows. Every completion presents the same fitting features, cross-validation problem and selected head. The deployed table shape and online computation are also fixed. Thus a retrieval difference cannot be attributed to a better alignment fit, a larger deployed representation or additional online depth. It identifies the value of teacher behavior outside the observed fitting support.

The information advantage persists against a stronger completion search. We select among 30 affine/ridge, Nyström kernel and weighted-neighbor configurations using only source token-type out-of-fold error. The full table exceeds the source-selected champion on all three tasks: +.05165 ArguAna, +.01334 NFCorpus and +.08793 Android nDCG@10. The Android experiment uses a second encoder setting, E5-small-v2, and reproduces the result with its own source-selected models. These controlled comparisons establish offline teacher information as a consequential resource in compressed retrieval, even when the observed alignment fit is unchanged.

The completion search also reveals a mismatch between reconstruction precision and retrieval value. In both encoder settings, source selection prefers a nonlinear completion with lower coordinate error; its target rankings are worse than the original affine completion. A budget study then replaces predicted rows with exact teacher responses. Each replacement removes that row's reconstruction error, but relevance can fall along a nested reveal order. A source-error acquisition rule also trails fixed random orders throughout the intermediate NFCorpus budgets. These results expose a selection mismatch: equal-weight coordinate reconstruction rewards a different object from the query–document decisions that determine relevance.

Ranking geometry explains why these criteria can disagree. A representation perturbation matters for ranking through its projection onto competing document directions and through the margins it crosses, not through coordinate error alone. The same accurate response can help one query, be irrelevant to another, and disrupt a third ranking after pooling and normalization. We develop this decision-sensitive view in Section 8. It suggests selecting teacher information by its anticipated effect on deployment decisions, while the present experiments supply the counterexamples and controlled evidence that motivate such a criterion.

The same information-based comparison clarifies the roles of prior retention, readout fitting and online depth. Source-null prior retention can preserve useful behavior without altering fitted source predictions. Compatible consumers can recover quality hidden by an unsuitable readout. Once the offline representation is strong, deep full-query computation still adds .07660 ArguAna, .04840 NFCorpus and .11228 Android nDCG under matched source-ridge procedures. Compression therefore involves three interacting resources: information inherited before alignment, teacher information acquired offline, and computation performed on the actual query. A shallow or zero-layer endpoint is an informative baseline only after its offline information and consumer have been made competitive.

Using alignment [1,4,8], contextual-to-static extraction [24,26–29] and prior regularization [31,34] as controlled tools, we establish the following results:

1. **An intervention that separates observed alignment from deployment quality.** Source-identical table replacements retain the entire fitted alignment pipeline while exposing positive teacher-information increments across three tasks and two encoder settings, including a fixed 30-configuration source-only completion search in each setting.
2. **A precision–value distinction for teacher information.** Stronger source reconstruction and larger exact-response budgets do not order retrieval quality. The mismatch motivates acquisition criteria that account for retrieval decisions beyond coordinate error.
3. **An information-aware baseline for online computation.** Prior, consumer and depth controls establish what can be retained or purchased offline and what additional full-query processing still contributes. Complete-request comparisons then test whether that quality earns its online cost.

Section 2 defines the fixed-index contract. Section 3 isolates teacher information and tests stronger completions and response budgets. Sections 4–6 connect retained priors, consumer compatibility and

online computation. Section 8 develops the ranking interpretation and resulting selection agenda. The appendices retain every endpoint, unsuccessful control and reproducibility record. Section 2 specifies the evaluated populations and the statistical scope of the comparisons.

2. Fixed-index retrieval and the observed alignment fit

The principal system pairs a six-layer Leaf query encoder, hidden width 384 and a learned 768-dimensional Dense output, with a twelve-layer Arctic document encoder. Query Linear operators use per-channel dynamic INT8, with floating-point pooling, projection and normalization; PC2 and PC6 denote two- and six-layer execution at this precision. Pooling and readout are specified separately. Documents are normalized FP32 CLS vectors. A fixed source document rotation represents the query head's 385-dimensional support, with fifteen structural zeros at physical width 400; width 256 keeps the leading coordinates. Retrieval scores are inner products followed by complete-index ordered selection.

Alignment uses hash-grouped MS MARCO training queries [5]. R2048 denotes source groups 0–2,047; the nested supports contain 128, 512 or 2,048 groups. Teacher-fidelity evaluation scores against all 5,183 SciFact documents [6]. Natural tasks use ArguAna's 8,674 documents and NFCorpus's 3,633 documents [7,16,22]. The common fitting procedure is centered ridge with an unpenalized intercept, float64 SVD fitting and FP32 deployment. Five source-only folds select among eight fixed penalties. Natural relevance labels select none of these heads.

An E512 panel contains 512 source evaluation groups, but the name does not imply one universal population. The lexical and matched-depth comparisons use groups 2,048–2,559. The common-readout study uses previously unused groups 2,560–3,071; the additional Arctic prior study uses 6,656–7,167. Positions are zero-indexed and inclusive. Appendix A specifies the remaining panels, model revisions and numerical conventions.

The lexical and follow-up interventions are retrospective on the declared task populations. Source-only selection prevents target-based model choice within each fixed panel; it does not make a previously seen benchmark blind. Appendix A records Leaf exposure, and Appendices K–L and O distinguish current E5 execution from the original study record.

Teacher top-ten overlap measures agreement with the teacher, not human relevance. We report nDCG@10 with original judgment grades and the declared self exclusions. Intervals use 4,000 paired resamples of queries or the stated connected groups; family-adjusted contrasts are labeled. They condition on fixed fits rather than integrating research-wide selection, arbitrary training seeds or hardware-session variation.

3. Teacher responses outside the fitting support

3.1 An intervention that leaves the fit identical

For each of 30,522 literal vocabulary IDs, we apply the original six-layer teacher to that ID with the fixed retrieval prompt and special tokens, then cache the mean hidden state before the Dense projection. Online queries average the corresponding rows and apply a source-fitted affine head. This uses an established static-distillation construction [24,26], purchasing 30,522 teacher contexts offline rather than restricting total supervision to the 2,048 alignment examples.

Let S contain the vocabulary IDs visited by the prompted, truncated alignment inputs, and let H be the teacher table. Here $|S| = 4,555$. We construct alternatives $H^{(a)}$ satisfying

$$H^{(a)}[S] = H[S].$$

Every source query therefore has the same pooled feature under every arm. The original folds, cross-validation values and fitted head remain unchanged. Only rows outside S differ: the full table uses actual teacher responses; source-affine completion fits a raw-to-teacher affine map on S; eight-neighbor completion averages teacher rows from eight source-observed neighbors in raw-embedding cosine. An all-vocabulary affine map supplies an additional-information control because it uses teacher rows outside S. It is not a source-only competitor.

Teacher table outside source support	E512 overlap	ArguAna nDCG (973 queries)	NFCorpus nDCG (323 queries)
Original teacher rows	.433398	.508233	.304053
Source-only affine completion	.376758	.458332	.292018
Source-only eight-neighbor completion	.325391	.335272	.236144
All-vocabulary affine completion	.382422	.463038	.298410

Table 1. Four predeclared lexical completions with the same original source-ridge head. Only source-unseen rows differ. In the original lexical execution, all 1,296 natural baseline complete fingerprints (973 ArguAna and 323 NFCorpus queries) and 512 source baseline top-ten arrays reproduced exactly. This historical parity statement is distinct from the new pooling lane in Section 3.2. Intervals below are pointwise historical-population comparisons.

The full table exceeds source-affine completion by .04990 [.03782,.06276] ArguAna nDCG and .01203 [.00174,.02171] NFCorpus nDCG. Its advantages over eight-neighbor completion are .17296 [.14842,.19961] and .06791 [.05099,.08592]. Full minus all-vocabulary affine is .04520 [.03383,.05709] on ArguAna and .00564 [-.00438,.01461] on NFCorpus. The native full-query teacher reaches .59537/.35651 on these tasks, above the table's .50823/.30405.

All 77 NFCorpus queries without a source-unseen token return exactly the same output under every completion. Unseen occurrences constitute 26.36% of ArguAna tokens and 8.59% of NFCorpus tokens, distinct from the 85.08% unseen vocabulary fraction. The intervention thus combines exact fitting invariance with target changes where replaced rows are used. Held-out source overlap also changes, from .376758 under source-affine completion to .433398 under the full table: new source observations can distinguish these systems even though the existing fit cannot.

The comparison fixes deployment shape and the consumer while varying offline teacher supervision. Both source-only controls inherit pretrained raw embeddings, but neither reconstructs the useful predictions supplied by the additional teacher responses. The next two sections test a broader source-only search and partial teacher-response budgets to determine how that increment depends on the completion and acquisition rule. Appendix E.3 gives the original fitting, invariance and closed-vocabulary checks.

3.2 Source-selected nonlinear completion

We test the information increment against a broader source-only reconstruction family with 30 fixed configurations: six centered affine/ridge models, fifteen 512-anchor Nyström RBF kernel-ridge

models, and nine distinct weighted-neighbor models. Five token-type folds withhold 911 of the 4,555 source-observed types at a time. Candidate training receives only the other 3,644 teacher rows. Learned means, decompositions, anchors and bandwidths use training types only. The selection objective is equal-type out-of-fold mean squared error in the original 384 teacher coordinates. Neither outside-support teacher rows, target relevance, target token frequencies nor held-out source retrieval outcomes select a model.

One winner per family and one cross-family champion are selected before new target predictions. Nyström kernel ridge with bandwidth multiplier 1 and ridge coefficient 10^{-6} wins, with source OOF MSE 4.52966×10^{-5}, compared with 4.74396×10^{-5} for the unregularized affine candidate. After source selection, each map is refit on all S, compiled into a table and overwritten with exact $H[S]$. The original R2048 alignment head is not refit. Table 1a retains all three family winners and both historical source-only controls.

Completion endpoint	ArguAna nDCG@10	NFCorpus nDCG@10
Full teacher table	.508233	.304053
Historical source affine	.458332	.292018
Historical uniform eight-neighbor	.335272	.236144
Source-selected affine/ridge	.456761	.292098
Source-selected Nyström champion	.456580	.290709
Source-selected weighted neighbors	.342451	.238215

Table 1a. The fixed source-selected completion panel on the same 973 ArguAna and 323 NFCorpus queries. The champion is chosen by source token-type OOF error, not by target quality or a separate winner per task. All endpoints have the same deployed table shape and unchanged alignment head.

The full-minus-champion nDCG differences are .05165 [.03771,.06638] on ArguAna and .01334 [.00356,.02300] on NFCorpus. These two 97.5% intervals form the prespecified nominal 95% Bonferroni family; they condition on the fixed source selection and query populations. The source-validation improvement does not transfer into higher target quality than historical affine completion. The finite search strengthens the measured information comparison while keeping model choice independent of target relevance.

All six arms preserve the source fit exactly, and the original baseline metric arrays reproduce. Appendix N records the pooling lane, inactive-query identities and historical replay checks.

3.3 Exact responses and their decision value

We next hold the selected completion and alignment head fixed while revealing b additional exact teacher rows outside S. The deployed table uses H on $S \cup A_b$ and the frozen completion elsewhere. Budgets are $b \in \{0, 256, 1024, 4096, 16384, 25967\}$. One acquisition rule orders unseen types by the mean source OOF error of their eight nearest raw-embedding source neighbors. This rule was specified before the preceding panel's target outcomes. Three uniform random nested orders use fixed seeds 20261027, 20261028 and 20261029. No target occurrence frequency or outside-S teacher error enters acquisition; the champion is never refit after a reveal.

The source-error rule does not consistently beat the fixed three-seed random mean. On NFCorpus it is below that mean at every intermediate budget; at $b = 16,384$ the difference is -.00550, with a descriptive 95% interval [-.00990,-.00104]. Both policies are fixed independently of target outcomes, and the curves retain each seed alongside their mean (Appendix N).

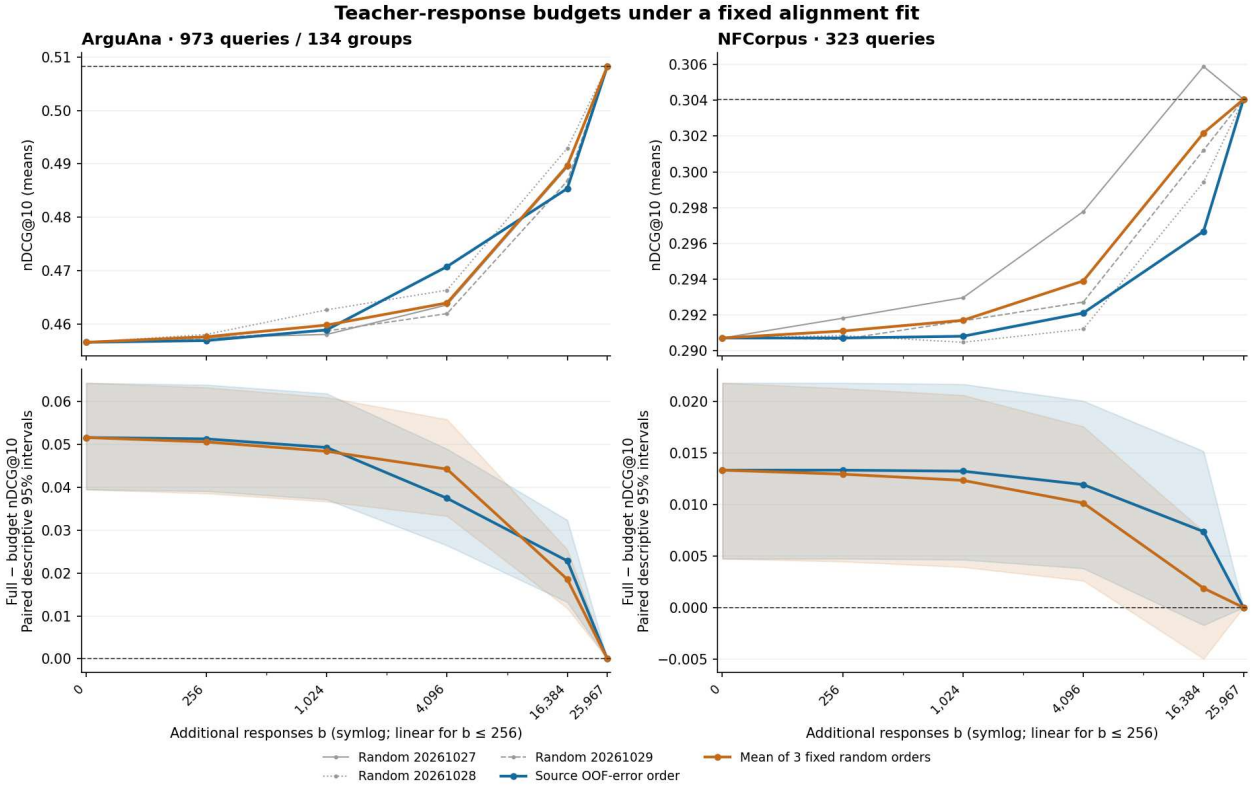


Figure 1: Fixed-head teacher-response budgets on the two Leaf tasks. All three random orders, their fixed mean, the source-error proxy and the full-table reference are retained. The horizontal axis uses base-2 symlog spacing with a linear region through 256 additional rows. Intervals are descriptive pointwise 95% paired intervals, not a simultaneous policy-selection test.

Retrieval is also nonmonotone along some nested reveal orders. NFCorpus seed 20261027 reaches .305889 at 16,384 additional rows, then falls to the full-table .304053 at 25,967. Each replaced row has zero reconstruction error afterward, but the fixed downstream ranking can become worse. The nonmonotonicity concerns the fixed acquisition-and-use pipeline. An oracle could ignore newly revealed rows, but the tested consumer uses them; the favorable partial-budget seed was not selected as a deployment policy.

The budget counts singleton teacher responses: $4,555 + b$ in total, in addition to the unchanged 2,048 full-query alignment targets. Because all teacher rows were already cached, this is an information-access replay with zero new teacher inference, not zero historical acquisition cost. It measures neither serving latency nor a paid-computation frontier. Source row MSE and response count are useful descriptors, but neither alone measures the relevance value of the acquired information.

3.4 The information increment extends to E5

We repeat the original lexical specification with E5-small-v2 on the complete BEIR CQADupStack Android collection: 699 queries and 22,998 documents. The new run uses the pinned twelve-layer, 384-dimensional model in FP32 with uniform masked-mean pooling and the original R2048 source query list and folds. Its tokenized fitting support contains 4,553 IDs. All four table completions preserve those rows, source features and the selected head; source-only cross-validation selects penalty 0.1. The native full-query teacher remains a descriptive reference rather than a matched static completion.

E5 endpoint in the new run	Android nDCG@10	Recall@10	MRR@10
Full teacher-token table	.359120	.461560	.358614
Source-only affine completion	.278784	.373893	.278218
Source-only eight-neighbor completion	.245782	.343221	.241306
All-vocabulary affine completion	.295485	.393865	.297523
Native full-query teacher reference	.469873	.580445	.475121

Table 2. E5/Android lexical endpoints on all 699 queries. The full table improves on source-only completions while the native contextual teacher remains stronger. The all-vocabulary affine endpoint has access to teacher rows outside source support and is therefore a different information condition.

The full-minus-source-affine gain is .08034, with a paired 95% interval [.06496,.09628]. All 56 queries without source-unseen tokens are unchanged across the lexical completions. This exact inactive control localizes the intervention to newly encountered lexical support rather than a change in the shared alignment pipeline.

We then transport the finite 30-configuration source-only search to E5. The 4,553 source types are split into five folds, with every type equally weighted in aggregate OOF MSE. Nyström ridge with bandwidth multiplier .5 and coefficient 10^{-6} wins with MSE .01498525, versus .01615017 for the unregularized affine candidate. The selected family maps are fixed before compiling their unseen rows; exact source rows are restored in all tables and the C1 head is unchanged.

E5 completion endpoint	Android nDCG@10	Recall@10	MRR@10
Full teacher table	.359120	.461560	.358614
Historical source affine	.278784	.373893	.278218
Historical uniform eight-neighbor	.245782	.343221	.241306
Source-selected affine/ridge	.278143	.373893	.277377
Source-selected Nyström champion	.271187	.362820	.270343
Source-selected weighted neighbors	.250991	.345415	.248078

Table 2a. E5 stronger-completion panel. The primary comparator is the source-selected champion, not the endpoint with the best Android score. All six endpoints preserve every R2048 feature and prediction under the unchanged head; all 56 inactive queries remain identical.

The full-minus-champion increment is .08793 [.07264,.10495], using the fixed C1 paired resamples. The champion's lower source-coordinate error again fails to produce higher target relevance than historical affine completion. Both encoder settings therefore show a positive teacher-response increment and the same source-to-target ordering mismatch: the best source reconstruction is not the best retrieval completion. E5's native teacher remains stronger than the full table, leaving a clear role for query-level computation (Section 5). Encoder and task vary together; the two settings provide replication of the intervention rather than a factorial attribution to model versus dataset. Appendix O gives the complete selection and audit record. The response-budget experiment remains Leaf-only.

4. A complementary witness: retaining a prior outside the source span

The lexical intervention changes unvisited representation rows. A retained readout can supply useful extrapolation through a different channel. At 128 source queries, the centered feature matrix has rank 127. Adding the existing readout's component orthogonal to that source span leaves all source predictions unchanged to float64 numerical precision, yet raises held-out teacher top-ten overlap from .1340 to .2322. This is a same-prediction witness for teacher fidelity, not a second lexical intervention or a natural-relevance estimate.

The broader prior-centered fit compares zero-centered ridge with

$$\min_{W,c} \sum_i \|Wz_i + c - y_i\|_2^2 + \lambda \|W - W_0\|_F^2,$$

where W_0 is the projected native Dense head with its original calibration scale. Features, targets, affine function class and penalty are fixed. For centered source features X_c,

$$\widehat{W}_{\text{prior}}^\top - \widehat{W}_{\text{zero}}^\top = \lambda (X_c^\top X_c + \lambda I)^{-1} W_0^\top .$$

The Hessian, target-response derivative and effective degrees of freedom agree. Unlike the isolated source-null intervention, the full prior-centered fit generally changes source predictions. At the common penalty 1.0, overlap rises from .133984 to .2673, a .13330 [.12383,.14238] gain; source selection chooses penalty 10 and gives .271680. The unchanged prior already reaches .1945, and an

intercept-only fit reaches .2420. The remaining source-selected gain is .02969 [.02490,.03447]. The warm-versus-zero improvement thus combines retained prior quality with learning from 128 new queries.

Natural-task incumbents make the new-learning increment explicit. Unchanged, intercept-only and source-selected priors give .2664/.2717/.2813 ArguAna973 nDCG and .2492/.2797/.2905 NFCorpus323 nDCG. Fitting beyond the intercept adds .00957 [.00097,.01845] and .01083 [.00528,.01695]. At 2,048 queries, both centers select .1 and have nearly identical source overlap, yet prior centering gives .3471 versus .4119 ArguAna nDCG: -.06482 [-.08103,-.04893]. Both curves improve with support; retaining the prior has a relative cost on this population. Counts are incremental to shared pretraining, an earlier 140-query learned pool and fixed projection selection.

The additional Arctic query-backbone study and the E5/Android supporting analysis test other priors. At 128 queries, Arctic sixth-layer identity centering raises overlap from .0752 to .1250 on 512 new source groups and NFCorpus nDCG from .0411 to .1043; second-layer centering changes little. On E5/Android, identity centering adds .06783 [.05022,.08656] nDCG at the same penalty under the three-comparison family correction. Appendix L retains the original interval and the new replication record. Its source-selected prior exceeds both unchanged and intercept-only identities. Appendix B retains every support, incumbent, source-null, objective and capacity control; Appendix L retains all E5 endpoints. Together, the prior and lexical interventions show how useful behavior can survive outside the directions or inputs constrained by an alignment sample.

5. What full-query processing adds above the offline table

5.1 A compatible consumer is part of the comparison

A fixed representation can support useful predictions that its original head fails to produce. Replacing only the consumer of each fixed adapted Leaf prefix with a common source-only ridge procedure recovers .11510 [.09885,.13185] and .12626 [.10880,.14451] natural nDCG for dedicated and shared training. On previously unused source groups, their width-400 overlap exceeds frozen-prefix ridge by .02760/.03021. The new E5 three-seed replication raises mean Android nDCG from .11920 to .18067 by replacing the trained KL head with source-only ridge: +.06147 [.04864,.07539] under the three-comparison family correction. Its descriptive .01003 improvement over frozen-prefix ridge is positive in every seed (Appendix M). Consumer replacement changes objective, solver, regularization and fitting features together, rather than isolating a loss-function effect.

The token table supplies an especially direct compatibility control. Source-only mean removal, without changing the native weight matrix or adding query distinctions, raises native deep-table nDCG from .00514 to .37628 on ArguAna and from .02582 to .29825 on NFCorpus. A common query-vector offset becomes a fixed document-score bias. Full ridge adds .13195 [.11251,.15247] on ArguAna, whereas its .00580 [-.00378,.01568] NFCorpus increment is unresolved. Much of the native-table failure is recoverable by adapting the consumer to information already present in the representation.

Appendix C examines the adapted-feature result through complete-ranking coordinate swaps and crossed query/index controls. It retains the negative repair tests: uniform coordinate supervision does not remove the common-readout residual, and source-determined pooling changes reduce every tested natural-relevance endpoint despite improving some transport diagnostics. The later ArguAna973 residual is also smaller without changing models. Compatible consumers must therefore be evaluated through retrieval outcomes, not geometric diagnostics or the training loss alone.

5.2 Depth, input construction and readout

We cross extraction depth, singleton-table versus full-query input, and native Dense versus source-ridge readout. Every cell uses uniform mean and the same fixed index. Source-ridge cells share R2048, targets, folds, penalty grid and solver, but fit separate weights; the six-layer full-query cell selects .01 rather than .1.

Depth	Input	Readout	Source E512 overlap	ArguAna nDCG (973 queries)	NFCorpus nDCG (323 queries)
2	Token table	Native Dense	.0102	.0003	.0118
2	Token table	Source ridge	.3207	.3841	.2701
2	Full query	Native Dense	.0574	.0394	.0533
2	Full query	Source ridge	.3582	.2751	.2890
6	Token table	Native Dense	.0318	.0051	.0258
6	Token table	Source ridge	.4334	.5082	.3041
6	Full query	Native Dense	1.0000	.5954	.3565
6	Full query	Source ridge	.8566	.5848	.3525

Table 3. Depth, input construction and readout under the fixed index, using 973 ArguAna and 323 NFCorpus queries. Every cell uses uniform mean. Native Dense is the unchanged original head trained for full-query layer six; its other cells deliberately test head compatibility. Source-ridge cells share fitting queries, target coordinates, folds and penalty grid, but fit separate weights. New cells have no separately measured service latency.

Increasing offline extraction depth from two to six improves table nDCG by .12414 [.10439,.14390] on ArguAna and .03398 [.02181,.04661] on NFCorpus. Full-query six-layer processing then adds .07660 [.05149,.10096] and .04840 [.03353,.06407] over that deep table under source ridge. E5 adds .11228 [.08787,.13643] Android nDCG from twelve-layer full-query ridge under the three-comparison family correction; Appendix L preserves both the original and newly replicated estimates. Native deep E5 is already strong: .46987 versus .47140 for full-query ridge. The positive increment compares full-query processing with cached singleton responses under source ridge; it is not a general benefit from adding another readout.

The shallow contrast changes sign. Full-query two-layer mean minus its table is -.10901 [-.13676,-.08132] on ArguAna and +.01894 [.00727,.03061] on NFCorpus. The original PC2 attention pool is stronger than the uniform full-query mean, reaching .42730/.31377 nDCG; replacing it with the mean of the same states loses .15221/.02476. Source teacher overlap nevertheless favors the shallow full-query mean over its table. Source fidelity and shallow joint processing therefore do not by themselves predict the direction of natural retrieval gains.

The singleton/full-query contrast changes joint computation, positions and prompt multiplicity together. It identifies an input-construction increment at matched depth and fitting procedure, not an isolated causal effect of attention or semantic context. The new factorial consumers have no separately measured service latency. On NevIR, the original full-query native and ridge endpoints each solve 38/225 pairs, versus 27/225 for table ridge; paired success is weak, and all original models fail the sole exact token-multiset reversal. Appendix E.2 separates these weak fine-grained capability results from the positive complete-index relevance gains.

A later fixed, mass-preserving position-weighted table consumer worsens nDCG on both historical populations: position minus uniform is -.06415 on ArguAna and -.03410 on NFCorpus, with paired

97.5% intervals [-.07768,-.05222] and [-.04877,-.02001]. Both prespecified gain tests fail. This negative result concerns one slope-0.5 recipe, not every position-aware compressor, and does not isolate a semantic-context contribution. Appendix E.4 gives its fixed design, independent saved-result audit and numerical-parity boundary.

6. The stronger offline starting point changes the cost comparison

The teacher table occupies 46.88 MB. Constructing its 30,506 rows beyond sixteen qualification rows takes 98.43 post-import CPU seconds; source fitting and evaluation take 2.33 seconds. It retains the raw-token table's 384-dimensional online feature width, while held-out source overlap rises from .30625 to .43340. These are additional offline resources, separate from the cost of serving a request.

Matched complete requests on ArguAna973 give teacher-table, raw-table, PC2-ridge and native-PC6 nDCG of .50823/.38818/.42730/.59537 at 1.958/1.998/15.565/39.816 ms. On NFCorpus323 the corresponding values are .30405/.26756/.31377/.35651 at .557/.559/2.119/5.226 ms. The table exceeds PC2 by .08093 [.05719,.10336] on ArguAna; its NFCorpus contrast is -.00972 [-.02413,.00453]. Stronger offline responses improve the cheap starting point without making every middle endpoint redundant.

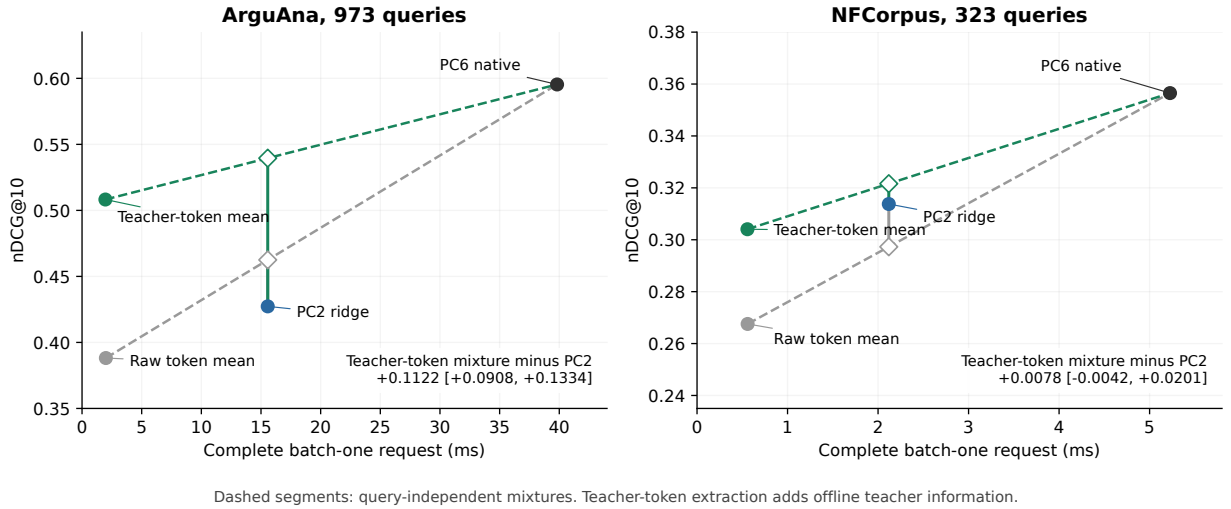


Figure 2: Offline teacher information moves the zero-layer budget curve. Dashed segments mix each lookup endpoint with the original six-layer endpoint independently of the query. Diamonds use the measured mean PC2 budget; the inset reports teacher-token mixture minus PC2 and its paired interval. All endpoints in a panel share one complete-request timing session.

A paid request includes tokenization, lookup or encoding, readout, full-index scoring, selection, sorting and any observation or continuation. Model loading, index construction and offline extraction are separate. Costs are warm singleton measurements on a shared CPU host. A query-independent mixture of same-session static endpoints supplies the same-average-cost comparator; target-selected mixtures describe the benchmark frontier, not a deployable target-free selector.

Quality complementarity does not ensure a new frontier point. A joint table/PC2 head improves nDCG over the table by .01471/.02126, while an affine prediction of its PC2 input fails to retain the same quality. Actual complete requests, however, give same-cost gaps of -.02204 [-.03340,-.01035]

on ArguAna and +.00008 [-.00703,.00676] on NFCorpus against five strong statics. Appendix E.2 retains the fixed-consumer compilation, all endpoints and their costs.

Useful allocation information can likewise coexist with an unresolved paid advantage. On 2,890 NFCorpus queries, two regressors sealed before new endpoint outputs use the same ten cheap normalized scores and log token count. At exactly 963 calls, signed-gain and negative cheap-quality predictions improve over random allocation by .00773 [.00528,.01041] and .01062 [.00778,.01304]. Their difference, -.00289 [-.00541,.00001], establishes no advantage for direct gain supervision. Subsequent complete-request execution retains both regressors and the source-CDF rule; all three family-adjusted paid-envelope intervals span zero. Table S16 in Appendix F.11 reports every online rule, not just the best point estimate.

The static action bank also matters. The unchanged cached-first source-CDF policy initially exceeds six neural statics on historical ArguAna by .01217 [.00144,.02256] nDCG at 17.730 ms, but loses on NFCorpus by -.00877 [-.01756,-.00158]. Adding unchanged BM25/RRF actions in a common ArguAna execution changes its gap from +.01178 against six statics to -.00760 against nine. The simultaneous intervals are [-.00031,.02387] and [-.02442,.00923]. Neither paid superiority against the expanded bank nor inferiority is established. Appendix F preserves the full comparisons and cost accounting; Appendix G gives the lexical and compilation implementations.

Offline teacher information raises the reference point for buying more computation. Deep processing can add relevance, and cheap observations can help allocate it; the complete-request measurements show how observation costs and stronger static alternatives absorb part of that gain.

7. Related work

Query-side alignment [1,4,8] establishes the fixed-document route to efficient retrieval. Smaller asymmetric query encoders [23] and contextual-to-static extraction [24,26–29] also precede this study. LightRetriever caches instruction-conditioned token means and studies lightweight queries with a trained document model [24]; contextual-to-static reduction and Model2Vec provide related extraction recipes [26–28]. We use this established offline-to-online route to test a source-identical intervention: only teacher responses outside visited lexical support change while the entire alignment fit remains fixed. Leaf natural-retrieval tests measure the resulting increment, including a larger source-selected completion search; the E5 extension reproduces both the positive teacher increment and the source-precision/target-value mismatch in a second encoder setting.

Prior-centered regularization [31,34] supplies the retention estimator, while data-span projection work [32,33] motivates the source-null control. These tools expose retained behavior rather than adding an affine function class. Feature/readout separation under shift [9,10] motivates the adapted-prefix study. Its complete-ranking block swaps and crossed query/index comparisons test retrieval consequences; they do not identify semantic factors or a unique training-loss cause. Classical mean removal [35] supplies a strong unlabeled consumer control. Width/depth methods [2,3] motivate the resource axes, with dedicated-small and strong-affine comparisons distinguishing gains from sharing and gains over a stronger endpoint.

Query-performance prediction, ranker selection and budgeted execution [13–15,17,18,25] motivate the allocation studies. Complete-request measurement and static mixtures determine the paid comparison, while classical lexical fusion [37,38] and offline score compilation [39–41] change the available alternatives. Tangent adaptation [19], margin distillation [20] and candidate-conditioned pooling [21] supply related mechanisms for the supporting shallow-correction experiments. Appendix

J retains the detailed comparisons with all 41 references, including the distinction between established methods and the questions tested here.

8. Why information precision is not retrieval value

The source-identical intervention constructs several deployed retrievers that are indistinguishable to the observed alignment procedure. Their disagreement on retrieval shows that alignment fit alone does not specify the deployment function. Teacher responses on unvisited inputs choose how that function extends beyond the fitting support. The completion search makes this extension problem explicit: source validation can prefer one extrapolation rule while the deployment task prefers another.

For a fixed document index, coordinate changes affect rankings through document-pair margins. Let u be the normalized query output, $\tilde{u}$ the output after changing its table responses, and d_i, d_j two document vectors. The change in their score margin is exactly

$$\Delta m_{ij} = (\tilde{u} - u)^\top (d_i - d_j).$$

This follows by subtracting the two inner-product score differences. The ranking effect depends on the direction of the output change and the pre-existing margin; its relevance effect also depends on which documents cross the evaluation cutoff and their judgments. Equal-weight source-coordinate MSE contains none of these deployment-specific weights. Token averaging and output normalization further couple several row changes within one query. Consequently, replacing an erroneous row by an exact teacher response removes that row error but supplies no general sign guarantee for the relevance change under a fixed consumer.

The experiments identify this mismatch at two distinct levels. The 30-model searches select lower source-MSE completions whose target quality trails the original affine rule in both encoder settings. The response-budget panel then changes the information itself while freezing the completion and consumer: exact rows can move the ranking in an unfavorable direction. Both results concern how the deployed consumer uses information. An optimal policy could ignore an unhelpful response; a fixed pipeline instead passes every acquired row through the same pooling and ranking computation.

Teacher acquisition should therefore account for the retrieval decisions a response is expected to change. A source-only implementation could combine held-out query occurrence weights with document-direction sensitivity and near-boundary ranking margins, then freeze the criterion before evaluating a separate target population. This proposed selector remains to be evaluated against the source-error proxy and random orders tested here.

Retained information, offline acquisition and online computation interact through the consumer and deployment population. The prior, head-replacement, budget and depth controls show how changing one resource alters the value of another. This is why a stronger offline endpoint can change the meaning of a paid-computation comparison without eliminating the benefit of deep query processing. The practical baseline is the best justified use of available offline information, and the practical objective is retrieval quality at the complete-request cost.

The protocol and interpretation conditions are collected in Section 2 and Appendices A and K–O, including the fixed-fit uncertainty, coupled interventions, execution provenance, original-versus-

new interval record, resource accounts and artifact availability. The E5 C1–C4 replications and stronger-completion extension now have complete independently audited evidence.

9. Conclusion

Offline teacher information contributes retrieval quality beyond the observed alignment fit. Source-identical interventions isolate this contribution, which persists against a stronger source-selected completion search on three tasks across two encoder settings. Better source reconstruction and more exact responses do not reliably order the resulting retrieval quality. Their value depends on the rankings produced by the consumer. Retained-prior and readout interventions explain how useful behavior can be preserved or obscured, while matched-depth comparisons establish a substantial deep-query increment above the offline table. These findings make the information available before a request central to query compression: it defines the starting point for online computation and motivates acquisition guided by the retrieval decisions that information can change.

References

[1] Yuxuan Wang and Lyu Hong. 2023. Query Encoder Distillation via Embedding Alignment is a Strong Baseline Method to Boost Dense Retriever Online Efficiency. Proceedings of SustaiNLP, pp. 290–298. https://aclanthology.org/2023.sustainlp-1.23/

[2] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka Representation Learning. https://arxiv.org/abs/2205.13147

[3] Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2024. 2D Matryoshka Sentence Embeddings. https://arxiv.org/abs/2402.14776

[4] Zhuchenyang Liu, Yao Zhang, and Yu Xiao. 2026. NanoVDR: Distilling a 2B Vision-Language Retriever into a 70M Text-Only Encoder for Visual Document Retrieval. Version 1. https://arxiv.org/html/2603.12824v1

[5] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. https://arxiv.org/abs/1611.09268

[6] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. Proceedings of EMNLP, pp. 7534–7550. https://aclanthology.org/2020.emnlp-main.609/

[7] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. https://arxiv.org/abs/2104.08663

[8] Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. 2023. EmbedDistill: A Geometric Knowledge Distillation for Information Retrieval. Version 2. https://arxiv.org/abs/2301.12005

[9] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. ICLR. https://arxiv.org/abs/2202.10054

[10] Zheda Mai, Arpita Chowdhury, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, and Wei-Lun Chao. 2024. Fine-Tuning is Fine, if Calibrated. NeurIPS. https://arxiv.org/abs/2409.16223

[11] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. ICML, PMLR97:3519-3529. https://proceedings.mlr.press/v97/kornblith19a.html

[12] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token Merging: Your ViT But Faster. ICLR. https://arxiv.org/abs/2210.09461

[13] Niranjan Balasubramanian and James Allan. 2010. Learning to Select Rankers. SIGIR. https://ciir-publications.cs.umass.edu/getpdf.php?id=937

[14] Lingjiao Chen, Matei Zaharia, and James Zou. 2020. FrugalML: How to Use ML Prediction APIs More Accurately and Cheaply. NeurIPS. https://arxiv.org/abs/2006.07512

[15] Maria Vlachou and Craig Macdonald. 2023. On Coherence-based Predictors for Dense Query Performance Prediction. https://arxiv.org/abs/2310.11405

[16] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the Best Counterargument without Prior Topic Knowledge. ACL, pp. 241–251. https://aclanthology.org/P18-1023/

[17] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. ACM Transactions on Information Systems30(2), Article11. https://doi.org/10.1145/2180868.2180873

[18] Khaled Albishre. 2026. Risk-Aware Selective Neural Refinement for Reliable Dense Retrieval. IEEE Access14:92876–92893. https://doi.org/10.1109/ACCESS.2026.3705625

[19] Tian Yu Liu, Aditya Golatkar, and Stefano Soatto. 2024. Tangent Transformers for Composition, Privacy and Removal. ICLR. https://proceedings.iclr.cc/paper_files/paper/2024/file/7adf155d85da5d2ef23c307f999b1f98-Paper-Conference.pdf

[20] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation. arXiv:2010.02666, version 2 (2021). https://arxiv.org/html/2010.02666v2

[21] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. ICLR. https://arxiv.org/html/1905.01969v4

[22] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A Full-Text Learning to Rank Dataset for Medical Information Retrieval. ECIR, pp. 716–722. https://webserver.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/

[23] Nachshon Cohen, Yaron Fairstein, and Guy Kushilevitz. 2024. Extremely Efficient Online Query Encoding for Dense Retrieval. Findings of NAACL, pp. 43–50. https://aclanthology.org/2024.findings-naacl.4/

[24] Guangyuan Ma, Yongliang Ma, Xuanrui Gou, Zhenpeng Su, Ming Zhou, and Songlin Hu. 2026. LightRetriever: A LLM-based Text Retrieval Architecture with Extremely Faster Query Inference. ICLR. Author version 5. https://arxiv.org/html/2505.12260v5

[25] Siqin Huo, Andrew Parry, Charlie Clarke, and Debasis Ganguly. 2026. Optimal Re-Ranking Depth. SIGIR, pp. 3799–3804. https://eprints.gla.ac.uk/383304/

[26] Stephan Tulkens and Thomas van Dongen. 2024. Model2Vec. Software and method documentation. https://github.com/MinishLab/model2vec. DOI: 10.5281/zenodo.17270888.

[27] Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. ACL, pp. 4758–4781. https://aclanthology.org/2020.acl-main.431/.

[28] Prakhar Gupta and Martin Jaggi. 2021. Obtaining Better Static Word Embeddings Using Contextual Embedding Models. arXiv:2106.04302. https://arxiv.org/abs/2106.04302.

[29] Yingrui Yang, Yifan Qiao, and Tao Yang. 2022. Compact Token Representations with Contextual Quantization for Efficient Document Re-ranking. ACL, pp. 695–707. https://aclanthology.org/2022.acl-long.51/.

[30] Orion Weller, Dawn Lawrie, and Benjamin Van Durme. 2024. NevIR: Negation in Neural Information Retrieval. EACL, pp. 2274–2287. https://aclanthology.org/2024.eacl-long.139/.

[31] Xuhong Li, Yves Grandvalet, and Franck Davoine. 2018. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. ICML, PMLR80, pp.2825–2834. https://proceedings.mlr.press/v80/li18a.html.

[32] Sharan Vaswani, Reza Babanezhad, Jose Gallego-Posada, Aaron Mishkin, Simon Lacoste-Julien, and Nicolas Le Roux. 2020. To Each Optimizer a Norm, To Each Norm its Generalization. https://arxiv.org/abs/2006.06821.

[33] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. 2017. The Marginal Value of Adaptive Gradient Methods in Machine Learning. NeurIPS30. https://papers.neurips.cc/paper_files/paper/2017/hash/81b3833e2504647f9d794f7d7b9bf341-Abstract.html.

[34] Ilja Kuzborskij and Francesco Orabona. 2013. Stability and Hypothesis Transfer Learning. ICML, PMLR28(3), pp.942–950. https://proceedings.mlr.press/v28/kuzborskij13.html.

[35] Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. ICLR. https://openreview.net/forum?id=HkuGJ3kCb.

[36] Yiling Chen and Bo Waggoner. 2016. Informational Substitutes. FOCS; full version, 2017. https://arxiv.org/abs/1703.08636.

[37] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. SIGIR. https://cormack.uwaterloo.ca/cormacksigir09-rrf.pdf.

[38] Xing Han Lù. 2024. BM25S: Orders of Magnitude Faster Lexical Search via Eager Sparse Scoring. https://arxiv.org/abs/2407.03618.

[39] Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient Open-Domain Question Answering via Sparse Transformer Matching Retrieval. NAACL-HLT, pp.565–575. https://aclanthology.org/2021.naacl-main.47/.

[40] Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning Passage Impacts for Inverted Indexes. SIGIR. https://doi.org/10.1145/3404835.3463030.

[41] Roger Weber, Hans-J. Schek, and Stephen Blott. 1998. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. VLDB, pp.194–205. https://www.vldb.org/conf/1998/p194.pdf.

Appendix A. Fixed-index protocol, exposure and reproducibility

A.1 Retrieval setting

Let $d_1, \ldots, d_N \in \mathbb{R}^D$ be fixed document vectors, $q_T(x)$ the teacher query vector, and $q_\theta^{(b)}(x)$ a student vector at computation budget b. Scores are inner products. A budget specifies an encoder depth, a readout width, and a numerical precision. The final engineering objective is judged retrieval quality at an actual request cost, including any computation used to choose the budget.

We use a six-layer Leaf query encoder with hidden width 384 and a learned 768-dimensional Dense output, paired with a 12-layer Arctic document encoder. The deployed teacher uses per-channel dynamic INT8 Linear operators, with floating-point pooling, projection, and normalization. We call this query precision PC. Documents are FP32 CLS embeddings with L2 normalization. A fixed projection used by the existing retrieval system is retained throughout.

The cheap representation is a frozen token-attention pool with a rank-16 residual applied to the second-layer hidden states. An affine readout maps its 384 coordinates to the retrieval space. An orthonormal basis of the query Dense support contains 385 active coordinates. Width 400 pads this support with 15 zeros for the measured physical operator. Narrow widths use the first 128 or 256 coordinates of the fixed document-PCA rotation. Document coordinates are not renormalized after projection or truncation.

A.2 Query populations and document indices

The teacher-fidelity studies use query text from the official MS MARCO training-query source [5], evaluated against the complete 5,183-document SciFact geometry [6]. These are distillation-transfer measurements on a fixed index, not MS MARCO relevance evaluations. We canonicalize query text by Unicode normalization, case folding, and whitespace normalization, then select distinct groups by a fixed hash order. The nested training sets contain 128, 512, or 2,048 groups.

The initial staged comparisons fix their models before opening each new source panel. The initial objective study uses 512 held-out groups. Joint-width, prefix-training, and pooled-capacity studies each use a later, disjoint set of 512 groups. The alignment study uses the next 1,024 groups and reevaluates the previously fixed competitors on the same queries. The initially reserved 512 groups are subsequently opened for the common-readout mechanism study, after all twelve new readouts are fixed. The models can share a development history, but these later source panels are disjoint from their fitting queries and earlier source panels. The subsequent prior-centered and source-null interventions revisit exposed panels and are retrospective mechanism tests. Their source-only fitting and fixed comparisons do not make those panels newly independent.

Source query count is our operational measure of coverage. It increases the number of sampled query contexts from one distribution. It does not independently manipulate semantic diversity or sequence length. Effective source lengths in the initial experiment range from 12 to 50 tokens, including the query prefix.

For cross-domain analysis, we evaluate the same 48 fixed source-study endpoints in four cells: 1,024 MS MARCO queries and 1,406 natural ArguAna queries, each scored against complete SciFact and ArguAna indices. The ArguAna index contains 8,674 documents encoded by the same frozen document model. All readouts retain the source coordinate map, including its document-PCA rotation; there is no target refit, new checkpoint selection, or normalization choice after target results. The 48 endpoints include the original PC6 teacher, two warm readouts, 12 raw/normalized

ridge variants, 18 factorial models, three pooled MLPs, three old affine controls, and nine prefix controls. Each query is paired across indices, while the two query populations contain different texts and are not paired with one another. This design changes task, query lengths, corpus size, and document geometry; it does not identify a pure corpus-size effect.

A.3 Metrics and uncertainty

Teacher fidelity is measured by KL over all documents, raw and centered score MSE, teacher top-10 overlap, and top-1 agreement. Top-10 overlap is the fraction of teacher top-10 documents recovered by the student. Ranking ties use UTF-8 document identifiers. Seed means describe uniform random selection among the three fixed fitted models, not an ensemble of their vectors.

Source intervals use 4,000 paired query bootstrap draws. They are pointwise and conditional on the fitted models. They quantify variation across the evaluated query population, not uncertainty over all possible training seeds. The objective study's sign reversal is presented through its prespecified pairwise contrasts; we do not substitute a retrospectively selected omnibus significance test.

Judged relevance uses historical SciFact development material with 427 queries and 484 positive judgments. The primary deployed comparison uses 142 queries in 90 components formed by joining prior semantic-group and shared-relevant-document links, with 4,000 component-bootstrap draws. These labels were exposed in earlier research. The remaining historical rows retain their declared training and out-of-fold roles. We therefore use relevance to test the direction of practical benefit, with fresh source teacher fidelity providing a separate generalization measurement.

Natural ArguAna relevance is evaluated only on its own document index. We retain all 1,406 queries, including five whose judged positive document IDs are absent from the corpus; their missing positives remain in the relevance denominators. Exact same-ID document exclusion applies to 1,298 queries. Intervals use 4,000 paired resamples of 175 historical groups. ArguAna was exposed in earlier research, so this is a frozen-model transfer test rather than a fresh benchmark claim. Its 25 earlier production static endpoints are also rescored, yielding 73 human-quality endpoints in this stage. Those older heads were fitted on the historical 427-query source material, whereas the new study models inherit their declared source-fold or MS MARCO fitting populations.

A.4 A later, group-separated validation task

We subsequently freeze a paid continuation policy before consuming new original ArguAna validation queries [16]. This is a separate within-task validation, rather than another dataset or a previously untouched version of the official BEIR test. The original archive contains 1,290 validation points and 1,287 paired counterarguments. We remove unpaired points and entire connected debate components that overlap the exposed original test through query/counterargument text or truncated token inputs. Deduplicating the remaining effective inputs yields 973 queries in 134 connected groups. All are also disjoint from the 6,656 previously consumed source-query inputs. Their text already exists in the fixed document index; the new exposure is as queries in this research.

Every new query has exactly one indexed positive and one exact-content self row, which is excluded. This canonical-self rule differs from the earlier official query-ID exclusion: 108 old query copies lack their own ID in the index but have an exact-text self row under another ID. The new result therefore changes both query population and this task contract. We preserve the official old results and test the self-exclusion difference separately, with old vectors and labels fixed. Unknown model-pretraining exposure remains outside the separation claim.

The frozen comparison contains 125 existing static endpoints and eight fixed paid policies. The primary observes the unit-scale standard deviation of the cheap top-ten scores and continues below .016067809465228212. Its threshold is the source E512 60th percentile; choosing this feature and percentile used the exposed old ArguAna population. The final policy, competitors, 973-input population and identifier-hash 64-query cost panel are fixed before any new output. No model, document representation or readout is refitted. Quality intervals resample 134 connected groups. An initial actual-cost experiment uses the 52 groups in the fixed 64-query panel. A subsequent execution measures all 133 conditions on all 973 queries with an equivalent ten-address observer; its intervals resample the same 134 groups. The gate and model bank remain fixed after quality becomes known.

A.5 A prospective graded task

After the shallow-correction models are fixed, we evaluate all 323 official NFCorpus test queries against its complete 3,633-document BEIR corpus [7,22]. Eleven existing static endpoints and the unchanged std-q60 continuation gate are selected before target payload access. They include original and adapted-prefix readouts, two original ridge widths, native PC4/PC6 and adapted continuation. No target relevance or timing selects models, seeds or thresholds. Predictions and complete-request costs are saved before grade-bearing label parsing.

The original 11,758 grade-1 and 576 grade-2 judgments enter nDCG with linear ordinal gains; recall and MRR count positive grades. All queries have judgments. Effective query lengths, including the fixed instruction, average 15.0 tokens and range from 11 to 36. Domain, index and judgment structure also change. Two corpus documents exactly match SciFact inputs; no normalized query text matches the 2,048 fitting queries. All document IDs remain in retrieval. Four thousand paired-query resamples describe uncertainty conditional on the models, without assuming that unique query URLs imply independent topics.

A.6 Training and representation details

The KL-based readout experiments use AdamW with learning rate .001, weight decay .0001, batch size 32, global gradient clipping at 1, and no dropout or learning-rate schedule. Each fit performs 500 updates. Training permutations are determined by seeds 20261018, 20261019, and 20261020. The trainable prefix uses learning rate 2×10^{-5} for the first two Transformer layers and .001 for its affine head. The shared objective averages shallow and continued-deep KL with equal weight. The later layers remain fixed but transmit gradients to the prefix. The subsequent coordinate intervention retains these settings and substitutes half squared Euclidean distance in the fixed orthonormal 385-dimensional teacher support; the fifteen padding outputs are structurally zero. Shared coordinate training averages the shallow and continued-deep losses. The nine new fits use 1,262.37 CPU seconds, with no model selected from their evaluation outcomes.

Alignment uses dimensionless penalties $\{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. The actual ridge penalty multiplies this value by $\mathrm{tr}(X_c^\top X_c)/384$, computed inside each training fold. The fold assignment hashes the raw query identifier with a fixed namespace, and exact cross-validation ties select the smallest penalty. Width 256 fits 98,560 affine parameters; width 400 fits 148,225 active parameters. The float64 teacher reconstruction error in the shared support basis is at most 3.42×10^{-8} on the 2,048 fitting queries.

For the Leaf-based experiments, query texts use the prefix “Represent this sentence for searching relevant passages:”, right truncation, and a maximum length of 512 tokens. The query checkpoint is

`MongoDB/mdbr-leaf-ir`, available from the official fixed-revision file tree at revision 4262131b32c3182bd06e67e92ae69d7bd66e0c5c; the Leaf and Dense SHA256 identities begin `82691b5531ec8323` and `b3e7c0e1ef65e39a`, respectively. Documents use Snowflake/snowflake-arctic-embed-m-v1.5 at revision e58a8f756156a1293d763f17e3aae643474e9b8a, title plus text, no document prompt, FP32 eager execution, CLS pooling, and L2 normalization. The complete SciFact document asset has SHA256 a254d089f66ce379fcdbb04f0bc766ce3ef81ccec9949115912d91cc69a24c38. The historical download ledger records successful byte/hash verification of the Leaf weights, Dense head, tokenizer and supporting configuration files. The public file tree identifies the checkpoint; it does not establish that every source array or transitive dependency is publicly accessible, or that the four core comparisons have been independently replayed from a clean environment.

A.7 Experimental sequence and integrity

The normalized-query hash sequence allocates positions 0–2,047 to training. The original objective evaluation uses 2,048–2,559. Positions 2,560–3,071, initially reserved, become the fresh common-readout mechanism panel after its twelve fits seal. Joint-width, trainable-prefix, and pooled-capacity evaluations use 3,072–3,583, 3,584–4,095, and 4,096–4,607. Alignment evaluation uses 4,608–5,631. The coordinate-objective intervention uses the next 1,024 groups, positions 5,632–6,655, only after its nine training fits and twelve ridge fits finish. All positions are zero-indexed and inclusive in this description.

Independent numerical audits reconstruct the initial objective metrics, joint-width comparisons, prefix fidelity, and the 100-mode historical serving study from their saved arrays. The serving audit performs 4,012 checks, including quality metrics, timing aggregation, paired comparisons, bootstrap intervals, and independently computed static convex envelopes. Its largest floating-point discrepancy is 4.66×10^{-15}. The subsequent 89-mode serving audit independently reproduces 5,572 checks, including its 70-endpoint static envelopes, with maximum discrepancy 1.11×10^{-16}. A separate independent alignment audit reconstructs all 49,152 source endpoint score/rank fingerprints, with exact equality, and reproduces metrics and paired intervals to numerical tolerance. The natural-transfer audit independently reconstructs 134,976 full score/rank fingerprints across the two indices, all 73 by 1,406 human-quality arrays, and all 23 prespecified contrasts in four metric families. The largest interval discrepancy is 2.9×10^{-15}. It checks the stored encoder provenance and outputs without repeating their forward passes. The completed 101-mode natural serving audit also reconstructs all 284,012 human-quality values, the three-process cost aggregation, and the prior-86 and older-52 static envelopes. Its largest numerical difference is 1.39×10^{-16}; it verifies the recorded analysis rather than repeating the latency experiment.

The runtime comparison uses one warmup and two timed repeats per query/mode in each of three fresh serial processes. Per-query costs are the median of the two repeats, averaged over the processes. All 85,200 timed requests reproduce the corresponding warm output arrays. Independent auditing verifies the saved timing and quality analyses, while 24 older static modes lack stored full-vector fingerprints for independent reconstruction of the aggregate request hash.

Resource failures remain in the experimental record. The alignment analysis retained full ranking arrays through top-10 slice views and reached 3,357,280 KiB peak RSS. Copying only the top-10 slices preserves every numeric result and reduces the successful replay peak to 669,544 KiB. No query or document was re-encoded during recovery. Successful replay uses 30.58 CPU seconds. Complete CPU totals for the failed attempts were not captured and are not represented as zero. Scientific conclusions use the same predeclared endpoints and comparisons throughout recovery.

The natural-query encoding retained the initial execution failure: at the first of eight preselected direct-forward checks, the qualification-only head call omitted its required width argument. The correction adds the intended width 400 without changing vector generation, and preserves the first 50 complete rows. A supplemented execution freeze records the change before result analysis. The resumed 1,356 rows use 170.10 CPU seconds. The failed process's saved first-50-row total of 4.83 CPU seconds is only a lower bound, not its complete resource cost. Eight direct-check queries then pass all 32 encoder/readout equivalence assertions. The complete ArguAna document matrix is reused; no document is re-encoded.

Appendix B. Supervision, inherited priors and capacity

B.1 What restricted targets omit

The teacher distribution over the corpus is

$$p_T(j \mid x) = \frac{\exp(q_T(x)^\top d_j/\tau)}{\sum_{k=1}^{N} \exp(q_T(x)^\top d_k/\tau)}, \qquad \tau = 0.05.$$

Full-distribution distillation minimizes $D_{\mathrm{KL}}(p_T \| p_\theta)$. Candidate distillation selects a set $S(x)$ and minimizes KL between separately renormalized distributions on that set. Our candidate set combines the teacher's top 32 documents with 32 documents selected by a fixed hash rule, with duplicates removed.

The KL chain rule gives a useful interpretation. Write $a = p_T(S)$ and $b = p_\theta(S)$. Then

$$\begin{aligned} D_{\mathrm{KL}}(p_T \| p_\theta) = {} & D_{\mathrm{KL}}(\mathrm{Bern}(a) \| \mathrm{Bern}(b)) \\ & + a D_{\mathrm{KL}}(p_T(\cdot \mid S) \| p_\theta(\cdot \mid S)) \\ & + (1-a) D_{\mathrm{KL}}(p_T(\cdot \mid \bar{S}) \| p_\theta(\cdot \mid \bar{S})). \end{aligned}$$

Candidate-only training directly supervises the conditional term on S. It leaves the total probability assigned to S and the relative distribution on its complement unspecified. This is a change of objective, not an unbiased estimate of the full-softmax gradient. We use the standard identity to locate observed errors. Its terms do not by themselves establish which objective will generalize better.

B.2 Embedding alignment as a strong comparator

Query embedding alignment is an established distillation approach [1,4]. It transfers the teacher vector directly and avoids constructing document-score targets during fitting. To test that approach within the same cheap affine family, we solve centered multivariate ridge regression:

$$\min_{W,c} \sum_{i=1}^{n} \|W z_i + c - y_i\|_2^2 + \lambda \|W\|_F^2.$$

Here z_i is the frozen PC two-layer pooled vector and y_i contains the teacher coordinates retained at the chosen width: 256 at width 256, or all 385 support coordinates at physical width 400. The intercept is unpenalized. For each training-query count and width, five folds within the training set

select a penalty from a fixed logarithmic grid. We then refit on all training rows with a float64 SVD solver and deploy FP32 weights. The coordinate map is shared with the preceding experiments and was constructed using the existing document geometry. The alignment fit itself requires no document scores.

We also retain a fixed inherited weight matrix W_0 by replacing the penalty with $\lambda\|W - W_0\|_F^2$. Centering on pretrained parameters is established transfer regularization [31]. Both objectives leave the intercept free and use the same features, source folds and eight-penalty grid. We compare equal penalties and each center's separately selected penalty; target outcomes select neither. The 256-coordinate target is a truncation of the teacher support, not a lossless full teacher vector.

For row-oriented centered features X_c, let $H_\lambda = X_c^\top X_c + \lambda I$. At positive λ, the difference between the two coefficient matrices is

$$\widehat{W}_{\text{prior}}^\top - \widehat{W}_{\text{zero}}^\top = \lambda H_\lambda^{-1} W_0^\top.$$

The unpenalized intercept subtracts the source-feature mean times this difference. Thus this penalty retains the prior slope, while recalibrating its offset: two prior heads differing only in their intercept give the same fitted function. In a source singular direction with value s, the retained prior weight is $\lambda/(s^2 + \lambda)$; a source-null direction retains it fully. The Hessian, derivative with respect to source targets, and effective degrees of freedom are identical at a common penalty. Thus the controlled change concerns prior-dependent bias, not additional capacity or reduced conditional estimation variance. At zero penalty we use the corresponding minimum-change pseudoinverse solution.

We retain both raw and query-normalized predictions. Positive rescaling preserves real-number ranking, while changing the softmax temperature implicit in the scores. Reporting both separates agreement in ranking from agreement in score scale.

B.3 Objective and capacity controls

The main factorial crosses three query counts, two objectives, and three seeds. All 18 models use the same frozen PC2 pool, 256-coordinate affine readout, initialization, and 500 AdamW updates of 32 queries each. The two objectives are candidate KL and full-corpus KL. Increasing query count reduces repeat exposure to individual queries while preserving 16,000 total presentations. The deployed computation is identical across these objective arms. Offline arithmetic and optimization convergence are not equalized.

The alignment comparison fits six deterministic models at the same three query counts and widths 256 and 400. Training-only folds select its penalty. This supplies a stronger optimization baseline for comparison with the preceding 500-step KL fits, whose global optimality is unknown.

We then study where to train additional capacity. Nine models cross three seeds with a frozen prefix, a dedicated trainable two-layer prefix, and a trainable prefix with equal shallow/deep KL objectives. The pool, embeddings, later four layers, and native Dense output remain fixed. Frozen-prefix training updates 148,225 affine parameters. The dedicated and shared variants update 3,697,153 parameters. These fits train in FP32 and are subsequently converted to PC for deployed evaluation.

A separate pooled-capacity control adds a 384-to-4,608-to-384 GELU residual before the affine output, totaling 3,692,161 trainable parameters. Its encoder and pool remain unchanged, and its residual output initializes to zero. It trains on cached PC features for the same 500 updates. This

nearly matches parameter count while changing optimization geometry, computation, and training precision.

B.4 The full-versus-restricted objective reverses

Table S1 establishes the support-by-objective interaction within the fixed KL training family. With 128 training queries, the full objective reaches .2184 teacher overlap, below the candidate objective's .2415. At 2,048 queries, the ordering reverses: .3408 versus .2556. The full-minus-candidate contrast moves from −.02318 with a 95% interval [−.03360, −.01315] to +.08522 with interval [.07630, .09453].

Distinct training queries	Candidate KL	Full-corpus KL	Candidate overlap	Full-corpus overlap
128	.473723	.507304	.241536	.218359
512	.478905	.437474	.243424	.253711
2,048	.430056	.307780	.255599	.340820

Table S1. Full-teacher KL and teacher top-10 overlap on the same held-out 512 queries. Values average three fixed training seeds; lower KL and higher overlap are better.

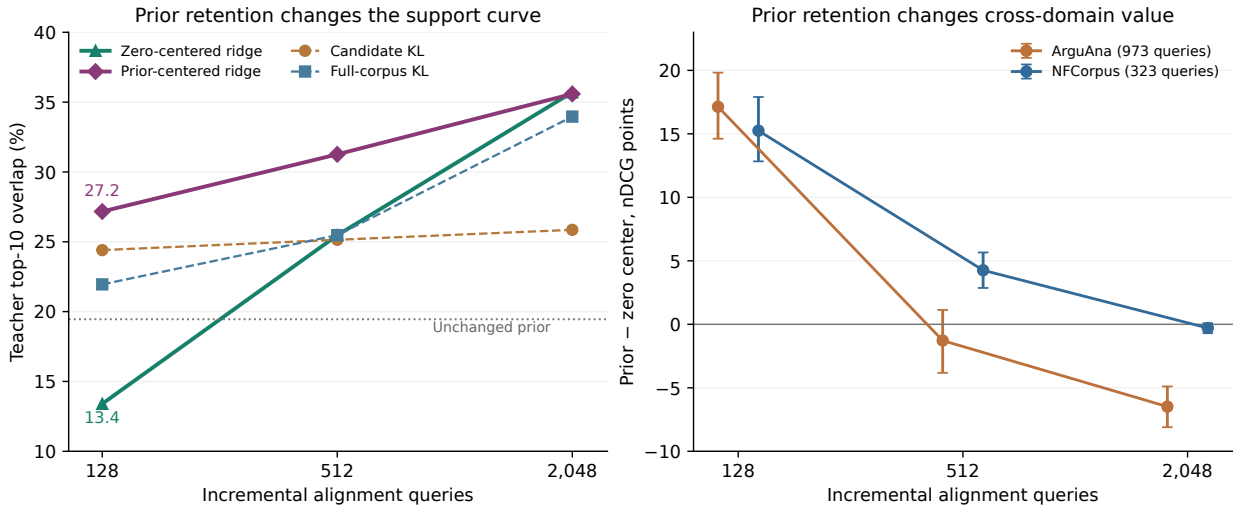


Figure 3: Retaining prior information changes the apparent support requirement and its cross-domain value. Left: teacher overlap on the 1,024-query panel; each ridge center uses its own source-selected penalty. The fixed KL competitors and all heads use width 256; the dotted line is the unchanged prior. Right: prior minus zero center at the same zero-center-selected penalty, with pointwise 95% paired intervals on historical ArguAna973 and NFCorpus323. Support is incremental to the shared pretrained encoder, earlier learned pool and fixed geometry. The prior-centered comparisons revisit exposed populations.

At 2,048 queries, full-teacher KL falls by 28.4% relative to candidate training. Increasing support from 128 to 2,048 improves full-KL overlap by .12246, compared with .01406 for the candidate objective. Both experiments use the same number of updates. The gain therefore reflects how repeated exposure is distributed over distinct queries and how their targets constrain the readout.

Training losses help interpret the small-support failure. The 128-query full-KL fits achieve training KL between .0254 and .0321, yet held-out KL stays near .507. At 2,048 queries, training KL is higher, about .211, while held-out KL drops to .308. The small-support failure is compatible with a training/test gap, rather than an inability to optimize the full objective on its observed queries.

B.5 Most of the KL improvement occurs outside the candidate set

The initial evaluation's candidate sets contain a median .1132 of teacher probability mass. Their selected documents can be highly ranked while still covering little of the softened teacher distribution. At 2,048 queries, the full objective improves total held-out KL by .12228. Of this arithmetic difference, .09900 comes from the teacher-mass-weighted complement term, .01274 from the weighted candidate term, and .01054 from mass allocation. The complement accounts for 81.0% of the total difference.

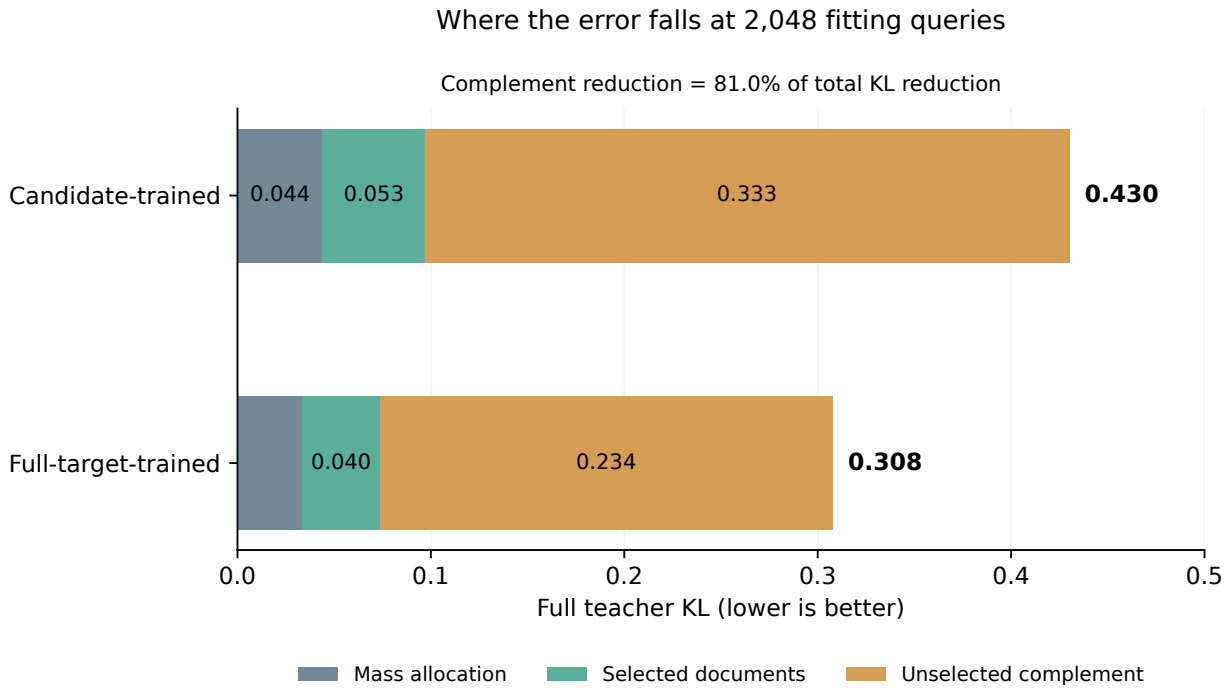


Figure 4: The full objective improves all three KL components at 2,048 training queries. The complement term accounts for 81.0% of the total reduction. The decomposition is an exact accounting identity applied after the experiment.

Conditional error inside the selected set also falls, from .3492 to .2638 before teacher-mass weighting. Full supervision thus better matches the selected-set distribution while learning information that the restricted objective omits. At 128 queries, all three held-out components are worse for full KL. In this KL family, the additional supervised terms improve held-out predictions at larger support. The prior-centered control below tests whether the same support requirement extends to another way of retaining teacher information.

B.6 Preserving the prior changes the small-support result

Incremental fitting queries	Candidate KL, 256	Full KL, 256	Zero-centered ridge, 256	Prior-centered ridge, 256
128	.244043	.219499	.133984	.271680
512	.251432	.254753	.254688	.312598
2,048	.258561	.339616	.357031	.355859

Table S2. Teacher top-ten overlap on the same 1,024-query panel. KL values average three fixed seeds. Each deterministic ridge center selects its penalty using only its fitting queries. The prior-centered column is a later retrospective control; its targets retain 256 teacher coordinates. Counts are incremental supervision, not total training budgets.

The next 1,024 queries reproduce the KL reversal using the original fixed models. Zero-centered ridge reaches .3570 overlap at width 256 and 2,048 fitting queries, exceeding full KL by .01742 [.01221,.02279]. Its full-teacher KL is also lower, .2836 versus .3102. At 128 queries, however, the same ridge procedure gives only .1340 overlap. That small-support failure changes when regularization retains the existing affine prior (Table S2).

The prior is exactly the projected native Dense head, with the fixed calibration scale used to initialize every KL fit. It is not a later head fitted to 128, 512 or 2,048 queries. Both ridge centers share a pool previously trained on 140 historical SciFact queries, projection and family selection using the earlier 427-query development population, and document-PCA geometry from the fixed index. The pool fit plus its held-row prediction took 1.66 CPU seconds from cached features; encoder-feature construction and upstream model pretraining are additional inherited costs. The new 15-configuration centered-ridge comparison, including source cross-validation, takes 4.52 CPU seconds from existing arrays and adds no teacher forward for its prior.

At 128 queries, holding the dimensionless penalty at the zero-centered choice, 1.0, raises overlap to .2673: a gain of .13330 [.12383,.14238]. Selecting the prior-centered penalty on source folds chooses 10 and yields .2717. This is a large change within the same globally solved affine family. The unchanged prior gives .1945; keeping its weights and fitting only a source-mean intercept gives .2420. Prior-centered source selection adds .02969 [.02490,.03447] beyond that intercept-only control. Rich targets are useful with few incremental queries when their fit preserves an appropriate existing function.

A source-null control identifies one part of this effect. The centered 128-row feature matrix has rank 127. Adding only the prior component orthogonal to that source span leaves every source prediction unchanged to float64 numerical precision, yet raises held-out overlap from .1340 to .2322. The same-penalty prior-centered solution adds another .03506 [.02979,.04033]. At 512 queries the input matrix is full rank and prior centering still improves overlap by .05791 [.04902,.06689]. Exact nullspace retention therefore explains neither all of the small-support gain nor the entire prior effect; weakly constrained observed directions also matter.

Prior retention has a different value on natural tasks. At 128 queries and the same penalty, ArguAna973 nDCG rises from .0940 to .2653 and NFCorpus323 from .1358 to .2883. Much of this improvement preserves existing quality: the unchanged prior already gives .2664 and .2492, and its source-intercept update gives .2717 and .2797. Prior-centered source selection reaches .2813 and .2905, exceeding the intercept controls by .00957 [.00097,.01845] and .01083 [.00528,.01695]. These smaller increments are distinct from the large comparison against zero-centered ridge.

At 2,048 queries both centers select penalty .1, and source overlap is nearly unchanged (.3570 versus .3559). On ArguAna, prior centering lowers nDCG from .4119 to .3471, a difference of -.06482 [-.08103,-.04893]; NFCorpus changes from .3162 to .3133 with an interval spanning zero. Both centers' quality curves improve with more queries. This is a relative cost of retaining the prior, not evidence that more data damages either model. Source coordinate cross-validation at this support also favors the zero center; penalties were selected separately within centers, so the experiment does not show that joint source selection would choose the harmful alternative. All intervals condition on fixed fits and resample queries or the 134 ArguAna debate groups.

A second query checkpoint tests how much this effect depends on the retained function. We use Arctic-m-v1.5, the existing document tower, as the query encoder too. Normalized CLS features from its fixed second and sixth layers predict normalized final-layer CLS in the original 768-dimensional document space. The identity center is an explicit early-exit hypothesis, rather than a trained Dense readout. At 128 source queries, sixth-layer identity centering raises overlap from .0752 to .1250 on the next 512 unused source groups, a gain of .04980 [.04336,.05625]. On historical NFCorpus323 it raises nDCG from .0411 to .1043 and exceeds a fitted scalar-plus-intercept control's .0861 by .01814 [.00701,.02968]. Second-layer identity centering changes neither metric appreciably. At 2,048 queries the sixth-layer differences shrink to about .0033 on each metric. Both centers select the same penalties at each support, and all fixed depths and incumbent controls are retained. The original twelve-layer model reaches .3631 NFC nDCG. This comparison extends the prior test to another query backbone while showing that a shared coordinate dimension alone supplies little guidance about which early readout to retain.

The original width-400 zero-centered results remain .1435, .2754 and .3812 across the three support counts. Query normalization leaves all overlap scores unchanged but can worsen teacher KL; at 2,048 queries and width 256, KL changes from .2836 to .3340. These controls separate prior-dependent extrapolation, ranking fidelity and score calibration.

B.7 Source-null optimizer drift is only one contribution

Adaptive optimization can move a linear predictor outside the training-feature span [32,33]. We test that channel in the existing 18 KL models by projecting only the change from their weight-decayed initial affine function onto the augmented source span, including the intercept. This established type of data-span intervention [32] preserves the source predictions. All original source top tens remain unchanged, and the 512- and 2,048-query full-rank controls are exact identities.

At 128 queries, projection changes the full-minus-candidate overlap gap on the original E512 panel from -.02318 to -.01855. Its improvement is .00462 [.00046,.00859], while most of the original difference remains. The final raw gradient has negligible source-null energy, whereas its Adam-preconditioned direction does not. This classical projection quantifies one optimizer channel while leaving most of the reversal unexplained. Bias inside the observed span remains a possible contributor, and Adam itself is unchanged. Together with prior-centered ridge, the result narrows the claim: the tested KL family has a real support interaction, but that interaction is insufficient to infer a general coverage requirement for rich teacher supervision.

B.8 Token-level adaptation improves shallow representations

On the prefix study's fresh 512-query panel, training the first two layers raises PC shallow overlap from .3408 to .3828. The paired gain is .04206 with interval [.03509, .04902]. Sharing the objective with the continued deep representation yields .3812 shallow overlap and improves the continued

deep overlap from .8073 to .8428. The shared-versus-dedicated deep gain is .03555 with interval [.02936, .04193].

Prefix training	Shallow teacher KL	Shallow overlap	Continued deep overlap
Frozen prefix, trained affine	.300073	.340755	N/A
Dedicated shallow objective	.250368	.382812	.807292
Shared shallow/deep objective	.251796	.381185	.842839

Table S3. Three-seed PC means on the prefix study's disjoint 512-query panel. A shared objective reduces drift after the changed prefix. Its continued deep output remains a changed model, not the original PC6 teacher.

The stronger alignment baseline narrows the apparent benefit of changing the encoder. On the later common 1,024-query panel, ridge400 reaches .3812 overlap and the dedicated prefix reaches .3931. The remaining difference is .01191 in favor of the prefix, with paired interval [.00544, .01813]. The earlier KL-trained affine reaches .3516. Thus a substantial portion of the gap attributed to head-only learning can be removed by a stronger affine fit.

The large pooled MLP is an informative control. On its own fresh 512-query panel, it attains .3408 overlap, below both the preceding affine's .3596 and the dedicated prefix's .4001. Its final training KL is also higher than the affine control by .0213, .0218, and .0305 across seeds. Since the MLP contains the affine subfamily, this outcome exposes an optimization or parameterization limitation of the tested recipe. It cannot establish that pooled information is insufficient. The comparison motivates improving the cheap fit before paying for token-level adaptation.

B.9 Shared widths must face dedicated small models

Joint training at widths 128, 256, and 400 improves narrow prefixes relative to truncating a head trained only at width 400. On a separate 512-query panel, joint training raises overlap by .01406 at width 128 and .00612 at width 256 relative to those truncation controls. Dedicated small models remain stronger controls: joint versus dedicated overlap is .28294 versus .28815 at width 128 and .35208 versus .35221 at width 256.

These results separate useful parameter sharing from a superior quality–cost frontier. A nested representation can offer several usable budgets while a dedicated model still defines the best point at a particular budget. Comparisons against post-hoc truncation alone would overstate the benefit.

Appendix C. Consumer replacement, coordinate geometry and input controls

C.1 The coverage effect reaches judged retrieval

The natural ArguAna task preserves the larger-support advantage of complete supervision (Table S4). Full-minus-candidate nDCG changes from −.01170 at 128 fitting queries, with interval [−.02387, .00070], to +.03153 at 512 queries and +.06235 at 2,048. The latter two intervals are

[.01837, .04545] and [.04968, .07555]. The small-support natural-quality contrast is uncertain even though its point estimate has the same sign as the source-fidelity result.

Distinct fitting queries	Candidate KL, 256	Full KL, 256	Ridge, 256	Ridge, 400
128	.245392	.233696	.090208	.087981
512	.230527	.262060	.262943	.295268
2,048	.211748	.274098	.387051	.419567

Table S4. Natural ArguAna nDCG@10 for fixed source-fitted models. KL columns average three fixed seeds. Ridge values use the raw prediction; the prespecified unit-normalized versions return the same ordered top ten and identical relevance metrics. No model is refitted on ArguAna.

At 2,048 fitting queries, width-256 ridge improves on full KL by .11295 nDCG, with interval [.09544, .13133], and by .15837 Recall@10. Width-400 ridge reaches .4196 nDCG and .6693 Recall. Its source teacher-fidelity advantage over the old affine head therefore survives a substantial change of retrieval task. The teacher and strong deep endpoints still retrieve better: original PC6 reaches .5779 nDCG and .8549 Recall; the highest observed value among the 25 prior static controls is .5818. The alignment result concerns the cheap representation family, not the highest achievable task quality.

C.2 Prefix adaptation wins at source fidelity, alignment wins after joint transfer

On the later 1,024-query source panel, dedicated prefix adaptation exceeds width-400 ridge in teacher top-10 overlap: .3931 versus .3812. Natural ArguAna reverses this comparison. Ridge reaches .420 nDCG, dedicated-prefix adaptation .271, shared-prefix adaptation .278, and the matched frozen-prefix head .250. The ridge-minus-dedicated difference is .14890, with interval [.13009, .16911]; the ridge-minus-shared difference is .14117, with interval [.12169, .16148].

Dedicated and shared adaptation still improve on their matched frozen-prefix control. The practical question is whether those gains outperform a better-solved fixed-feature baseline. The alignment comparison changes the target, regularization, optimizer, and trained capacity together. It supports a distribution-dependent ordering of these complete training procedures, rather than a causal claim against any single loss or against token-level training.

Query population	Document index	Ridge, 400	Dedicated prefix	Ridge minus prefix
Source MS MARCO	SciFact	.381152	.393066	-.011914
Source MS MARCO	ArguAna	.388672	.391406	-.002734
Natural ArguAna	SciFact	.328236	.319654	+.008582
Natural ArguAna	ArguAna	.523969	.403580	+.120389

Table S5. Teacher top-10 overlap for the same fixed models in the crossed evaluation. Dedicated-prefix values average three fixed seeds. The natural-query contrasts use historical-group bootstrap; source-query contrasts use the source query bootstrap.

The four cells locate the largest change in the natural query–index pairing (Table S5 and Figure 5). The source-cell difference has interval $[-.01813, -.00544]$. Changing only the index gives $[-.00970, .00433]$; changing only the query population gives $[-.00328, .02053]$. With both natural queries and their index, the difference is $+.12039$ with interval $[.10094, .14099]$. The two mixed cells are not individually decisive. This pattern motivates testing adaptation jointly against query and document shifts; it does not establish a semantic explanation for their interaction.

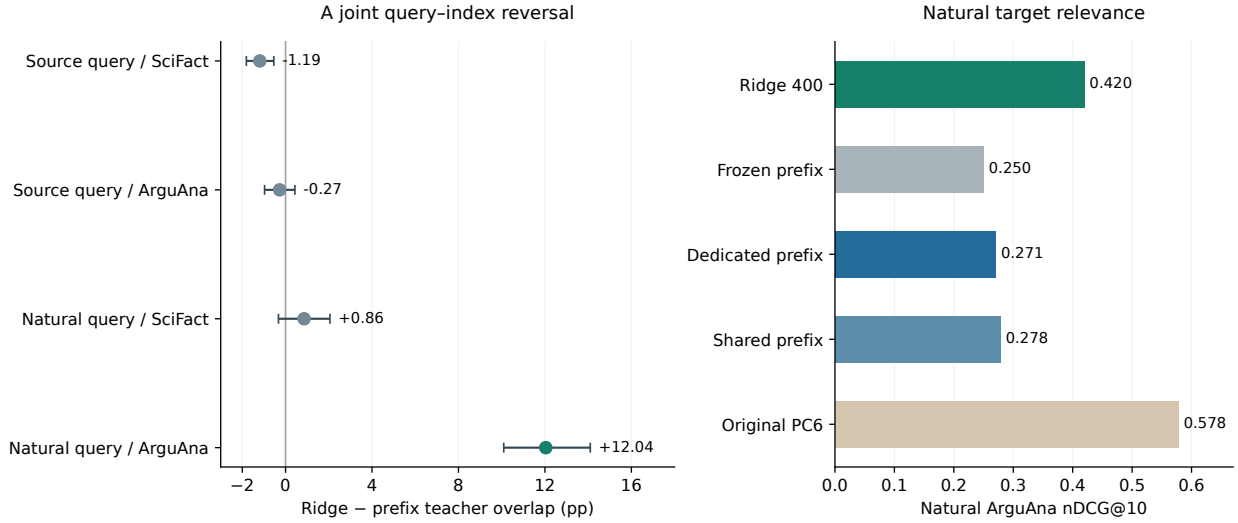


Figure 5: Transfer changes the alignment–adaptation comparison. Left: the teacher top-10 overlap difference between width-400 ridge and dedicated prefix in the four fixed-model cells. Right: natural ArguAna relevance for ridge, the three prefix controls, and the original PC6 reference. These panels show different metrics; neither is a matched-cost frontier.

C.3 Distribution calibration explains an apparent fidelity conflict

Raw teacher KL initially suggests the opposite ordering from target ranking and relevance. On natural ArguAna, width-400 ridge has KL 1.8879, worse than the dedicated prefix's 1.5486, despite its higher teacher overlap and nDCG. The already-declared unit-normalized ridge resolves much of

this apparent conflict: KL falls to 1.1346 while its complete ordered top ten, nDCG, and Recall remain identical for every query.

Natural ArguAna endpoint	Full teacher KL	Teacher overlap	nDCG@10
Ridge, 400, raw	1.887867	.523969	.419567
Ridge, 400, unit-normalized	1.134610	.523969	.419567
Dedicated prefix	1.548575	.403580	.270668

Table S6. A prespecified scale control separates full-distribution calibration from the ranking comparison. All values use the same natural queries and ArguAna index.

All six raw/normalized pairs return exactly the same ordered top ten on both indices, although FP32 near ties change some lower-ranked positions. Normalization had increased KL in the earlier source width-256 example; here it improves KL. A fixed-temperature KL comparison is therefore sensitive to how predicted query norms transfer. These controls support reporting distribution agreement, ranking agreement, and judged relevance together. They do not support dismissing KL or attributing the transfer reversal solely to the distillation loss.

C.4 A common readout recovers most of the transfer gap

Prefix and readout	Width	Fresh source/SciFact overlap	Natural/ArguAna overlap	Natural nDCG@10
Frozen + ridge	256	.365430	.495306	.387051
Dedicated + ridge	256	.385417	.478758	.381929
Shared + ridge	256	.386263	.488928	.397124
Frozen + ridge	400	.391211	.523969	.419567
Dedicated + original KL head	400	.400456	.403580	.270668
Dedicated + ridge	400	.418815	.495187	.385763
Shared + original KL head	400	.402930	.416335	.278400
Shared + ridge	400	.421419	.509981	.404661

Table S7. A common readout separates accessible feature gains from the original head's transfer performance. Adapted rows average three fixed seeds; all rows use raw predictions. Fresh source uses 512 previously unused groups, while natural target relevance uses all 1,406 historical queries. Width 400 contains 385 active coordinates.

The original prefix comparison changes the feature map and its readout together. We now hold each of the six adapted prefixes fixed and fit the same ridge procedure used for the original frozen prefix: the 2,048-query fitting set, the same five training folds, penalty grid, widths, and solver. All twelve new heads are fixed before opening 512 previously unused source groups. Existing natural ArguAna queries provide an exploratory mechanism comparison. The original frozen-prefix ridge CV results and FP32 weights reproduce bitwise, and each old adapted head applied to the newly captured natural features reproduces its prior vector exactly.

The same adapted features support substantially better natural retrieval after head replacement. Dedicated nDCG increases by .11510 [.09885, .13185] and shared nDCG by .12626 [.10880, .14451].

With the common readout, their fresh source overlap also exceeds frozen-prefix ridge by .02760 and .03021 at width 400, with positive intervals; gains persist at width 256 and on the second document index. The large original transfer deficit therefore includes a recoverable readout component.

A smaller distribution-dependent difference remains. At width 400 the aligned dedicated and shared models trail frozen ridge in natural nDCG by .03380 [.02038, .04755] and .01491 [.00358, .02628], respectively; natural teacher overlap has the same ordering. At width 256 both nDCG contrasts span zero. Shared preservation improves on dedicated adaptation under the common readout at either width. These results identify a useful design target: preserve the source feature gain while retaining the stability of the readout across domains. They do not equate the remaining deficit with irreversible information loss.

The intervention jointly changes the target, numerical solver, regularization selection, and numerical features used for head fitting. Original heads were trained jointly in FP32 before PC deployment; the new ridge fits use the actual PC pooled features. The result establishes a readout dependence, with those components still coupled. No natural label enters a fit, and no new serving-time claim is attached to these twelve heads.

C.5 Fixed coordinate swaps separate source gains from target losses

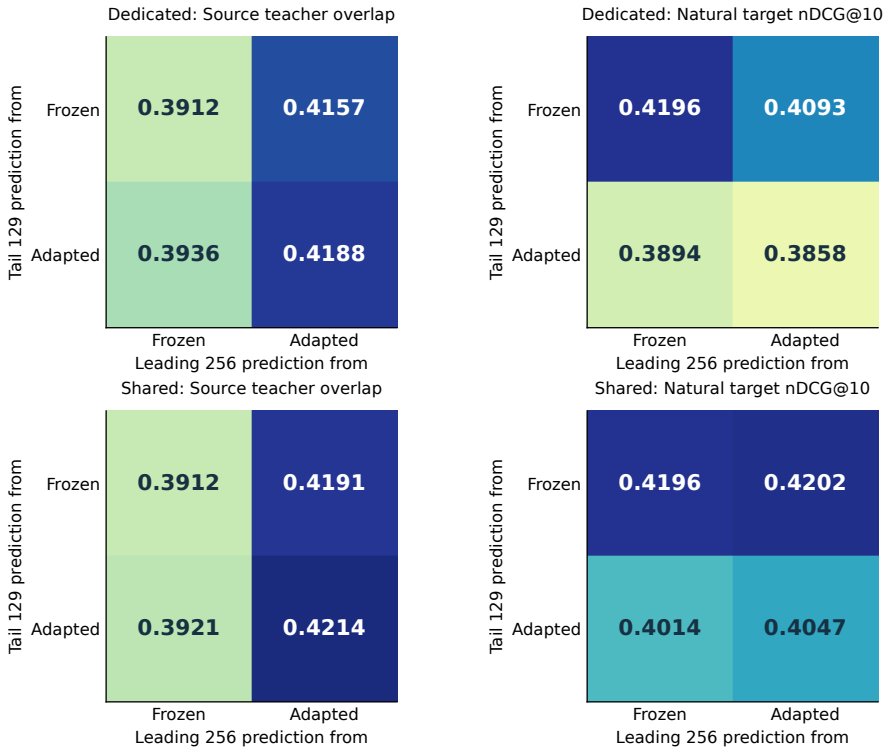


Figure 6: Fixed prediction-block swaps separate source gains from natural transfer losses. Each cell uses all leading and tail coordinates from the indicated prefixes; adapted cells average three seeds. Keeping the adapted leading block while restoring the frozen tail retains most source fidelity and largely recovers target relevance. These two-prefix interventions diagnose the predictions; their serving cost has not been established.

The width comparison supplies a direct intervention. Both widths read the same 384-dimensional pooled features. All seven prefix families select ridge penalty alpha=.1, and every width-256 raw prediction is bitwise equal to the leading 256 coordinates of its width-400 prediction. The larger head adds 129 active outputs and fifteen structural zeros. Denote the leading block by H and the tail by T. We evaluate all four combinations of frozen-prefix predictions F and adapted-prefix predictions A: (F_H,F_T), (A_H,F_T), (F_H,A_T), and (A_H,A_T). The cut at 256 is the previously fixed budget, not a newly selected boundary. This is a post-result intervention on the existing source and natural panels.

Keeping the adapted leading block and restoring the frozen tail raises dedicated/shared natural nDCG from .38576/.40466 to .40934/.42022. Their differences from the all-frozen .41957 are -.01023 $[-.02202, .00185]$ and +.00066 $[-.00976, .01126]$. These intervals do not establish equivalence, but show a substantial reduction in the observed deficit. The same swaps retain source/SciFact overlap gains of .02448 [.01647, .03262] and .02786 [.02038, .03561], compared with the original full-adaptation gains of .02760 and .03021. Replacing only the frozen tail by the adapted tail has the reverse effect: natural nDCG falls to .38941 and .40144, with both decreases resolved.

Averaging the two replacement orders provides an additive allocation of the full adapted-minus-frozen difference. The tail contribution to natural nDCG is -.02686 $[-.03350, -.02072]$ for dedicated and -.01685 $[-.02260, -.01153]$ for shared adaptation; the leading-block contributions span zero. This allocation is specific to the fixed coordinate intervention, and a nonzero interaction remains. The complete-ranking swaps, rather than a single score-error statistic, provide the evidence that the two blocks carry different parts of the source benefit and target deficit.

The tail's geometry explains why a low source weight need not imply a low transfer cost. The coordinate rotation is the descending eigensystem of the SciFact document uncentered second moment; its isotropic shrinkage leaves those eigenvectors unchanged. The tail accounts for 7.49% of source document energy and 8.28% of centered covariance trace, but 17.87% and 20.95%, respectively, on ArguAna documents. The source-query-averaged teacher Fisher assigns it 8.71% of its trace, although it occupies 33.5% of the active coordinates. By comparison, teacher query tail energy is stable across fitting, fresh source, and natural queries: 22.28%, 22.59%, and 21.56%. The observed shift is in how the document index weights the block, not an increase in the teacher queries' tail energy.

Prediction error changes at the same time. On source queries, tail squared error decreases from .11980 for frozen features to .11664/.11659 for dedicated/shared features. On natural queries it increases from .18093 to .22009/.19969. Natural tail cosine also falls, while predicted tail norms rise relative to the teacher. Direction and block-scale changes therefore both occur; the result does not identify information erasure. A small average tail energy cannot protect a near-tied document pair: its ranking depends on the full margin plus the leading and tail errors dotted with that pair's document difference.

There is no exact missing source direction within this full support. The source centered document Gram has rank 385. In exact arithmetic, zero full-index KL requires the student-teacher score difference to be constant across documents, which here implies equal query coordinates. Coordinate squared loss has the same per-query optimum. Their local weights differ: the KL Hessian at the teacher is $\mathrm{Cov}_{p_T}(d)/\tau^2$, whereas half squared coordinate loss has identity Hessian. The source-query-averaged Fisher has condition number approximately 934. Finite fitting support, finite-capacity approximation, regularization, and optimization can thus change which errors remain even when the ideal optimum agrees.

Finally, a common Euclidean ridge procedure is not invariant to every invertible change of feature coordinates. If centered features obey $X_A = X_F M$, then the penalty on an equivalent frozen-coordinate coefficient B becomes $\lambda\|M^{-1}B\|_F^2$. Our trace-scaled penalty removes scalar scale and orthogonal-rotation dependence, but not general anisotropy. The residual can therefore include finite-sample readout geometry as well as a change in the features. The coordinate swaps establish a useful design target: retain the source-leading-block gain while stabilizing the transfer-sensitive tail within a single cheap encoder.

C.6 Uniform coordinate supervision does not repair the common-readout residual

The block intervention suggests a training prediction: giving every teacher coordinate equal weight might retain the adapted leading-block gain while protecting the transfer-sensitive tail. We test that prediction with nine new fits, replacing full-index KL by half squared teacher-coordinate distance. The fitting queries, initial models, three seeds, 500 updates, optimizer settings and capacity placements remain fixed. Shared training replaces both shallow and continued-depth terms. Each new adapted prefix then receives the same R-only ridge procedure at widths 256 and 400, with all twelve fits selecting alpha=.1. A further 1,024 previously unencoded source queries are opened only after all training and ridge fitting finish. Natural ArguAna remains an exposed mechanism population.

Prefix and objective	New-source overlap, ridge400	Natural nDCG, trained head	Natural nDCG, ridge400
Frozen ridge reference	.387109	–	.419567
Dedicated / KL	.419043	.270668	.385763
Dedicated / coordinate	.417122	.291306	.388577
Shared / KL	.414518	.278400	.404661
Shared / coordinate	.416374	.297440	.394442

Table S8. The objective intervention has different effects before and after a common readout. Source overlap uses the SciFact index and a new 1,024-query population; natural nDCG uses the complete ArguAna task. Adapted rows average all three fixed seeds. All shallow heads have width 400 with 385 active coordinates. These new models are not included in the earlier serving-time experiments.

Coordinate supervision improves the trained heads by .02064 $[.01377, .02773]$ and .01904 $[.01227, .02643]$ nDCG for dedicated and shared prefixes. Even the frozen-prefix affine-only head improves, from .25002 to .26016. After the common ridge fit, however, the dedicated objective contrast is only .00281 $[-.00394, .00949]$, and the shared contrast reverses to -.01022 $[-.01727, -.00366]$. Both remain below frozen-prefix ridge400. Thus a better directly trained head need not imply more transferable features for a common deployed consumer.

The source benefit persists: both new ridge400 families exceed the frozen reference on the new source queries by approximately .030 overlap, with positive intervals, and retain a similar advantage when those queries score the ArguAna index. Their changes relative to KL on source overlap are small and uncertain. Width 256 gives natural nDCG .38832/.39936 for dedicated/shared coordinate training; no fixed endpoint restores the frozen400 result. Across all 26 raw/unit ridge pairs, normalization changes no ordered top ten in any of the four query/index cells.

The coordinate diagnostic narrows the explanation. On new source queries, the dedicated/shared ridge tail error decreases from .11561/.11549 under KL to .11359/.11361 under coordinate training. On natural queries, shared tail error instead increases from .19969 to .21141, while its tail cosine decreases from .41818 to .40273. Dedicated natural tail error improves slightly, but leading-block error grows. The earlier swaps identify consequential prediction errors; they do not identify a loss replacement that generalizes beyond the fitting population. Continued depth-six relevance remains near .578-.582, with coordinate-minus-KL intervals spanning zero, so the large shallow deficit is not shared by every consumer of the adapted token states.

This intervention changes a complete training procedure. Equal updates do not match gradient scales, clipping or convergence, and FP32 training followed by INT8 prefix deployment remains common to both families. The experiment therefore rejects the simple repair prediction in this setting without identifying one unique optimization cause. It strengthens the case for evaluating representation changes through a common consumer and through the input distributions they must serve.

C.7 Source-fitted coordinate maps also face distributional extrapolation

A single invertible feature-coordinate change would preserve an appropriate linear consumer, while changing an isotropic ridge estimator. We therefore fit paired affine least-squares maps in both directions between frozen and each of the six old KL-adapted pooled representations, using only R2048. The relative singular-value cutoff is fixed at 10^{-6}; every source matrix and every fitted map remains full rank. Representation invariance alone is not a task-performance measure [11], so this diagnostic reconstructs features on other populations rather than interpreting a similarity score as retained information.

On the earlier independent source-512 panel, residual squared error is 5.2-6.3% of the predicted feature's energy around its R mean. On natural queries it rises to 20.4%/24.1% for frozen-to-adapted shared/dedicated maps, and 28.9%/40.9% in the reverse direction. The source-fitted maps are worse there than simply using the other representation unchanged: their errors are 3.5-7.3 times the actual adaptation-displacement energy. Approximately 86-88% of natural residual energy is a systematic residual-mean component, compared with about .2% on source evaluation. The estimated maps extrapolate poorly on natural queries; alternative approximate maps and nonlinear recovery remain open.

The input distributions expose an important boundary. R and source evaluation contain 12-50 and 12-38 effective tokens, respectively; natural queries contain 62-512, with median 227. Every natural query's source-covariance leverage exceeds its corresponding R maximum. The source matrices are full rank, so these queries are not outside their linear span; they are outside the observed covariance range. Within the natural population, length correlates strongly with residual magnitude and leverage, but length, content and task are jointly shifted. The data cannot identify length as the cause. They motivate matched-input interventions before another target-domain fit, and extend the coverage question from how many distinct queries are fitted to which input regimes those queries represent.

C.8 Input composition separates token states from pooled transport

The source and natural length ranges do not overlap, so their ordinary correlation cannot distinguish input form from task content. We intervene on the exact token sequences of 16 fixed source queries and 16 natural queries preselected to have at least 256 effective tokens. Source inputs contain 15–20

tokens; the natural subset contains 284–417. We cycle each source query's content to lengths 64, 128, 256 and 512, and truncate each natural query's content to lengths 32, 64, 128 and 256, retaining the template and boundary tokens. These exposed-query tests use all six old adapted prefixes and the original source-fitted maps. Repetition and truncation change content statistics as well as length; relevance is not evaluated on these artificial inputs.

Both directions move consistently with the input intervention. In deployed PC execution, dedicated-prefix forward relative affine error rises from .0566 on source originals to .1869 at repeat512, while natural error falls from .2820 to .0499 at truncate32. The shared-prefix values move from .0474 to .1512 and from .2405 to .0408. The pattern holds in FP32, whereas masked right-padding leaves FP32 pooled features within 5.67×10^{-7} maximum absolute difference. Figure 7 reports the full fixed trajectories.

The template's pooling mass changes sharply with these inputs. For the frozen prefix, its eight tokens carry 26.64% of the weight on short source inputs and .83% after repetition. A fixed-hidden-state 2×2 intervention exchanges this mass between paired inputs while preserving the relative weights within template and complement groups. On repeated source inputs, restoring the short-input mass reduces dedicated forward squared residual from 4.247 to 2.104, but its relative error remains .1869 versus .1872. Natural long inputs likewise improve in absolute residual, from 2.469 to 1.720, while the relative error remains .2820 versus .2843. Pool mass contributes to the error magnitude; the altered token states and their interaction with pooling remain consequential.

A second factorial separates absolute positions, self-attention token multiplicity and pooling multiplicity. Every copied body token is assigned its original position class, with count m_j. We independently cycle position IDs back to their originals, subtract $\log m_j$ from each layer's key logits, and subtract the same term from pooling logits. The softmax multiplicity identity is established, including its proportional-attention use in ToMe [12]; here it supplies an exact control. With position and attention corrections, every repeated FP32 hidden state matches its original counterpart within 1.05×10^{-5}. Yet uncorrected pooling leaves dedicated forward relative error at .1137, compared with .0437 on the short original. Correcting the pool restores .0437, as the identity requires. Position correction alone reaches .1258 from the native repeated value .1864, and attention correction alone reaches .1807. Appendix I gives the complete factorial.

This construction changes pooled transport while preserving the token-state set. A fixed attention-pooling vector and its nonlinear residual are part of the representation comparison, not neutral observers of it. Source coverage therefore concerns the input statistics presented to the consumer as well as the number of distinct queries. The artificial-sequence recovery establishes a mechanism of pooled-map failure. We next test whether source-determined pooling changes improve natural retrieval.

C.9 Improving feature transport can still harm retrieval

We compare three pooling consumers on the unchanged original and six KL-adapted prefixes. The native consumer retains its learned attention weights. Content-only pooling renormalizes those weights over body tokens. Source-fixed mass retains both the contextualized format-token mean and body-token mean, but combines them using the format mass averaged over R2048. Here the format group includes CLS, the eight prompt tokens and terminal SEP. All three consumers apply the original nonlinear residual after pooling. Each receives the same source-only ridge selection and fitting procedure at widths 256 and 400; cross-validation computes fixed mass within its training fold. The experiment reuses the earlier source512 and natural1406 populations.

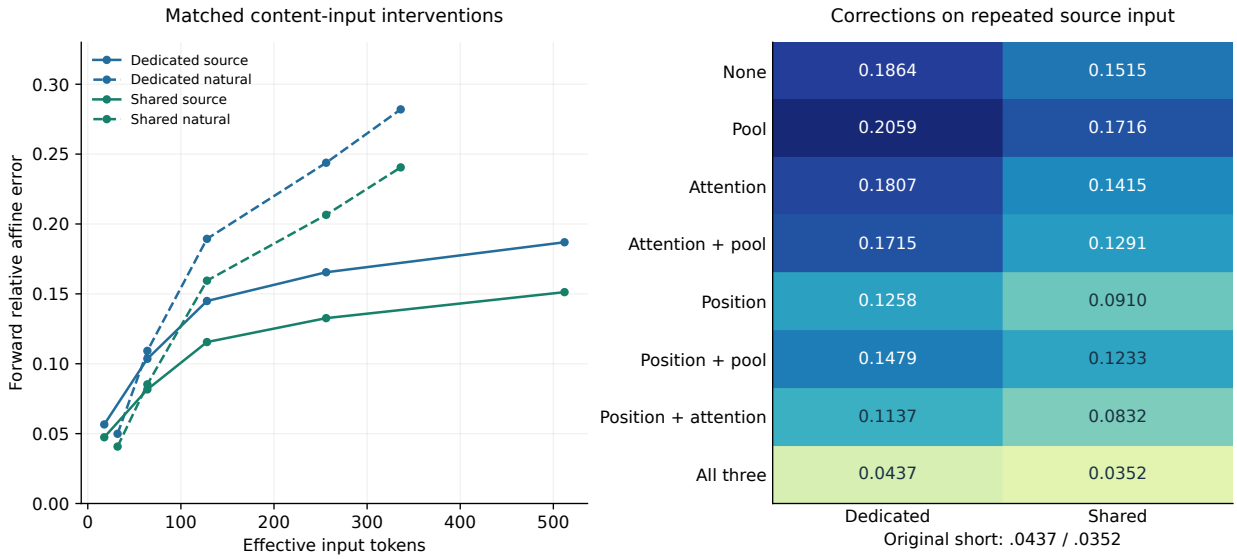


Figure 7: Input form changes source-fitted pooled transport. Left: fixed source-content repetition and natural-content truncation trajectories, PC execution. Right: the full position, self-attention and pool-multiplicity factorial on the 16 source queries, FP32 execution; numbers are three-seed mean forward relative affine errors. Original short-input references are .0437 (dedicated) and .0352 (shared). All-corrected recovery is a mathematical control.

Width	Pool consumer	Source overlap: frozen	Dedicated	Shared	Natural nDCG: frozen	Dedicated	Shared
256	Native	.3654	.3854	.3863	.3871	.3819	.3971
256	Content only	.3582	.3786	.3781	.3246	.3314	.3314
256	Source-fixed mass	.3611	.3811	.3820	.2995	.2369	.2605
400	Native	.3912	.4188	.4214	.4196	.3858	.4047
400	Content only	.3838	.4070	.4056	.3521	.3607	.3606
400	Source-fixed mass	.3871	.4109	.4108	.3402	.2657	.2830

Table S9. Source-determined pooling consumers with a common ridge budget. Adapted columns average the three fixed seeds. Source overlap uses E512 against SciFact; natural nDCG uses all 1,406 ArguAna queries. These populations were exposed before this mechanism experiment.

Every alternative loses natural relevance against native pooling: all twelve consumer-minus-native contrasts have pointwise 95% intervals below zero. At width 400, content-only pooling lowers dedicated and shared nDCG by .02508 and .04403. Source-fixed mass causes larger losses, .12008 [.10255, .13762] and .12163 [.10466, .13877]. Content-only adapted prefixes slightly exceed their equally modified frozen control, but all three endpoints have deteriorated. The smaller adaptation

gap therefore comes at the cost of lower absolute retrieval quality. Source teacher overlap also declines, although the adapted prefixes retain their source advantage within each consumer.

The geometric diagnostic moves differently. Source-fixed mass reduces natural frozen-to-adapted affine squared residuals from 1.928/1.528 to .863/.690 for dedicated/shared prefixes. This change also compresses feature energy: errors relative to energy around the R mean improve much less, from .2408/.2035 to .2311/.1786. Nevertheless, even the lower relative errors accompany worse teacher ranking, raw and unit-scale KL, and judged relevance. The original fourteen native ridge fits and all native predictions reproduce exactly. One new width-256 content-only model has an FP32 near-tie exchange at ranks eight and nine after normalization; its top-ten set and every raw/unit relevance value remain unchanged.

These results give a concrete limit to geometric correction. The template and body statistics influence the source-fitted feature relation, but making that relation easier to reconstruct does not identify the document directions or ranking margins that matter to retrieval. The native consumer remains the stronger endpoint across this fixed comparison.

C.10 New query groups narrow the common-readout residual

The later 973-query validation changes the size of the transfer gap without changing any model. Dedicated and shared shallow native-prefix means rise to .36667 and .36193, while frozen-prefix ridge400 reaches .42730. The same source-fitted common readouts raise the adapted-prefix means to .42193 and .42506. Their differences from frozen ridge are now -.00537 and -.00224, substantially smaller than the old official-population differences of -.03380 and -.01491. The new paired intervals are [-.01646,.00645] and [-.01355,.00913]. Relative to the canonical-self old task, the dedicated gap shrinks by .03015 [.01257,.04825], while the shared shrinkage is .01402 [-.00193,.03042]. These are post-result comparisons between two populations, with independent group draws across populations and paired endpoints within each. Continued adapted-prefix means are .60524/.60016, against .59537 for original PC6.

The benefit of replacing the trained shallow head persists, while the remaining adapted-feature deficit is much less pronounced. Thus the earlier tail-coordinate intervention identifies consequential errors in its evaluated population, not an inevitable loss from prefix adaptation. The canonical-self change can alter old aggregate nDCG by at most .02835 for any endpoint: excluding one nonrelevant item moves the positive at most one rank, and only 108 of 1,406 old queries are affected. That bound is already too small to explain most of the native-prefix mean increase. The exact old-vector sensitivity raises old frozen ridge by .00526 and the native dedicated/shared means by .00224/.00240. Their common-readout gaps become -.03552/-.01626, slightly larger than before; this rule change does not explain the new population's smaller residual. These results sharpen the coverage question: useful feature changes, their readout and their conditional retrieval consequences must be evaluated over the query populations that will use them.

Appendix D. Shallow correction and its transfer limits

D.1 Query-conditioned correction beyond pooled affine recalibration

The pooling interventions show that stable feature transport and useful retrieval corrections are different objectives. We next ask whether the geometry of a shallow correction improves retrieval after allowing a strong readout of the existing pooled representation. All comparisons retain the frozen quantized two-layer encoder, original attention and residual pooling MLP, source-fitted

coordinate-ridge head, and fixed document vectors. The new coefficients predict teacher-score residuals around this existing endpoint.

Let h_t denote token states, $\alpha_t = \mathrm{softmax}_t(u^\top h_t)$ their original attention weights, and $p = \sum_t \alpha_t h_t$. The existing pool is $z = g(p) = p + \mathrm{MLP}(p)$. Its attention derivative is the weighted covariance $C_q = \sum_t \alpha_t (h_t - p)(h_t - p)^\top$. With $J_q = \partial g / \partial p$ and readout matrix W, a displacement a of the attention parameter gives the first-order query correction

$$\delta v_q = W J_q C_q a.$$

The 384 coefficients in a are shared across queries, while the covariance and Jacobian make the correction query-dependent. This is a fixed tangent model evaluated at the original parameters. Pretrained attention linearization and regularized tangent fitting are established by TAFT [19]. Here the small pooling subspace permits an analytic query-side correction and preserves single-vector retrieval over cached documents.

We fit the correction on the original 2,048 source queries. Each query contributes its original cheap top-K documents, with $K = 100$ and no relevance-based inclusion. The original query norm scales baseline scores and correction features. Teacher-minus-baseline residuals and features are centered within this fixed set. For residual errors e_j, $\sum_j (e_j - \bar{e})^2 = K^{-1} \sum_{j<k} (e_j - e_k)^2$. This connects centered regression to established Margin-MSE distillation [20]. The baseline coordinate-ridge prediction remains the zero-correction prior, so local residual fitting complements the earlier complete-coordinate supervision.

The pooled-only comparator fits $\delta v_q = M[z_q; 1]$ across all 385 active output coordinates, giving 148,225 residual coefficients. It uses the same scores, candidates, norms and five source folds, penalizing all residual coefficients including the intercept. Its broad penalty grid includes the exact zero correction and selects $\alpha = 100$ internally, away from either boundary. The final float64 matrix-free solve reaches a relative normal-equation residual of 3.73×10^{-9}. Both studies validate residual coefficients conditional on a base head already fitted on all 2,048 queries. Their fold losses are therefore conditional correction-selection scores, rather than fully out-of-fold estimates of the entire training procedure. The affine correction is fused into the original deployment weight and bias.

On the existing source E512 panel, teacher top-ten overlap rises from .38145 to .39785 with the shared tangent and to .38359 with the affine correction. The corresponding centered score MSEs are .001253, .001183 and .001242. Judged retrieval separates the corrections more sharply (Table S10). On historical ArguAna973, shared tangent reaches .44168 nDCG@10, compared with .42730 for the baseline and .42726 for pooled affine. Its advantage over affine is .01442 [.00793,.02107], using 134-group paired bootstrap resampling. On historical SciFact300, the source-index relevance check, tangent reaches .58725 versus .57629 for affine, a difference of .01095 [.00135,.02142] under paired-query resampling. These populations had already been exposed during development. The intervals condition on the fitted models and the stated resampling units.

Table S10. Fixed shallow corrections under a common teacher-score residual target. E512 measures teacher top-ten overlap; ArguAna973 and SciFact300 measure nDCG@10 on historically exposed populations. SciFact is a source-index relevance check. Request time covers all 973 ArguAna queries in one matched four-endpoint session. Correction coefficients exclude the unchanged encoder, pool and base head.

Endpoint	Correction coefficients	E512 overlap	ArguAna nDCG	SciFact nDCG	Request ms
Original PC2 ridge	0	.38145	.42730	.57378	15.634
Pooled-affine residual	148,225	.38359	.42726	.57629	15.666
Shared token tangent	384	.39785	.44168	.58725	16.007
Original PC6	N/A	1.00000	.59537	.70364	40.377

The tangent correction yields a retrieval advantage under the common residual target. The two families are nonnested: pooled affine can modify every active output direction, whereas the tangent remains in col(W) but depends on token covariance and a query-dependent Jacobian. Parameter count consequently does not order their effective capacity. This advantage is relative to the source-selected pooled-affine family; other nonlinear functions of the pooled features may recover the useful signal.

The execution structure also matters. Matched complete requests over ArguAna973 take 15.634 ms for the baseline, 15.666 ms for fused affine, and 16.007 ms for shared tangent. The tangent increment is .374 ms, or 2.39%, in this shared-machine session, with all 3,892 timed returns checked. A separate six-endpoint session evaluates a 384-coefficient candidate-conditioned quadratic correction: it reaches .42896 ArguAna nDCG and costs 90.768 ms when applied across the corpus, or 16.855 ms within the cheap top 100. Candidate-dependent attention has precedent in Poly-encoders [21]. This quadratic restriction lacks the full Poly-encoder function class and yields no advantage here. Shared global scoring and top-100 correction return identical top tens on all 973 queries. The observed gain therefore concerns nearby ranking corrections, while the single-vector implementation avoids candidate-wise computation across the index.

The tangent can also be realized approximately inside the original graph. If u is the scalar-attention parameter, replace it by $u + a$ with the fitted displacement and a fixed unit step. This recomputes the original softmax pool and nonlinear readout rather than explicitly adding their first-order response. No additional trainable parameters or serving operators are introduced. On historical ArguAna973, this finite update reaches .43894 nDCG, retaining 80.9% of the tangent's improvement over .42730. Its difference from tangent is -.00275 [-.00507,-.00061]. A matched three-endpoint session measures 15.513 ms for the finite update, 15.502 ms for the original endpoint and 15.919 ms for tangent; finite minus original is .0109 ms with paired-group interval [-.0854,.1063] ms. All 2,919 timed outputs agree with saved predictions. The finite update has slightly better source E512 teacher overlap (.39902 versus .39785), so source fidelity does not order the two natural-quality outcomes. This realization was examined after the twelve-method NFCorpus protocol was fixed and is not added to that prospective comparison.

D.2 Correction transfer under graded relevance

The fixed NFCorpus task attenuates the historical correction advantage (Table S11 and Figure 8). Tangent improves nDCG from .31377 to .31678, a difference of .00301 [-.00088,.00681]. Affine reaches .31529 and the constant control .31594; tangent minus affine and tangent minus constant

both have intervals spanning zero. The original width-256 ridge attains .31616 at slightly lower request cost than the width-400 baseline. The prospective data therefore support a small positive tangent point estimate, with unresolved superiority over these cheap alternatives.

Table S11. Every prospectively retained NFCorpus endpoint. All 323 official test queries and 3,633 documents are used. Original relevance grades enter nDCG; recall and MRR use positive judgments. Times are matched complete singleton requests in one shared-host execution. The adapted prefix is the preselected existing dedicated seed20261019 model.

Endpoint	nDCG@10	Recall@10	MRR@10	Request ms
Original PC2 ridge400	.31377	.15354	.50115	2.088
Constant correction	.31594	.15415	.50487	2.094
Shared token tangent	.31678	.15346	.50279	2.271
Pooled-affine residual	.31529	.15444	.50293	2.114
Original PC2 ridge256	.31616	.15477	.50886	2.060
Native PC4	.08707	.04559	.17631	3.539
Ridge PC4	.18105	.08158	.34411	3.648
Native PC6	.35651	.17393	.55488	5.540
Adapted PC2, original head	.30669	.14312	.51301	2.187
Adapted PC2, common ridge	.31484	.15149	.50851	2.071
Adapted native continuation	.35961	.17229	.55306	5.054
Fixed paid std-q60 gate	.32535	.15624	.52367	3.345

A diagnostic fixed before the target outputs asks whether the correction still follows the local teacher residual. On each original cheap top100, let r be the centered native-teacher minus original-cheap score vector, and c the centered correction, with every shallow head divided by the same original query norm. The reduction in score MSE is exactly $(2r^\top c - \|c\|^2)/100$. Tangent reduces this error from .00148525 to .00142879, or 3.80%, compared with .68% for the constant correction and .73% for affine. Median directional cosine is .176; 86 queries have negative alignment and 32 more have positive alignment but a correction too large to reduce this local squared error.

Local fidelity improvement does not ensure relevance improvement. Of the 205 queries whose tangent score MSE improves, 46 gain nDCG and 32 lose it. Of the 118 whose MSE worsens, 22 gain nDCG and 17 lose it; the others tie. These fixed-output results distinguish partial transfer of the teacher-margin geometry from the smaller judged-relevance benefit. All corrections and selection rules remain unchanged in this fixed-output analysis.

Appendix E. Offline support and matched online computation

E.1 Offline teacher information moves the first useful budget

This appendix details the information available before a request begins and its relation to online continuation. The key distinction is between keeping the alignment sample fixed and keeping the total teacher information fixed: vocabulary extraction can add useful responses that no alignment query contains. A source-identical completion control below makes that contribution directly testable.

A continuation policy depends on what it can return before buying further computation. We compare two fixed 384-dimensional lookup means: the original pretrained token embeddings and a table extracted from the six-layer teacher. Both retain the complete prompted, truncated query

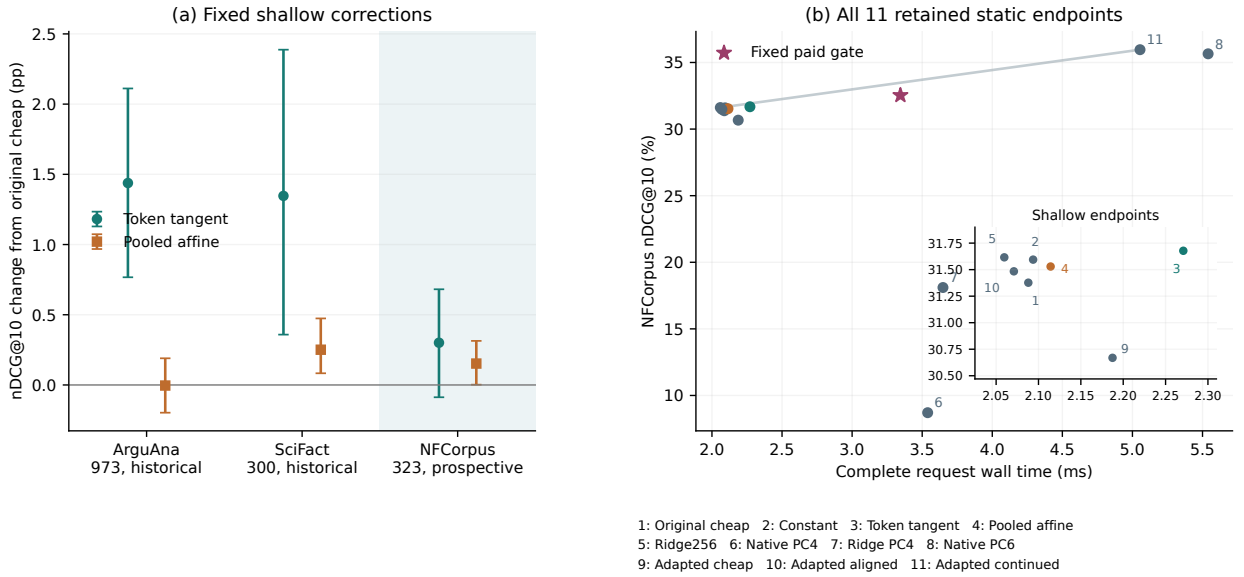


Figure 8: Fixed correction gains shrink on the prospective graded task. Left: paired nDCG differences from the original cheap endpoint, with historical ArguAna group intervals and SciFact/NFCorpus query intervals. Right: all eleven retained static endpoints and the unchanged paid gate on NFCorpus, at matched complete-request costs. The line is the target-descriptive static-mixture envelope; the inset resolves the shallow endpoints.

and all special tokens. The same 2,048 source queries, teacher coordinates, five folds and eight ridge penalties determine their readouts. Both select penalty .1. Their online function class remains a mean lookup followed by an affine map.

The teacher table follows the established static-distillation idea [24,26]. For each of the 30,522 vocabulary IDs, we form the original retrieval prompt followed by that literal ID, with the original special tokens. Each sequence contains eleven active positions. We execute the original six-layer INT8 query encoder at batch one and store its mean final hidden state before the Dense projection. Online queries average those cached rows. The table adds 30,522 isolated-token teacher contexts to the offline information budget; no natural-query or document encoding is added. It occupies 46.88 MB. Constructing the 30,506 rows beyond the sixteen reused qualification rows takes 98.43 CPU seconds, and source fitting plus evaluation takes 2.33 seconds. These clocks begin after imports. The source-held-out teacher overlap rises from .30625 for the input table to .43340 for the teacher table, above the original PC2 ridge's .38145.

Table S12. Complete-population static lookup, two-layer and six-layer endpoints. All four actions are executed in the same task-specific request session. Costs include tokenization, query computation, full-index scoring, self exclusion and ordered retrieval. Model and index loading precede request clocks.

Task / endpoint	nDCG@10	Request ms
ArguAna973: teacher-token mean	.50823	1.958
ArguAna973: raw token mean	.38818	1.998
ArguAna973: original PC2 ridge	.42730	15.565
ArguAna973: original PC6	.59537	39.816
NFCorpus323: teacher-token mean	.30405	.557
NFCorpus323: raw token mean	.26756	.559
NFCorpus323: original PC2 ridge	.31377	2.119
NFCorpus323: original PC6	.35651	5.226

The teacher table improves ArguAna nDCG over the raw table by .12005 [.09932,.14002] and over PC2 by .08093 [.05719,.10336]. Its request time is about one eighth of PC2's. NFCorpus also improves over the raw table, by .03650 [.02299,.05060]; its contrast with PC2 is -.00972 [-.02413,.00453]. This interval does not establish equivalence. Both comparisons retain identical feature width, source fitting queries and document vectors, while the teacher table purchases additional offline teacher information.

At the mean PC2 cost, random selection independent of query content can afford 35.94% PC6 calls on ArguAna when paired with the teacher table. The mixture reaches .53955 nDCG, exceeding PC2 by .11225 [.09084,.13337]. On NFCorpus, 33.45% deep calls yield .32160, a .00783 [-.00425,.02012] contrast. The raw-table mixtures measured in these same sessions instead give +.03521 [.01306,.05629] and -.01648 [-.02839,-.00497]. Thus the earlier finding that NFCorpus favors the middle endpoint over a zero/six-layer mixture depends on the weak raw table. A stronger offline representation moves that comparison to an unresolved difference (Figure 2). The actions also have distinct readouts, so this is a comparison of deployed retrieval functions rather than an isolated effect of depth.

A fixed affine projection probes whether the new table is an affine reparameterization of the original input rows. We fit the whole vocabulary once, with equal row weights and an intercept, as $H = EA + b + R$. Averaging token rows preserves this decomposition exactly. The residual contains 44.14% of centered whole-table feature energy, but this percentage is not a share of the retrieval gain. An affine-only table, refitted with the same source ridge protocol before reading natural results, gives .37730 ArguAna and .27572 NFCorpus nDCG. The full table exceeds it by .13093 [.11169,.14932] and .02833 [.01489,.04191]. The selected penalties remain .1. The fitted map has a condition number of 2.82×10^8. Its nearly null output direction aligns with the final LayerNorm's fixed affine constraint: the corresponding unit vector has absolute cosine above .9999999999 with the normalized inverse LayerNorm scale. Formal full rank here arises from floating-point perturbations. The comparison therefore tests an affine replacement with different regularization geometry, including this architectural near-degeneracy. This fixed affine substitute fails to recover the teacher table's gain. The residual-energy fraction alone cannot explain the size of that gain.

Earlier controls clarify the scope of the input-table result. Averaging BERT's embedding-module output, including positions, types and LayerNorm, has almost the same source teacher overlap as averaging raw token rows, .30605 versus .30625, but only .16065 ArguAna nDCG versus .38818. One

fixed intervention clamps position IDs to the source range without refitting anything. It reduces ArguAna nDCG by another .10009 [-.11825,-.08359]; NFCorpus vectors remain byte-identical because those inputs already lie within range. Restricting position support alone therefore does not repair that transfer. It also changes position frequencies, leaving other positional and normalization effects unresolved.

The stopping endpoint changes the meaning of a fixed gate as well. For the original raw table, replacing PC2 on stopped queries changes allocation by $\mathrm{Cov}(g, Q_2 - Q_0)$. This term is -.01179 [-.01748,-.00651] on NFCorpus, reducing zero/six-layer allocation to -.01066 [-.01756,-.00388]. The first two layers improve stopped queries by .06182 on average, compared with -.00197 among continued queries. ArguAna retains positive allocation, .02354 [.01099,.03551]. Absolute policy quality nevertheless falls by .02876 and .04670 when stopped PC2 outputs are replaced by raw-table outputs. These identities use the original raw endpoint; the original gate still requires PC2 observations. They do not supply a free routing signal for either lookup table.

The teacher table uses a restricted extraction and fitting recipe compared with Model2Vec and LightRetriever. Both natural populations are historically exposed at this stage. The single-thread matched execution verifies all 5,184 timed and eight warm complete outputs against frozen fingerprints. Its two post-import processes use 66.01 CPU seconds including setup. The measurements retain shared-host conditions. Four thousand paired bootstrap draws resample the original 134 ArguAna groups or the 323 NFCorpus queries, jointly varying costs and quality and reselecting the random-mixture proportion. These are average-budget comparisons within one hardware session. The experiment evaluates the four static actions rather than a new adaptive policy or an expanded 125-endpoint bank.

Classical source-IDF pooling does not improve this endpoint. Each source fold recomputes its frequencies; both the IDF head and an equal-content-weight control with identical per-query format mass select ridge penalty .1. On ArguAna, uniform, IDF and mass-control pooling reach .50823, .44849 and .43810 nDCG. IDF minus uniform is -.05974 [-.07406,-.04541], while IDF minus the mass control is +.01039 [.00308,.01787]. On NFCorpus, the same three values are .30405, .30586 and .30236; IDF minus uniform is unresolved. Independently fitted heads prevent an additive causal decomposition. Fixed lexical weights can be compiled into the table, but no new serving-time advantage is claimed.

A separate intervention identifies information that the alignment queries never observe. Their effective token lists cover only 4,555 of 30,522 vocabulary IDs. We retain those teacher rows exactly and complete every other row either by source-only raw-to-teacher OLS or by the mean of its eight nearest source-observed raw embeddings. Every original R2048 feature, CV result and readout stays unchanged. Yet the full table exceeds source-affine completion by .04990 [.03782,.06276] ArguAna nDCG and .01203 [.00174,.02171] NFCorpus nDCG; the nonlinear neighbor control is weaker. Actual source-unseen token occurrences are 26.36% and 8.59%, rather than the vocabulary's 85.08% unseen fraction. All 77 NFCorpus queries containing no such token return identical outputs under every table. Relative to these fixed completion priors, offline lexical teacher responses contribute retrieval value beyond the query-alignment sample. The broader source-selected completion search subsequently retains this information increment, as reported in the main text. Sections E.2–E.3 retain the complete comparison and its relation to the full-query increment.

E.2 Matched constructions retain a positive online deep increment

A strong token table sets the baseline against which to measure the value of joint query processing. We compare both input constructions at both depths under both readouts, retaining every cell. This separates a useful offline-depth increment from a positive matched online-deep increment, while revealing that the shallow comparison depends on the task.

The teacher-table result leaves open which computation supplies its quality. We first fix two additional representations before fitting or target evaluation: the same singleton-token extraction stopped at layer two, and a uniform mean of the original full-query layer-two states. Both use the same R2048 queries, teacher coordinates, source folds and eight ridge penalties as the existing table; both select penalty .1. We then complete a fixed eight-cell comparison by adding the missing full-query-six mean with one source-only ridge fit, selected at penalty .01, and applying the original Dense readout to all four representations. Table 3 includes every cell.

Increasing offline extraction depth from two to six raises ArguAna nDCG by .12414 [.10439,.14390] and NFCorpus by .03398 [.02181,.04661]. The two-layer cache has unresolved differences of -.00409 and +.00252 from the raw table, whose nDCG is .38818 and .26756. The stronger zero-layer online endpoint therefore contains useful computation performed deeper in the teacher offline. This locates the offline-depth benefit within established contextual-to-static extraction [27,28].

Full-query computation still adds quality when depth and readout are matched. At layer six, full query minus singleton table under source ridge is +.07660 [.05149,.10096] on ArguAna and +.04840 [.03353,.06407] on NFCorpus. The full-query-six ridge retains .85664 teacher overlap and reaches .58483 and .35245 relevance, versus .59537 and .35651 for the native endpoint. The latter differences are descriptive; they do not establish equivalence. The common-readout comparison therefore identifies an additional quality benefit from processing the full query at depth six.

The shallow result is different. Full-query PC2 mean minus singleton-PC2 mean is -.10901 [-.13676,-.08132] on ArguAna but +.01894 [.00727,.03061] on NFCorpus. The original PC2 attention pool, outside this uniform-mean factorial, reaches .38145 source overlap and .42730/.31377 nDCG. Replacing it with the mean of the same hidden states lowers nDCG by .15221 and .02476. All 1,296 original attention-PC2 complete returns reproduce exactly from those shared states. Moreover, full-query PC2 mean has better source teacher overlap than the shallow table (.35820 versus .32070), despite its worse ArguAna relevance. Neither a source-fidelity improvement nor a shallow-context effect alone predicts cross-domain action value.

Separate endpoints can nevertheless hide a useful complement. One source-selected block-ridge head combines the deep table mean with the original PC2 attention pool, reaching .52294 ArguAna and .32532 NFCorpus nDCG, above table-only by .01471 [.00127,.02888] and .02126 [.01260,.03039]. We freeze this joint consumer and replace its PC2 input by a source-fitted affine prediction from the table. Compiling that substitution back into one static head gives only .50254 and .30318. Actual PC2 therefore supplies a useful residual relative to this particular source predictor, without establishing that a nonlinear static model cannot recover it. Complete requests then separate complementarity from affordability. Fusion costs 16.021 ms on ArguAna and 2.069 ms on NFCorpus. Against five matched static controls, including the compiled head and the stronger adapted-six endpoint, its same-cost differences are -.02204 [-.03340,-.01035] and +.00008 [-.00703,.00676]. All 7,776 timed returns reproduce the fixed predictions. The shallow feature has task value conditional on the cache, but these measurements establish no new superior budget point.

A fixed-consumer prior cross-check further shows that ArguAna's live increment falls from .04157 to .02040 when source-unseen teacher rows are supplied, with the decline reproduced by holding the old correction fixed while moving its static baseline (-.02298, simultaneous six-contrast 95% interval [-.03809,-.00787]); reduced marginal ranking value need not mean comparable online information has disappeared. Appendix E gives the fitting and substitution details.

Readout compatibility accounts for much of the native head's failure on singleton means. Classical mean removal [35], using only the 2,048 source inputs and no new teacher-coordinate labels, raises the six-layer table from .00514 to .37628 ArguAna nDCG and from .02582 to .29825 on NFCorpus. Its weight matrix is unchanged. Subtracting a common query-vector mean is equivalent to adding a fixed document-score bias, so this recovery acquires no new query distinctions. Full ridge still adds .13195 [.11251,.15247] on ArguAna; its .00580 [-.00378,.01568] NFCorpus increment is unresolved. A source-fitted shared gain and intercept improves source teacher overlap more than mean removal, yet reaches only .28693 NFCorpus nDCG. The native-by-input interactions in Table 3 therefore combine useful supervision with consumer geometry and scale. The within-ridge comparisons remain the appropriate controlled comparison of the two input constructions.

The singleton/full-query construction also changes positions and prompt multiplicity together with within-query interaction. Each ridge uses its own source-selected regularization. The evidence identifies this complete construction's dependence on depth and readout, without naming a single causal context variable. The initial depth/pooling controls use 71.39 CPU seconds after imports; completing the eight cells adds 79.41 seconds, including 57.42 for the missing six-layer means, 1.24 for the new fit and 20.75 for all-cell evaluation and bootstrap analysis. All document vectors and teacher targets are reused. Original native vectors and all 1,296 natural full fingerprints reproduce exactly, including the mean-versus-sum numerical check. All eight output arrays are retained, and the new fit is sealed before any new evaluation. These remain historical-population mechanism comparisons, separate from actual request timing.

A paired capability check examines the finer distinction available from the full endpoint. On all 225 official NevIR development pairs [30], the historically projected native full-query-six and full-query-six ridge each answer both questions correctly on 38 pairs (16.89%); table ridge succeeds on 27 (12.00%). Native minus table is .04889 with a 115-paragraph-group interval [-.00446,.10268]. The contextual endpoints agree in aggregate but differ on eight pairs. Public text previews were seen while resolving the source, and pretraining overlap is unknown.

For document-one-minus-document-two margins a, b, define $k = (a - b)/2$ and $c = (a + b)/2$. Paired success requires $k > |c|$. The native full-query model has positive k on 157 pairs but succeeds on only 38. Applying the same source-mean correction to the inherited table head makes this distinction concrete: successes rise from 5 to 30 pairs, a difference of .11111 [.06667,.15789], while positive k remains 131/225. A common offset changes c, leaving k unchanged in real arithmetic; the numerical residual is below 6×10^{-19}. Both-wrong pairs also rise from four to fourteen. This retrospective, fixed-consumer check reuses every existing input and document, with no target fitting or new encoder call. Its contrast with table ridge is unresolved. All original models fail the sole exact token-multiset reversal, so neither the full-index context gains nor the offset recovery establishes reliable fine-grained reversal capability.

A fixed native-tower comparison separates that projection from the model pairing. Reusing all 450 raw Arctic document vectors, native Leaf queries solve 48/225 pairs and native Arctic queries solve 37/225, a difference of -.04889 [-.08772,-.01339]. With the unchanged source-direction projection applied to both sides, each solves 38/225. The projection lowers Leaf accuracy by .04444

[.00889,.08072]; its Arctic effect is unresolved. All four arms have positive k on 153–157 pairs, yet prefer the same document twice on 170–179. Every historical Leaf output is reproduced bitwise, and double precision changes no preferences. This historically exposed, fixed four-arm check changes architecture, checkpoint, depth and query precision together; it shows that neither a deeper native tower nor the inherited projection automatically supplies the missing paired capability. No new head is fitted.

E.3 Exact support invariance and conditional complement

This subsection describes the original lexical and conditional-complement executions. Its exact parity and cost statements refer to those executions, not the new FP64-pooling completion and budget panels in Appendix N.

The source-observed set is reconstructed from the original prompted, truncated R2048 and verified against the saved document frequencies. All four tables preserve its 4,555 rows bitwise and reproduce all 2,048 source pooled features exactly. Source OLS uses equal row weights and the declared float64 numerical-rank cutoff; all 384 raw coordinates remain, with condition number 21.75. Eight-neighbor completion uses raw-vector cosine and ID tie-breaking, without normalizing teacher rows. The global affine map was previously fitted to every vocabulary row, so it is an additional-information control rather than a source-only competitor.

Full minus neighbor completion is .17296 [.14842,.19961] on ArguAna and .06791 [.05099,.08592] on NFCorpus. Full minus global affine is .04520 [.03383,.05709] and .00564 [-.00438,.01461]. Global affine changes both the teacher-information support and its fitting distribution; its lower vocabulary MSE does not make it a universal upper bound. Construction and evaluation take 11.35 and 9.21 single-thread CPU seconds, with no encoder forwards or new request-latency measurement.

A post-result cross-read of the fixed context/readout comparison examines the 77 NFCorpus queries where this intervention is exactly inactive. Their table nDCG is .31277; full-query-six ridge reaches .33684, difference .02406 [-.00392,.05544], while original native-six reaches .34591, difference .03314 [.00621,.06419]. These small-stratum intervals are descriptive. The native contrast shows remaining endpoint benefit without missing lexical rows; the uncertain matched-ridge contrast does not isolate a pure interaction effect. Query coverage, offline lexical responses and full-query processing remain distinct parts of the explanation.

The conditional-complement study uses the same R2048 and source folds. Table and PC2 feature blocks are scaled by training-fold centered RMS, then weighted by square roots of $1-\rho$ and ρ for $\rho \in \{0, .25, .5, .75, 1\}$. The ridge penalty retains total kernel trace divided by 384, so the two endpoints reproduce the existing separate ridge families. Five-fold teacher-coordinate error selects $\rho = .25$ and $\alpha = .1$. A separate five-fold/eight-penalty ridge predicts PC2 features from the table. Writing the joint head as $F(z,p) = W_z z + W_p p + b$ and the prediction as $\hat{p} = Az + a$, the compiled control is $C(z) = (W_z + W_p A)z + (b + W_p a)$. In real arithmetic $F - C = W_p(p - \hat{p})$; FP32 compilation differs from its float64 two-stage expression by at most 2.65×10^{-7}. This is a fixed-consumer intervention, not an independently optimized static head or an information bound.

Source fitting and fixed evaluation take 18.23 and 9.80 CPU seconds, including .452 seconds to recover the 323 missing NFCorpus PC2 pools. All 2,592 original table/PC2 natural fingerprints remain unchanged. The transport residual relative to the fixed source-mean predictor's error is .427 on source E512, .750 on ArguAna and .466 on NFCorpus; these are approximation diagnostics, not fractions of task information. Both cost envelopes choose table/adapted-six mixtures. Their deep

probabilities are 37.45% and 38.13%, and cost/mixture are reselected in each paired bootstrap. The six-endpoint requests run once with another research task's single-CPU log replay active; uncertainty describes query composition within that shared-host session. Full coefficients, four-arm quality, six-arm clocks and per-query fingerprints remain in the evidence package.

E.4 A fixed position-weighted consumer does not close the deep-query gap

A subsequent historical-population test retains the deep teacher-token table and the complete original indices, and replaces uniform pooling by one prespecified mass-preserving rule. With effective input length L and $m = L - 10$ content positions, content position $j \in \{0, \ldots, m-1\}$ receives weight $1 + .5(2j/(m-1) - 1)$ for $m > 1$; for $m \leq 1$ its weight is one. The first nine formatting positions and terminal position retain weight one, and the pooled sum is divided by L. The source-only ridge head uses the original R2048 queries, five folds and penalty grid; it selects .1. This is a new positional consumer fit, not a source-identical lexical completion arm. The slope, population and comparison rule were fixed before new outcomes; no alternate slope follows the negative result.

Uniform and positional nDCG are .50823334/.44408630 on ArguAna973 and .30405278/.26995122 on NFCorpus323. The paired differences are -.06414703 [-.07767821,-.05222394] and -.03410155 [-.04876793,-.02001148]. Four thousand paired resamples use the original 134 ArguAna groups with query-weighted means and the 323 NFCorpus query units. The two 97.5% intervals implement the prespecified nominal 95% Bonferroni family. Both positive-gain gates fail, so the subsequent quarter-gap claim tests do not open. The original full-query-six ridge values remain .58483 and .35245. These exposed populations are not fresh confirmation, and failure of this one recipe does not rule out every position-aware model or identify the residual as a pure context effect.

The numerical contract was fixed before execution: bounded FP32 differences plus exact ordered top-ten agreement for the uniform replay, not bitwise equivalence of new pooling and affine arithmetic. The original uniform refit on saved source features reproduces the saved CV, weight and bias bytes; newly accumulated uniform source features instead differ by at most $1.4901161193847656 \times 10^{-8}$. All 1,808 source and natural uniform top-ten rows agree. The new uniform lane matches 0/1,296 strict historical natural score fingerprints; scoring the saved original vectors reproduces all 1,296. Ordered token IDs agree between the configured tokenizer implementations on all 3,856 texts, and original ArguAna input hashes agree on all 973 queries; historical ordered-token hashes are unavailable for source and NFCorpus. These distinctions preclude a claim of a bit-exact new end-to-end replay.

The immutable execution record preserves the fixed protocol and all saved comparisons. Its independent audit recomputes metrics and bootstrap contrasts from saved rankings, verifies the saved numerical receipts and traces the tokenizer/score gates without rerunning tokenization, fitting or complete-index scoring. Recorded fit, prediction and label-opening times are internally consistent saved receipts, not an external timestamp attestation. Neither offline execution time nor this audit measures serving latency; no new same-cost or frontier claim follows.

Appendix F. Complete-request costs, allocation and static alternatives

F.1 Judged quality and strong static controls

The historical SciFact evaluation connects the prefix gains to retrieval judgments. The dedicated PC shallow model raises nDCG@10 from .5337 for the matched frozen-prefix control to .5601, a difference of .02643 with component-bootstrap interval [.00607, .04682]. Shared training reaches .5580. The three-seed means cost 3.607 and 3.621 ms per warm request, respectively, compared with 3.544 ms for the frozen control.

PC endpoint	nDCG@10	Recall@10	Warm request, ms
Frozen prefix, shallow	.533713	.674883	3.544
Dedicated prefix, shallow	.560144	.695227	3.607
Shared prefix, shallow	.557964	.699531	3.621
Dedicated prefix, continued deep	.718617	.812207	7.169
Shared prefix, continued deep	.720567	.815728	7.134
Original PC6	.714172	.793427	7.098

Table S13. Primary historical 142-query population. Costs include tokenization, encoder, readout, complete 5,183-document scoring, and sorting. Wall times are measured in three fresh serial processes on a shared CPU environment.

A bank of 52 pre-existing static endpoints includes numerical precisions, depths, readout families, and dedicated or joint widths. At the dedicated shallow model's actual cost, a randomized mixture of two old static endpoints attains .5655 nDCG, compared with .5601 for the trained prefix. Their difference has an interval crossing zero. The corresponding shared shallow comparison is also inconclusive. The gains in fidelity and over the matched frozen-prefix control do not establish an improvement over this stronger static envelope.

The continued deep endpoints show a metric-specific improvement. Their nDCG gains over original PC6 have intervals crossing zero. Recall gains are .01878 and .02230 for dedicated and shared training, with intervals [.00228, .04082] and [.00111, .04861]. The shared preservation loss improves teacher fidelity more clearly than judged quality in this population.

F.2 Alignment executes at the cheap endpoint's cost

A subsequent 89-mode measurement adds the fixed ridge and pooled-capacity models to the earlier controls, using the same historical 142-query population and three fresh serial processes. Width-256 ridge reaches .5573 nDCG at 2.698 ms, while the full-KL three-seed mean reaches .5393 at 2.684 ms. Their quality difference is .01802, with interval [−.00967, .04816], and their cost difference is .0140 ms, with interval [−.0044, .0362]. The fresh teacher-fidelity gain therefore accompanies comparable actual source request cost, but the historical human-quality difference remains uncertain.

Width-400 ridge reaches .5578 nDCG and .7289 Recall at 2.809 ms. Dedicated-prefix shallow retrieval reaches .5601 nDCG at 2.874 ms, and the pooled MLP reaches .5245 at 4.459 ms. A non-self-inclusive bank of 70 previously fixed endpoints gives same-cost quality differences of .0080 for width-256 ridge and -.0073 for width-400 ridge, both with intervals spanning zero. This expanded comparison establishes neither ridge nor prefix adaptation as a new static frontier.

These later measurements are internally contemporaneous; their absolute times should not be compared with Table S13's earlier run. They include 75,828 timed output-parity checks. CPU and thread CPU are also recorded over blocks of 16 singleton requests, avoiding the earlier study's zero-duration per-request CPU intervals; wall time remains the primary measure. Natural-query inference is longer and is not inferred from these source measurements. The next section measures all 101 fixed endpoints directly on natural queries.

F.3 A better static endpoint at the natural-query budget

The natural-query experiment runs 101 fixed endpoints in three fresh serial processes on 64 queries selected by an identifier hash before implementation execution. The complete prior 52-endpoint bank and all previously fixed prefix, width and objective controls are retained. Each request independently tokenizes, encodes, reads out, scores all 8,674 documents and sorts. The quality comparison in the last two columns of Table S14 uses the same 64 queries as the cost measurement; the all 1,406 column is reported separately.

Fixed endpoint	All 1,406 nDCG	Panel 64 nDCG	Panel 64 wall, ms
Ridge, width 256, raw	.387051	.353088	15.509
Ridge, width 400, raw	.419567	.400794	15.623
Full-KL, 256	.274098	.254615	15.241
Dedicated prefix, shallow	.270668	.236584	15.747
Shared prefix, shallow	.278400	.244144	15.763
Pooled MLP	.269043	.229982	17.586
Dedicated prefix, deep	.579557	.587199	39.568
Shared prefix, deep	.581795	.576005	39.014

Table S14. Natural ArguAna fixed-endpoint quality and actual warm request cost. Adapted, full-KL and MLP rows average three fixed seeds. Latency covers the preselected 64-query panel, whose 55 historical groups define its paired uncertainty; no full-population latency is inferred.

At width-400 ridge's 15.623 ms budget, the non-self-inclusive prior 86-endpoint static-mixture envelope achieves .27016 nDCG, versus ridge's .40079. The paired gain is .13064 [.04873, .20995]. At width-256 ridge's budget, the gain is .08447 [.01049, .16002]. The smaller old 52-endpoint bank gives nearly the same comparison. These banks exclude all new ridge and MLP candidates, and the prior 86-endpoint bank includes the deep adapted controls. The budget improvement therefore survives comparison with the strongest previously fixed bank, including its best randomized mixture.

Width-400 ridge's quality gain over the dedicated prefix comes at comparable cost: the wall difference is -.124 ms [−.297, .044]. Width-256 ridge costs .268 ms [.166, .381] more than full-KL, width 256 and gains .09847 [.02475, .17500] panel nDCG. For the focal 2,048-query ridge comparisons against the static envelopes, Recall intervals span zero. The measured gain is consequently metric- and budget-specific.

All 38,784 timed requests reproduce their complete same-process warm output arrays, and warm top-ten results agree with the sealed predictions. Positive block CPU/thread intervals avoid the earlier short-counter pathology; no block exceeds wall by 5% in process CPU. Known shared-host workloads are retained and documented. Mixture weights and costs are fixed at their panel estimates during the paired bootstrap, making the intervals conditional descriptive comparisons. The study executes fixed endpoints and establishes a better static option on this panel; it does not evaluate a new query-adaptive gate.

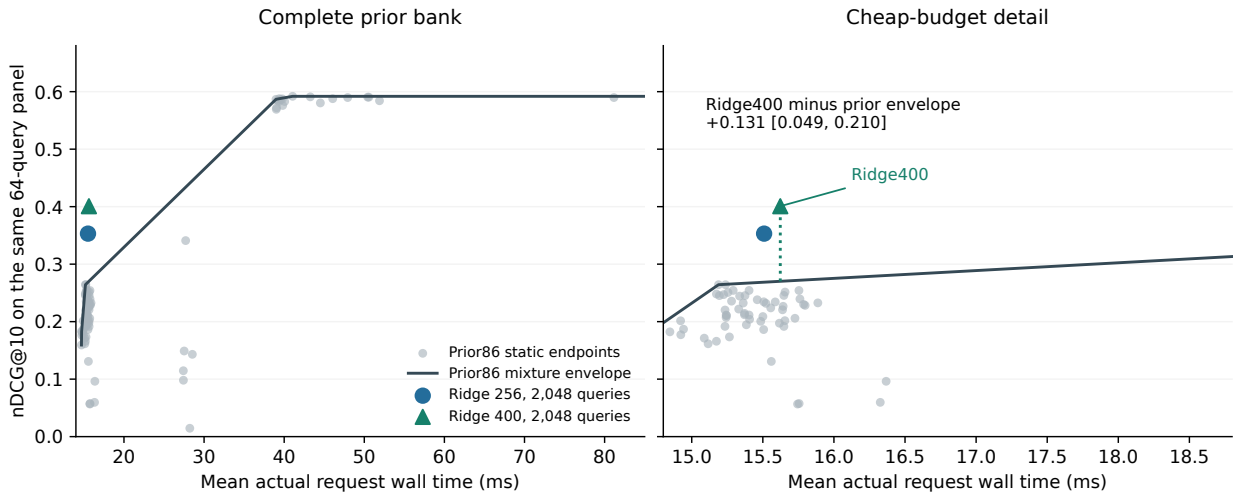


Figure 9: Natural-query actual cost and quality on the same 64-query panel. The complete prior 86-endpoint bank and its static-mixture envelope are shown with the two 2,048-query ridge endpoints. The enlarged cheap-budget region displays the paired width-400 ridge-minus-envelope nDCG interval; mixture weights are chosen on the panel.

F.4 Observation and continuation have separate prices

A budget decision made after scoring a shallow representation has already consumed the shallow readout, corpus scan, and sorting. We execute this path and then continue the encoder. Dedicated and shared continuation cost 8.479 and 8.474 ms, compared with 7.169 and 7.134 ms for direct deep execution. The shallow observation fee is therefore 1.310 and 1.339 ms.

Reusing the changed first-two-layer states saves 1.775 and 1.740 ms relative to restarting the same changed prefix. This compares identical final outputs. Restarting the original prefix returns the original teacher instead, so it is a different endpoint contract. Tokenization and embedding outputs are reused in both restart controls. The savings are positive for each query after averaging the three seeds; each arm has two nonpositive individual seed/query comparisons out of 426.

The experiment includes 85,200 timed requests and complete output-equivalence checks against each request's warm reference. Wall time is the primary cost. Individual process-CPU intervals contain 5,751 exact zeros and approximately 4 ms clustering, despite positive wall times. We retain those observations and use their aggregate CPU values only as diagnostics. No CPU-only speed claim is based on the anomalous short intervals.

F.5 Transfer changes the decision regime

A source-frozen adaptive policy using the earlier frozen shallow readouts was also executed on all 1,406 ArguAna queries over 8,674 documents. This experiment predates the new ridge and trainable-prefix models. The two-layer utility policy continues 23 queries, whereas the four-layer policy continues 1,352. Every target query is longer than the source fitting maximum of 57 effective tokens; the target minimum is 62, median 227, and maximum 512. Task and length support change together.

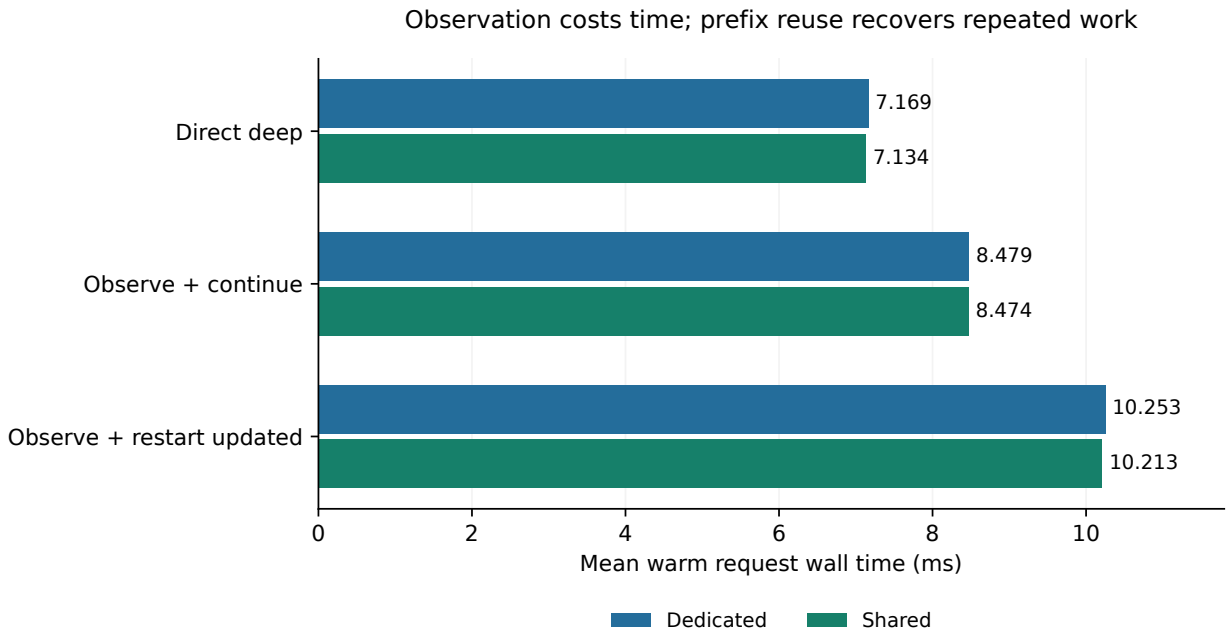


Figure 10: Whole-request costs for direct deep retrieval, a paid shallow observation followed by continuation, and a restart of the same changed prefix. Continuation saves repeated encoder work while still paying for the shallow observation.

On the fixed 256-query timing panel, the two-layer policy attains .27952 nDCG at 16.204 ms. The strong same-cost static envelope attains .27803, giving a difference of .00149 with interval $[-.00768, .01350]$. The four-layer policy costs more than original PC6 and is dominated by a fixed deep readout. These results identify the transfer of the decision regime as a practical bottleneck. They do not isolate length as semantic difficulty. ArguAna is historically exposed development material, and its labels were not used to retune the frozen policy after this run.

F.6 A fixed allocation rule transfers to new query groups

Let $Q_c(x)$ and $Q_d(x)$ denote cheap and deep nDCG, and let $g(x)$ indicate continuation. Relative to query-independent random continuation at the same fraction $p = \mathbb{E}g$, the quality gain is exactly

$$\mathbb{E}[Q_c + g(Q_d - Q_c)] - \mathbb{E}[Q_c + p(Q_d - Q_c)] = \mathrm{Cov}(g, Q_d - Q_c).$$

The standard covariance identity measures allocation beyond the advantage of a stronger average deep endpoint; the gain depends on how its query-specific effects align with the gate. The original PC2 ridge400 and PC6 native Dense endpoints differ in both encoder depth and readout.

On the new 973 queries, the fixed primary continues 462 (47.48%) and reaches .53886 nDCG, against .42730 for always cheap and .59537 for always deep. A query-independent mixture at the same continuation fraction reaches .50710. The allocation advantage is .03175 [.01994, .04372], with the fraction recomputed in every group-bootstrap draw. The primary gains .11155 [.08898, .13558] over cheap and remains .05651 below always deep. It continues 271 rescues and 53 damages, while skipping another 218 rescues and 82 damages.

All eight prespecified policies are retained. The four standard-deviation cuts yield allocation gains .01223/.02778/.03175/.03224, all with pointwise intervals above zero; the two lower gap cuts have intervals spanning zero. A descriptive random policy preserving continuation rates within five input-length bins reaches .50920, leaving .02965 of primary advantage. Coarse length composition explains little of the mean gain, but shallow retrieval quality and finer input statistics remain alternative explanations.

The exact shallow-rank decomposition locates much of this allocation advantage. Its between-rank contribution is .02976 [.02404,.03643], while its within-rank contribution is .00199 [-.00795,.01158]. The old within-rank estimate was .00997 [.00165,.01741]. Human rank is visible only to this post-result analysis. At shallow rank 1, continued low-std queries lose .21133 nDCG versus .09164 among skipped queries. A low shallow score spread therefore identifies a different opportunity regime, rather than universally identifying queries that deeper computation repairs.

A companion action contrast holds the final four layers and native Dense head fixed and replaces only the first two layers with the existing adapted prefix. Taking the difference between the two gate covariances cancels the common shallow quality. For the dedicated three-seed mean, the adapted-minus-original covariance is +.00263 [.00050,.00480] on old1406 and +.00258 [.00041,.00473] on new973; within exact shallow rank it is +.00263 and +.00220, with both intervals above zero. The shared-family contrasts are smaller and uncertain. This post-result association is more action-specific than the broad headroom decomposition, but much smaller than the overall .03175 allocation gain. It concerns the effect of a changed prefix on a fixed continuation, including the changed inputs to its later layers. Since the gate observes the original prefix, executing this adapted action would require prefix recomputation.

F.7 Allocation value after the full request is paid

The continuation request first pays for two encoder layers, a ridge readout, full-corpus scoring and sorting, and the score observation. It reuses those token states if selected, executes the remaining four original layers and native Dense, then scores and sorts the full index again. The original observer performed another complete unit-vector score scan. An equivalent implementation scores only the ten already returned addresses, using a conservative finite-precision bound and falling back to the original scan near the fixed threshold. All 973 by eight decisions agree; the primary falls back on seven queries. This change preserves the gate and its complete retrieval output.

We execute all 125 static endpoints and eight paid policies on all 973 queries in one fresh single-thread process. Each mode receives one global warmup and one measured request per query; mode order is randomized within 61 query blocks. Every actual output is checked after timing against its vector, all 8,674 scores, full ordered ranks and top ten. All 129,409 timed checks pass. The machine remains shared with other recorded scientific workloads. These mean costs describe the complete query population in this hardware session, rather than a per-query deadline or a distribution over future hardware sessions.

The primary reaches .53886 nDCG at 31.527 ms. The complete static-mixture envelope reaches .53431, a difference of +.00455 with reselected-mixture/cost interval [-.01109,.01613]. Process-CPU and thread-CPU comparisons give +.00449 and +.00455. The selected static mixture combines frozen ridge400/unit with a dedicated adapted-prefix continuation. All eight policy intervals span zero; std q40 has the largest point difference, +.00967 [-.00695,.01972], without replacing q60 as the primary. Thus the small positive point difference does not yet establish superiority or equivalence to the static frontier.

Let $\Delta Q = Q_d - Q_c$, $\Delta C = C_d - C_c$, C_g denote the executed policy cost, and $\alpha = (\mathbb{E}C_g - \mathbb{E}C_c)/\mathbb{E}\Delta C$. If F is the complete static envelope's advantage over the original cheap/deep line at that budget, the difference from the envelope is exactly

$$\mathrm{Cov}(g, \Delta Q) - (\alpha - p)\mathbb{E}\Delta Q - F.$$

For the primary, this is .03175 allocation minus .01673 relative price minus .01048 stronger static alternatives, leaving .00455. The price term includes the kinds of queries selected for continuation. Using the existing per-query times, its 2.551 ms excess over the same-rate average-cost mixture splits into .734 ms of selection/cost covariance and 1.818 ms of executed-path residual. The latter includes both extra path work and interleaved runtime variation.

A separately frozen observer-only comparison on all 973 queries, with five repetitions, reduces mean component cost from 218.46 to 21.78 microseconds. Eliminating that redundant scan saves .197 ms; it accounts for little of the overall price of this continuation request.

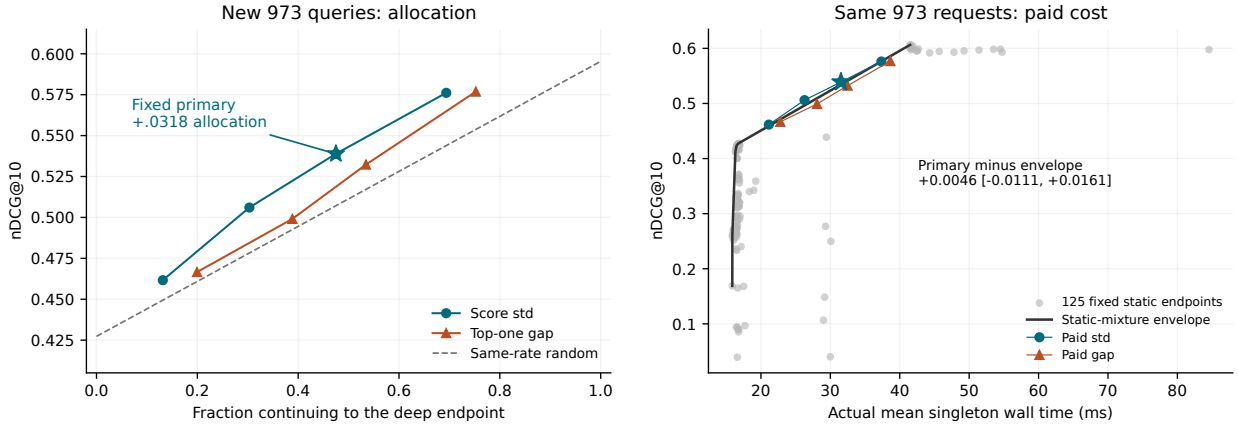


Figure 11: The fixed gate allocates computation usefully, while stronger static actions remain competitive at its measured budget. Both panels use the same 973 new queries. Left: all four fixed thresholds of each feature and the same-rate random line. Right: all 125 static endpoints, their mixture envelope and all eight paid policies at actual complete-request cost. Stars mark the unchanged primary; the displayed interval reselects the static mixture and its cost within paired 134-group resamples.

The earlier 64-query, two-process experiment remains informative about panel composition. Its primary difference was -.00646 [-.06446,.03784], with 34,048 timed full-output checks. Within the new execution, the same identifier-hash64 subset still gives -.00566, using the same pair of envelope endpoints as before. The full973 point estimate therefore cannot be explained as an observer-induced reversal. The timing protocols also differ: the earlier panel warmed each query/mode and measured two repeats in each of two processes, whereas the full population uses one request after global warmup. Absolute timing changes mix protocol, population and implementation effects; the matched component measurement isolates only the observer rewrite.

F.8 A new task changes the value of continuation

NFCorpus supplies a prospective test of the same observation, threshold and original cheap/deep actions. The gate continues 79 of 323 queries. Deep-minus-cheap nDCG is .04736 among continued

queries and .04125 among stopped queries, producing allocation value .00113 [-.00499,.00701]. The query-conditional advantage that was substantial across new ArguAna groups is weak on this task.

All 3,876 complete timed requests and twelve warm requests match frozen full-vector, full-score and ordered-return fingerprints. Tangent costs 2.271 ms versus 2.088 ms for original cheap, an 8.76% increment. The paid gate costs 3.345 ms and obtains .32535 nDCG. At that budget, the retained eleven-static envelope mixes ridge256 with adapted native continuation, using the latter on 42.91% of random queries, and reaches .33481. Gate minus envelope is -.00945 [-.01765,-.00326], with costs and mixture reselected in each bootstrap draw. This comparison retains the preselected eleven statics rather than reopening the earlier 125-model bank.

The accounting in Section F.7 now gives .00113 allocation minus .00511 relative price minus .00548 stronger static alternatives. Even a mixture of just the original cheap and deep endpoints could afford 36.40% deep calls at this cost, compared with the gate's 24.46%. Consequently the observed transfer loss involves both diminished action selection and the price of the executed continuation path. Thresholds and corrections remain fixed throughout this test.

F.9 Ranking movement survives while action value weakens

The NFCorpus transfer result leaves two possibilities: the gate may cease to predict the deep action, or the predicted action may cease to be useful. We distinguish them using the fixed rankings, original judgments and unchanged gate. Let $w_a(d)$ be the logarithmic rank discount of document d under action a, zero outside the top ten. Write $\delta w = w_{\text{deep}} - w_{\text{cheap}}$, and define

$$M = \frac{1}{2}\sum_d |\delta w(d)|, \qquad B = \frac{M}{\text{IDCG}}, \qquad R = \frac{\sum_d \delta w(d)\,\text{gain}(d)}{M}.$$

Both rankings have the same total discount mass. Hence M is the amount reassigned, while R is the gain difference between documents receiving and losing that mass. Their product gives the exact action benefit $G = BR$; set $R = 0$ when $M = 0$. This is a decomposition of the existing DCG functional.

The gate continues to select larger movements. Continued-minus-stopped mean M is 1.0058 [.8770,1.1325] on ArguAna and 1.1808 [.9825,1.3704] on NFCorpus. Yet their conditional nDCG-gain contrasts are .12733 and .00611. Expanding $\mathbb{E}[BR]$ within each gate group splits this contrast into a mean-amplitude term, a mean-direction term and the difference of within-group covariances. The three terms are .06758, .04980 and .00995 on ArguAna; they are .06718, -.00435 and -.05672 on NFCorpus. The last NFCorpus term has interval [-.09501,-.01976]. Larger ranking changes and favorable relevance directions cease to combine in the same way among continued queries.

This pattern survives replacing ordinal gains with binary gains and replacing query-specific IDCG with a common discount scale. Under binary nDCG the NFCorpus covariance difference is -.05679; under binary DCG divided by the common ideal top-ten mass it is -.02004 [-.03019,-.00972]. The weakness of the original gate therefore cannot be assigned solely to graded labels or normalization. First-hit MRR retains a larger, uncertain allocation gain, .00937 [-.00254,.02222], while binary nDCG gives .00043 [-.00560,.00618]. We retain all eight fixed functionals rather than selecting a replacement metric.

These are descriptive identities: B and R share M, and their covariance does not identify independent causal components. In particular, the unweighted NFCorpus direction contrast spans zero. The

result locates the missing predictive link more precisely than a generic uncertainty failure. Existing work already distinguishes query-performance prediction from optimal reranking depth [25]. Here the same frozen encoder action remains predictable in ranking movement, while its conditional relevance return changes across tasks.

F.10 Strong static alternatives set the price of selection

We next test one source-calibrated policy with complete execution. Its cheap action is the cached PC6 token table. After retrieval, it reads ten document addresses and computes the standard deviation of their float64 scores under a float64 unit query. Let z be this observation and $z_1, \ldots, z_{512}$ the same observations on the existing unlabeled source E512. The fixed continuation probability is

$$g(z) = \frac{1}{512} \sum_{i=1}^{512} \mathbf{1}\{z_i > z\}.$$

A predeclared task/query-ID hash supplies one nine-bit uniform draw. The request continues if that integer is below $512g(z)$. The sign, calibration, hash namespace, seed and actions are fixed before computing target policy outcomes. Both target populations are historically exposed at this stage, so this is a fixed development experiment. The source mapping is not fitted to their labels, and its conditional expected quality is reported separately from the single realized policy.

Every request pays for tokenization, cached retrieval, observation and the decision. Continued requests then run all six original layers and native retrieval; there is no Transformer-prefix reuse from a token table. Only tokenization and its tensors are shared. Six static controls run in the same randomized, sixteen-query-block schedule: cached and raw token means, original PC2 and PC6, and the existing dedicated and shared adapted-prefix continuations. These include the prior bank's two nondominated deep alternatives. The comparison is against this matched six-static subset, not a newly timed complete 125-endpoint bank.

Table S15. Actual source-CDF policy and same-cost six-static envelope. Each population is measured in one fresh single-thread process. The envelope reselects its mixture in each paired bootstrap sample.

Population	Deep calls	Paid nDCG	Paid ms	Static envelope	Paid minus envelope, 95% interval
ArguAna	365	.56167	17.730	.54951	+.01217 [.00144, .02256]
NFCorpus	100	.32102	2.586	.32979	-.00877 [-.01756, -.00158]

On ArguAna, the source rule's mean continuation probability is .36813. Its exact conditional expected quality is .56247, with .02216 [.01539,.02857] allocation gain above same-rate random continuation. The prescribed hash realizes .56167 quality at 17.730 ms. At that cost, the strongest static mixture sends 42.06% of queries to the dedicated adapted continuation and otherwise uses the cache, reaching .54951. The paid policy exceeds this six-static envelope by .01217 [.00144,.02256] (Table S15).

NFCorpus gives a different result under the same rule. Its mean probability is .31095 and conditional expected allocation gain is only .00168 [-.00263,.00606]. The realized policy reaches .32102 at 2.586 ms. Its static competitor can afford 46.32% dedicated continuation and reaches .32979, a

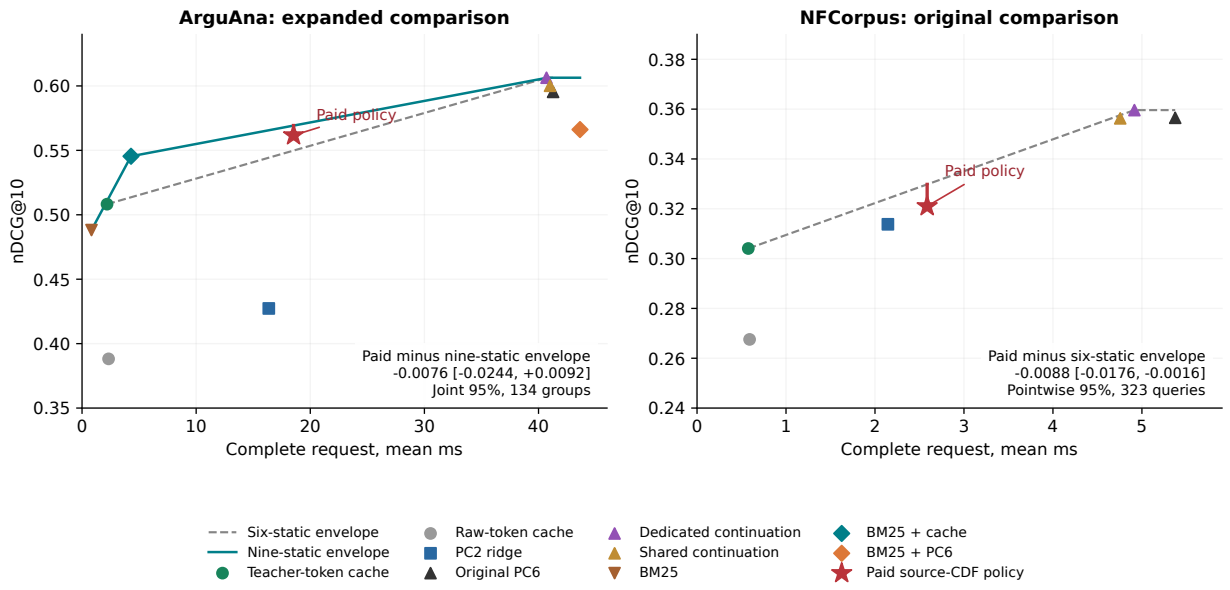


Figure 12: The static comparator changes the interpretation of the same fixed policy. Left: a new common execution on historical ArguAna retains all six neural statics and adds fixed BM25 and two RRF actions. The policy lies above the six-static curve but below the expanded curve in point estimate; the displayed simultaneous interval spans zero. Right: the original six-static NFCorpus execution remains negative. Each panel uses only its own session costs; intervals resample query groups or queries with mixture reselection, not hardware repetitions.

paid-policy deficit of .00877. The source calibration has no large-scale saturation: only three queries have probability zero and one has probability one. The weak action information, rather than a saturated gate alone, remains the important boundary. On realized continued versus stopped queries, deep-minus-cache nDCG averages .14246 versus .05393 on ArguAna, but .05479 versus .05142 on NFCorpus.

A supporting-line calculation makes the cost tradeoff explicit. Let $M_\lambda = \max_j\{E[Q_j] - \lambda E[C_j]\}$ over the six static actions, choose a supporting slope at the policy's measured mean cost, and define $R = C_{\text{paid}} - C_0 - GC_6$. Then

$$\begin{aligned} E[Q_{\text{paid}}] - F(E[C_{\text{paid}}]) = E[Q_0] - \lambda E[C_0] - M_\lambda \\ + E[G]\{E[Q_6 - Q_0] - \lambda E[C_6]\} \\ + \operatorname{Cov}(G, Q_6 - Q_0 - \lambda C_6) - \lambda E[R]. \end{aligned}$$

This is standard convex-envelope accounting. In the initial six-static comparison, the cache lies on the relevant supporting segment in both tasks, making the first term zero. ArguAna's price-adjusted selection covariance is +.02006, against -.00684 from mean serial escalation and -.00105 from the execution residual. NFCorpus has only +.00093 against -.00506 and -.00464. The residual includes observation, token reuse, runtime coupling and timing variation; it is not an isolated observer-cost estimate. Exact expected quality under g is also not assigned a fictitious measured latency: only one hash realization is executed.

All 9,072 timed complete returns and fourteen warm returns match frozen fingerprints. The 1,296 cheap observations reproduce the source-calibration arrays exactly. The two task processes use 167.57 CPU seconds including setup; the initial source lookup construction remains a separate offline expense. A short overlapping CT computation is retained in the shared-host record. Four thousand paired resamples use the original 134 ArguAna groups or 323 NFCorpus queries and jointly reselect cost and static mixture. The historical population and single hardware pass bound the statistical claim. Within that original six-static comparison, the cached-first construction has a positive pointwise contrast on ArguAna and a negative one on NFCorpus. The cheap endpoint and decision rule both differ from the earlier paid-PC2 pipeline, so the comparison does not isolate the cache as the sole cause. The broader comparison below changes the ArguAna interpretation.

A fixed classical control tests whether that apparent paid advantage survives a stronger inexpensive alternative. We transfer the NFCorpus BM25/RRF configuration in Section G.1 unchanged to all 973 historical ArguAna queries. BM25, BM25+table and BM25+PC6 reach .48814, .54542 and .56615 nDCG. Fusion improves the table by .03719 [.01588,.05851] but lowers PC6 by .02922 [.01111,.04734], under a four-contrast simultaneous group band. Complete raw-text lexical retrieval and the neural encoder's inherited 512-token interface remain distinct fixed actions.

A new common execution retains the original seven modes and adds those three static actions. All 9,730 timed outputs and all gate decisions reproduce their fixed references. The unchanged policy reaches .56167 at 18.536 ms. The contemporaneous six-static envelope is .54989, while the expanded nine-static envelope rises to .56927 by mixing the 4.291 ms BM25+table action with dedicated continuation (Figure 12). Policy minus envelope changes from +.01178 to -.00760. The simultaneous two-comparison 95% group intervals are [-.00031,.02387] and [-.02442,.00923]; the original-six pointwise interval remains positive. Thus the earlier conditional result survives as a comparison with six neural alternatives, while paid superiority is not established against the expanded benchmark. The negative point estimate does not establish inferiority either.

The selection signal itself remains useful. At the expanded supporting slope, price-adjusted selection covariance is +.02030. The cache now lies .03367 below the supporting line; mean serial escalation and the execution residual contribute +.00677 and -.00099. Their sum is the observed -.00760 gap. Stronger static evidence changes the opportunity cost of the same observation. The sparse index uses 5.85 MB of numerical arrays and takes 1.08 CPU seconds to build; the complete single-thread batch takes 211.39 CPU seconds on the shared host. The groups and models are historically exposed, and the actual-cost implementation follows the fixed quality screen. No gate or fusion parameter is changed after the result.

F.11 New queries separate action information from a paid advantage

Online rule	Native calls	nDCG	Mean ms	Static envelope	Gap, family 95% interval
Predicted signed gain	1,003	.32266	2.621	.32297	-.00031 [-.00353, .00284]
Negative cheap-quality prediction	943	.32345	2.424	.32048	+.00296 [-.00020, .00608]
Unchanged source-CDF/hash	916	.32005	2.393	.32009	-.00004 [-.00325, .00307]

Table S16. All retained online rules on the 2,890-query population. Each request pays for cached retrieval, observation, the decision and any original-six-layer continuation. Static mixtures use six same-session endpoints. The interval column applies a Bonferroni family adjustment over the three paid-envelope contrasts, using paired query resampling with mixture reselection.

The negative NFCorpus result concerns one population and one transferred rule. We next fix a direct comparison of two supervised targets using exactly the same cheap information. Historical official test 323 becomes development. The remaining official train and dev queries supply new model-outcome rows: all query text bytes had already been fetched, but these rows had not been executed or used for relevance-based decisions. Before outcomes are generated, exact normalized-text and prompted-token components remove five rows overlapping development and nineteen within-new duplicates. All 2,890 representatives remain, comprising 2,570 train and 320 dev queries. These exclusions address exact input duplicates, not semantic-topic dependence.

Both scalar regressors receive the ten ordered, normalized cheap top-ten scores and log(1 + prompted, truncated token count). One predicts cheap nDCG; the other predicts original-native minus cheap nDCG. Centered linear ridge uses the same eight penalties and five exact-token-group folds, with feature standardization inside each fold. Cheap-quality regression selects penalty .01 and signed-gain regression selects 1.0. Both fitted models are sealed before the new endpoint outputs. The new train/dev judgments are first parsed after all outputs have been saved. There is no expensive-action feature or target-side refit.

The first judgment parse reveals a material dataset distinction: the BEIR train TSV contains only grade 1, whereas dev and the old test contain grades 1/2. We retain every original grade. The original corpus describes query-level random splitting and relevance derived from website links [22]. This experiment therefore uses reassigned statistical roles within one corpus, rather than standard held-out BEIR test performance or a new human-label collection. A post-result split description finds the same mean ordering in both the binary and graded portions; it does not create an additional confirmation set.

The prespecified information comparison ranks the whole unlabeled batch and allocates exactly 963 native calls. Direct gain prediction improves nDCG over the equal-rate random expectation by .00773 [.00528,.01041]; negative predicted cheap quality improves it by .01062 [.00778,.01304]. The old negative-standard-deviation score gives .00809 [.00549,.01081], while query-hash ordering gives -.00092 [-.00328,.00150]. The primary gain-minus-difficulty contrast is -.00289 [-.00541,.00001], with 748 selection disagreements. Direct supervision on the desired benefit has no demonstrated advantage in this small, regularized fit. These batch rankings establish information value at a fixed call count; they do not yet pay for an online decision.

A subsequent fixed 256-draw source-group bootstrap tests sensitivity without changing the fitted function class or the two selected penalties. Across these perturbations, score-rank correlation with the original fit averages .9805 for difficulty and .9166 for gain; agreement with the original 963-call

set averages .9530 and .8894. Selection probabilities lie between .1 and .9 for 433 versus 1,027 queries. Coefficient-direction cosine is higher for gain, so stable coefficients alone do not imply stable decisions. The original gain-only and difficulty-only groups nevertheless retain substantially different selection probabilities. Sampling sensitivity is present, while systematic estimator differences remain. These are developmental perturbations around the same 323 labels, evaluated on already exposed arrays; they add no independent relevance or hardware evidence.

We then retain both fitted regressors and the unchanged source-CDF policy for complete execution. For each regressor, its final fitted scores on development 323 set the score/query-hash boundary selecting 107 development rows. The frozen boundary applies independently to new requests, without using an evaluation quantile or equalizing their observed call rates. The coefficients predate the new outcomes, but this online implementation is specified after the offline information result; it is a historical execution study. All three policies and six static alternatives are retained together.

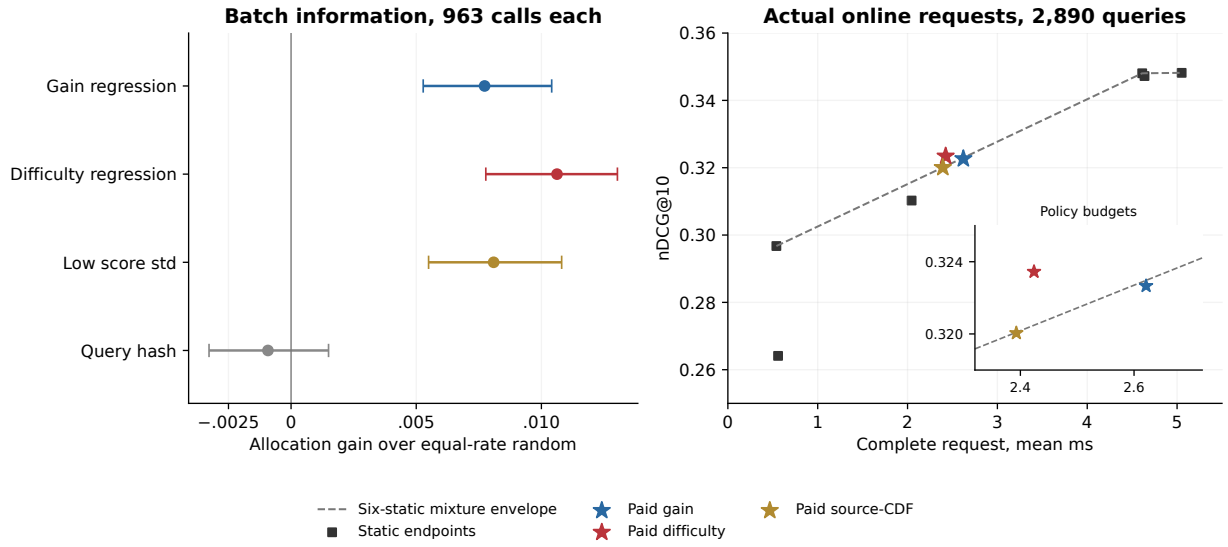


Figure 13: New-query action information and complete execution cost. Left: the four fixed batch rules use the same 963 native calls, with pointwise paired intervals for gain above random allocation. Right: all three historical-threshold online policies and six actual static endpoints; the inset enlarges the policy budgets. The best mean policy has a family-adjusted interval spanning zero.

The teacher-cache, original-PC2 and original-PC6 static endpoints reach .29674, .31025 and .34818 nDCG at .540, 2.045 and 5.051 ms. Dedicated continuation reaches .34806 at 4.612 ms and defines the active static segment with the cache. Difficulty has a pointwise positive contrast, +.00296 [.00032,.00544], which crosses zero after the three-rule adjustment (Table S16). A centered max-standardized bootstrap band retaining the correlation among contrasts also crosses zero, [-.00009,.00602]. The evidence places paid allocation near the strong static frontier without a jointly resolved advantage.

The cost accounting explains why a clear allocation gain can leave a small net result. On this active segment, let lambda = .01260249 nDCG per ms. The cache lies on the frontier, so the supporting-line identity reduces to

$$E[Q_{\text{paid}}] - F(E[C_{\text{paid}}]) = E[G\Delta] - \lambda E[GC_6] - \lambda E[R].$$

For the gain rule, these three terms are .02592, -.02290 and -.00333. For difficulty they are .02671, -.02090 and -.00285. Gain-selected native requests average 5.236 ms, versus 5.081 ms for difficulty and 5.051 ms unconditionally. Their difference in paid-envelope gaps is .00327: .00079 comes from quality and .00249 from the measured cost difference at the common frontier slope. This is exact arithmetic, not a causal attribution of runtime. It shows why predicting relevance improvement alone need not produce the most economical complete policy.

All 26,010 timed returns and nine warm returns reproduce their complete expected fingerprints. One single-thread process uses 73.88 CPU seconds including setup, with necessary independent research sharing the host. The paired intervals sample queries within this one execution, not hardware repetitions. The native action, document index and learned representations are unchanged. This larger population establishes transferable cheap action information, while preserving the earlier negative population result and the unresolved complete-cost comparison.

Appendix G. Candidate access, lexical fusion and score compilation

G.1 Evidence access and representation also change the budget

A fixed MS MARCO MiniLM-L6 cross-encoder raises table top20 retrieval from .30405 to .32934 nDCG on historical NFCorpus 323, below native deep's .35651. In that complete-request screen, paired FP32 interaction costs 4,401 ms CPU versus 7.277 ms for the INT8 query-only endpoint. Keeping the scorer and shared-pair logits fixed but supplying deep's top20 raises cross-encoder quality to .35583; the difference from table-pool interaction is .02649 with a four-comparison interval [.01233,.04065]. The larger union scores .35298. Access to more relevant candidates and the ability to order them are separate constraints; the new-pool experiment measures supplemental offline scoring cost, not a new serving latency.

A classical lexical source supplies a cheaper alternative. We fix stemmed BM25 (k1=1.2, b=.75) and equal-weight RRF60 [37,38] over each stream's first 100 documents. BM25, table, deep, BM25+table and BM25+deep reach .31874, .30405, .35651, .33639 and .35629 nDCG, respectively. Their same-run complete-request CPU times are .251, .363, 4.963, 1.065 and 5.945 ms. BM25+table improves over table by .03234 [.01518,.04949] under a four-comparison simultaneous query-bootstrap band, while remaining below deep. The sparse BM25 arrays occupy 3.17 MB and take .611 CPU seconds to construct. This fixed lexical fusion strengthens the cheap static comparison; it offers no mean quality increment when added to deep.

Execution can also change without acquiring new relevance information. For the frozen table scorer, let token counts be c_t, $n = \sum_t c_t$, and document vector d_j. Then

$$\left(W\frac{\sum_t c_t T_t}{n} + b\right)^\top d_j = \frac{1}{n}\sum_t c_t P_{tj}, \qquad P_{tj} = (WT_t + b)^\top d_j.$$

Vocabulary-score precomputation, quantized impacts and bound-driven refinement are established techniques [39–41]. In a separate seven-consumer session, full FP32 compiled rows reduce request CPU from .314 to .177 ms while increasing numerical-array storage from 53.31 to 443.57 MB. All 323 ordered top10 results remain unchanged. A folded-head control costs .343 ms at 54.65 MB. Four- and eight-bit rows with conservative refinement inspect only 81.36 and 11.37 original document rows on average, yet cost .453 and .404 ms after decoding, bounds and refinement. Compilation takes 2.07 CPU seconds for the dense in-memory control, and an index change requires corresponding

refresh work. Floating-point consumers have explicit reduction contracts; observed top10 agreement is not asserted for arbitrary reassociation.

Appendix H. Additional implications of the fixed interventions

H.1 Evaluate the increment over what is already available

Incremental query support, inherited information and the fitting rule jointly determine what a cheap model learns. The KL family's support reversal survives its source-null intervention, but prior-centered ridge invalidates a general claim that rich teacher targets require large query coverage. At small support, the large gain chiefly preserves available knowledge; a smaller gain remains beyond the unchanged and intercept-adjusted prior. At higher support, retaining the same prior can cost natural-task quality. Strong classical baselines therefore change both the explanation of data requirements and the preferred cheap endpoint.

Source adaptation yields real gains over the matched frozen-prefix control, yet the larger source teacher-overlap score does not determine the natural target ordering. The common-readout experiment further separates recoverable head loss from a width-dependent residual. Transfer evaluation therefore changes both the choice of a cheap endpoint and the interpretation of its capacity gains.

H.2 Preserve useful adaptation through the right consumer

The four-cell comparison points to an interaction between query errors and document geometry. For a fixed coordinate error e and uniformly sampled corpus documents, centered score MSE is exactly $e^\top G_D e$, where G_D is their covariance. This standard identity explains why the same embedding error can produce different score errors on different indices. It does not determine ranking without score margins, nor does uniform covariance explain the query-dependent softmax weighting in KL. Our calibration control supplies a direct, narrower result: scale alone can reverse a KL comparison without changing the top-ten retrieval decision.

Matched artificial-input interventions identify positional and token-multiplicity sensitivity in pooled feature transport, including a control that preserves the token states. The natural-task comparison then shows that source-fixed format mass can improve this geometric diagnostic while reducing retrieval quality. Transferable compression therefore requires task-relevant prediction directions and margins in addition to a stable feature-coordinate relation.

The shallow-correction comparison gives this distinction a practical consequence. A token-tangent correction improves retrieval beyond a full pooled-affine residual under the same local score target, while retaining one query vector and one index scan. Its advantage can reflect token covariance, the existing pooling nonlinearity, and the regularized function class jointly. Because affine and tangent corrections are nonnested, the comparison does not isolate incremental token information beyond a fitted affine predictor. A simple scalar counterexample makes the distinction exact: with token states $H = (0, 2z)$ and uniform attention, the pooled feature is z and token covariance is z^2. The tokens contain no information beyond z, yet a tangent correction can express az^2 while an affine correction cannot. Function-class and information comparisons therefore require different controls. The finite attention update retains most of the historical gain in the original graph; the prospective NFCorpus result then shows that its local teacher-margin counterpart can transfer more clearly than judged relevance.

H.3 Generalization and the experimental contract

The complete objective, capacity and serving study concerns the Leaf-query/Arctic-document pairing across three document corpora. The additional identity-center comparison uses Arctic queries with the same document tower and a new source panel. Query population, text length, corpus size, and task change together; neither coverage nor transfer is an isolated intervention on semantic diversity. The full-versus-candidate interaction is reproduced on later disjoint source panels and retained in the natural task's larger-support human-quality contrasts. The initial ArguAna and SciFact relevance labels are historically exposed. The later 973-query validation separates previously unconsumed query inputs and connected debate groups, while retaining the same indexed domain and declaring its canonical-self task rule. All intervals remain conditional on the fixed models and grouping. The prior-centered and source-null controls revisit exposed source and natural populations; their intervals do not include training randomness or research-wide model-search uncertainty. Their source-only rules are mechanistic controls, not new independent benchmark evaluations.

The shallow-correction study reuses the now-exposed 973-query ArguAna population and historical SciFact300 labels; its source folds select residual coefficients conditional on the already fitted base head. NFCorpus adds a prospectively fixed graded task and exposes weaker correction and allocation transfer. The complete objective and budget comparison remains to be replicated across other independently trained model pairs.

The capacity controls match updates and presentations, not all optimization conditions. The pooled MLP's higher training loss makes convergence a material alternative explanation. Ridge is globally solved for a penalized affine objective, not for the listwise KL objective or nonlinear family. Its initial comparison with prefix adaptation changes regularization and solver. The common-readout intervention recovers most of that target gap; its remaining width-400 difference is substantially smaller on the new query population. Head replacement still changes the target and deployment-precision alignment together.

H.4 Offline knowledge and online computation are complementary budgets

The cached-token comparison also separates offline and online budgets. It keeps the source alignment sample fixed but acquires one extra teacher context for every vocabulary ID. Its quality gain is therefore not a same-total-supervision-budget result. The fixed vocabulary and prompt restrict the cached contexts; richer extraction remains untested. The new E5/Android lexical run extends the checked intervention to a second model pairing (Appendix K). A new supporting prior/depth execution independently reproduces the reported C2/C3 means (Appendix L), without recovering the original artifacts. Neither new execution replicates the complete objective and budget study. The completed eight-cell comparison distinguishes offline lexical computation from query-level interaction: under common source alignment, deep online computation improves both tasks, while the shallow effect changes sign across them. Context is not one isolated variable in this construction.

A paid source-CDF policy initially exceeds six neural static alternatives on historical ArguAna and loses on the historical NFCorpus test population. Fixed lexical fusion subsequently raises the ArguAna static frontier enough that paid superiority is no longer established. The subsequent new-query experiment finds useful action information in the same cheap observations. The supervised target and selected execution cost then influence how much of that information becomes a paid advantage. This population change also changes the judgment-grade mixture, so it does not isolate one cause of the earlier failure.

The source-identical lexical intervention makes the information budget particularly concrete. All alignment features and fitted coefficients remain unchanged, yet replacing source-only completions with teacher responses improves both natural tasks. Together with the equal-penalty prior intervention, this establishes two distinct ways in which available knowledge can alter a compression result without increasing its stated alignment sample count. The complete matched-depth comparison then identifies a remaining online increment. Each contribution is measured against its specified alternatives, leaving other compression strategies open.

Improving offline knowledge also changes what an observation can usefully select. We hold the original continuation probability p and native deep action fixed, and replace source-affine estimates of unseen token rows by their actual teacher responses. The source features and fitted head remain identical. If δ is the resulting stopping-quality gain, the change in advantage over equal-rate random continuation is exactly $-\operatorname{Cov}(p, \delta)$, while absolute policy quality changes by $E[(1-p)\delta]$. On historical ArguAna the allocation advantage falls by .00497 [.00093,.00898]; on NFCorpus it rises by .00345 [.00048,.00731]. Absolute expected policy quality improves in both tasks. The NFCorpus final allocation advantage itself remains unresolved, and both changes have wider, zero-crossing intervals under the one realized hash decision. Added offline benefit lands on different parts of the fixed continuation rule. Information substitution is decision-dependent [36]; average endpoint improvement alone does not determine its direction. Here the observation still comes from the complete table, so the weaker-table cross-comparison identifies a changed opportunity rather than a free deployable observation.

H.5 Economic conclusions and measurement boundaries

Serving measurements concern warm singleton requests on one shared CPU environment. Model loading, quantization setup, and index preparation are separate offline costs. The static envelopes use target-selected mixtures as descriptive upper benchmarks. Earlier endpoint intervals condition on the selected mixture; the paid-policy analyses also reselect cost and mixture in each resample. The natural 64-query experiment establishes a same-cost nDCG advantage for fixed ridge endpoints. The later complete973 execution puts the fixed adaptive rule slightly above the 125-static envelope in point estimate, with uncertainty spanning zero. Its group intervals reflect population composition within one hardware session. They do not establish equal performance or hardware-session robustness. NFCorpus gives a negative same-cost gate contrast against its retained eleven-static envelope.

The later cached-first policy has a positive pointwise ArguAna contrast and a negative NFCorpus contrast against its matched six-static subset. Its ArguAna advantage is unresolved against the later nine-static benchmark with classical lexical fusion. Both populations are historical at these stages; neither subset is a complete-bank rerun. The later 2,890-query coefficients are fixed before new model outcomes, while their online implementation follows the offline information result. Its three retained paid-policy intervals jointly span zero, with the same query-composition rather than hardware-session scope.

Appendix I. Complete position and multiplicity factorial

The position switch maps every copied token to the original position whose token ID it copies, including the terminal SEP. Attention and pooling switches subtract the logarithm of that original-position class's multiplicity. Counts are per original position, not per vocabulary ID. With cycled positions, the uncorrected repeated sequence is also equivalent to a short sequence whose key and pool logits receive $+\log m_j$. A separate compact weighted execution on the first two fixed panel

queries reproduces all four corresponding FP32 pooling cells within 1.32×10^{-6}. The all-corrected 16-query outputs match the original short outputs within the same tolerance. PC retains small shape-dependent deviations, with maximum all-corrected output difference .02184.

Table S17. Complete forward relative affine-error factorial, three-seed means. P, A and G denote corrected positions, attention mass and pooling mass. The original short reference has no artificial copies. Reverse directions, absolute residuals, fixed-source leverage and per-query outputs are retained with the experiment.

P / A / G	Dedicated PC	Dedicated FP32	Shared PC	Shared FP32
Original short	.0566	.0437	.0474	.0352
0 / 0 / 0	.1869	.1864	.1512	.1515
0 / 0 / 1	.2039	.2059	.1689	.1716
0 / 1 / 0	.1840	.1807	.1454	.1415
0 / 1 / 1	.1767	.1715	.1352	.1291
1 / 0 / 0	.1343	.1258	.0998	.0910
1 / 0 / 1	.1538	.1479	.1278	.1233
1 / 1 / 0	.1247	.1137	.0931	.0832
1 / 1 / 1	.0565	.0437	.0473	.0352

The paired-input experiment, fixed-hidden prompt-mass experiment and position/multiplicity factorial use 61.42, 22.89 and 51.09 process-CPU seconds for their main forwards. The compact control uses another 4.24 seconds including runtime setup. These are offline mechanism costs. Native repeated and short pooling outputs reproduce the preceding saved values bitwise in all 448 factorial checks; the prompt-mass run reproduces 896 native values. Source maps, means and leverage geometry remain fixed throughout. The initial input run's import-name collision failed before any forward pass; its corrected execution retains the same inputs and protocol. Legacy runtime setup loads existing document vectors incidentally, but none of these diagnostics scores documents or reads relevance labels.

Appendix J. Detailed comparison with prior work

The contribution is a set of controlled information interventions and their retrieval consequences. Prior-centered fitting, common-readout diagnostics, contextual-to-static extraction and adaptive retrieval all have important precedents. The comparisons below locate our new empirical evidence within those established approaches.

EmbedDistill [8] already connects embedding alignment, a frozen teacher document tower, query augmentation and a teacher–student generalization analysis. Its NQ, MS MARCO and BEIR experiments establish precedents for alignment and broader query support. Our nested-support factorial tests the preference between candidate-conditional and complete-index supervision; the prior-centered control then limits how broadly that result can be interpreted. These controls establish how the objective comparison changes with support, prior retention and transfer, and what those changes mean for cost.

Prior-centered regularization is also established. Biased regularized least squares is a classical hypothesis-transfer method [34]; L2-SP [31] explicitly penalizes departure from pretrained parameters in neural fine-tuning. Our closed-form ridge control applies that principle with a fixed affine center and an unpenalized intercept. At the same penalty it changes extrapolation without changing the

estimator's response to its source labels. Wilson et al. [33] analyze different generalization behavior of adaptive methods; Vaswani et al. [32] use data-span projections to alter extrapolation while preserving training predictions. The source-null intervention in Section B.7 follows that principle. Here it quantifies one proposed explanation of the observed retrieval reversal.

Query-side embedding alignment with a frozen document encoder is a close predecessor [1]. Its results emphasize student initialization and show that shallow query encoders can retain much of a larger retriever's quality. NanoVDR extends query-centric alignment to visual-document retrieval and compares cosine alignment against ranking-based objectives across several backbones [4]. Our ridge control applies that established idea to a fixed pooled representation, separating the quality of the teacher target from the difficulty of training the encoder.

Asymmetric query encoding extends to much smaller function classes. Cohen et al. [23] pair efficient Transformer or recurrent query models with stronger passage encoders; their incremental RNN also exploits computation before query submission. Bommasani et al. [27] explicitly reduce contextual models to single-word or context-aggregated static vectors; Gupta and Jaggi [28] train static embeddings from contextual teacher representations. Model2Vec [26] forwards a vocabulary through a sentence transformer and postprocesses the resulting static token vectors. LightRetriever [24] caches instruction-conditioned token vectors for mean-lookup query encoding, combines dense and sparse retrieval, and trains its document model. Our input-table and teacher-table comparators apply these established ideas under a fixed heterogeneous document index and the same source-only ridge target. They measure how the offline representation changes the available budget curve. Batch-one complete-request costs address a different execution regime from large-batch GPU throughput. CQ [29] separates document-independent and document-dependent token contributions in late-interaction compression, including a control that replaces a high-layer static component by raw embeddings. Our depth controls measure how this established extraction route changes the budget comparison in the current fixed-index query system.

SPARTA [39] directly precomputes vocabulary-term/answer scores for lookup and summation, while DeepImpact [40] learns quantized additive passage impacts. VA-file [41] precedes compact conservative filtering followed by original-vector refinement. BM25S [38] likewise moves fixed term contributions into a sparse offline index. The fixed-scorer compilation experiment measures the resulting memory–cost trade-off; its lexical control uses classical RRF [37].

Matryoshka Representation Learning learns nested embedding dimensions for different resource constraints [2]. Two-dimensional Matryoshka Sentence Embeddings adds flexibility across Transformer depth [3]. These methods motivate the budget dimensions in our study. Our joint-width experiment focuses on the distinction between gains over truncation and gains over dedicated small models, while the executed serving experiment charges the observation needed to choose a later computation path.

Kumar et al. [9] show that fine-tuning can improve in-distribution accuracy while underperforming a linear probe under distribution shift, and study feature distortion in two-layer linear networks. Mai et al. [10] show that partial-class fine-tuning can preserve or improve class-discriminative features while its classifier loses accuracy through relative logit-scale bias. Together, these results motivate separating a feature change from its trained readout. Our retrieval comparison adds complete-index targets, crossed query/index evaluation, and executed budget comparisons. Its per-query positive rescaling differs from class-specific calibration: it preserves the exact within-query ordering and tests the scale sensitivity of teacher KL. The present prefix-versus-ridge comparison changes the

readout objective and solver as well as the features, so its ordering reversal alone does not identify feature damage.

BEIR provides heterogeneous retrieval tasks for studying transfer [7]. We use its SciFact, ArguAna and NFCorpus settings with a fixed document encoder and complete-corpus scoring. The core factorial holds the index fixed during training. Its subsequent crossed evaluation keeps the fitted models fixed while changing query population and index, revealing that source rankings of compression procedures need not survive the natural target task. Unlike a broad zero-shot benchmark, this study emphasizes a controlled model comparison and its metric and cost implications.

Query-dependent ranker selection and paid cascades have substantial precedent. Balasubramanian and Allan [13] directly predict the effectiveness difference between two rankers, using features obtained after executing both rankings. FrugalML [14] optimizes sequential prediction services under cost constraints. Score standard deviation is an established query-performance predictor, including query-drift estimation and NQC [17]; our unit-query top-ten statistic is not NQC's corpus-normalized formula. Dense query-performance prediction also finds that query types affect the reliability of score- and coherence-based predictors [15].

Huo et al. [25] directly study optimal reranking depth and report weak performance of many standard query-performance predictors. This is a close precedent for separating predicted initial quality from the value of a later action. Our fixed-ranking decomposition measures the survival of movement prediction after conditional relevance value weakens.

RA-SCE [18] is a close selective-retrieval precedent: it combines dispersion, entropy and score-gap signals with safety checks to choose cross-encoder refinement, including BEIR transfer experiments. Its reported effective cross-encoder latency weights measured reranker cost by the gating fraction. Our experiment measures shared-prefix continuation and both retrieval passes directly, and compares them with a complete static-mixture envelope. This comparison addresses the execution cost of the complete retrieval path relative to static alternatives. The new-group experiment asks whether a fixed shallow observation retains allocation value, and what that value buys after the complete request is executed.

Pretrained tangent features provide another route to adaptation. TAFT linearizes Transformer attention and fits regularized displacement parameters [19]. Our shared correction restricts this operation to the original scalar pooling-attention parameter, yielding an analytic query-side vector update. The centered score objective follows the margin-distillation principle established by Margin-MSE [20]. Candidate-conditioned pooling also precedes this study in Poly-encoders [21]. The present experiments compare correction geometry and executed retrieval cost within a fixed shallow system, including a full pooled-affine control and a restricted quadratic candidate model.

Appendix K. Original E5 record and newly executed C1 replication

K.1 What the original report establishes and leaves unavailable

The original E5/Android manuscript reports a locally timestamped, hash-sealed protocol, not public preregistration. It describes fixing the model, task, folds and lexical contrast before target execution, checking 738 prediction-artifact hashes before label acquisition, and independently reproducing the per-query metrics and bootstrap. Its bounded project-tree and selected-ledger audit reported no prior Android development use; that audit does not certify all project history or pretraining exposure. The completed C1–C4 prediction arrays, training/audit logs and final execution seals

were not recovered. These historical execution and audit statements remain reported rather than authenticated by the current package.

The old C1 means are retained alongside the new run in Table 2 because they agree at displayed precision. The old reported full-minus-affine interval is [.06528,.09580]. The old resource report gives 5,480.97 target-worker CPU seconds and describes a conservatively charged interrupted source process followed by checkpoint recovery. Those costs and events are not silently replaced with the new run's ledger. Appendix L retains the original C2/C3 report alongside a separate new audited reconstruction; Appendix M now supplies the new C4 replication alongside its original reported record. C2/C3 were not recomputed by the C1 execution itself.

K.2 Pinned specification and current execution

The new C1 uses intfloat/e5-small-v2 revision ffb93f3bd4047442299a41ebb6fa998a38507c52, FP32 eager execution, maximum length 512 and the symmetric query: prefix on both Android sides. Document text is title plus one space plus body, or body alone for an empty title; no edge whitespace is stripped. Source fitting reuses the exact original R2048 query bytes and folds. Its selected ridge penalty is 0.1. The support has 4,553 token IDs, and the source features, fit and source deployment predictions are identical across all four lexical arms within this run.

Android corpus and queries come from BeIR/cqadupstack revision e84b7220ce5a4e6c49aeb568983b0ec5a7584c2c. Judgments come from BeIR/cqadupstack-qrels revision 13595a9f221256a9a21e91c8b0fed3563382b6a7, Git blob 421f18e19f4b3bab05d1fd07378bfedc7ba8376f. All 22,998 documents and 699 queries are used, with unique IDs and no cross-role ID overlap. The source run includes 2,560 query vectors, all 30,522 singleton teacher rows and 5,183 SciFact document vectors. Qualification vectors reproduce the three available historical 64-vector hashes. That bounded parity does not establish byte identity for the missing historical full target run.

The isolated CPU runtime uses Python 3.12.14, Torch 2.8.0+cpu, NumPy 2.3.5, Transformers 4.55.4, tokenizers 0.21.4, safetensors 0.6.2 and PyArrow 21.0.0. Public asset identities and hashes are retained. The reconstructed target runner uses document batch 4 and query batch 8, fixed before its outcomes; the original completed target batch schedule is unavailable. The run uses two threads and a 6 GiB address-space limit. Frozen source/prediction/evaluation identities and conservative resource ledgers are retained; no scientific stage failed or required a post-outcome cap amendment.

K.3 Metrics, current seals and independent checks

All 699 queries enter the analysis. nDCG uses gain $2^{\text{grade}} - 1$; Recall and reciprocal rank treat strictly positive grades as relevant. Unjudged retrieved documents receive zero gain. Empty judgments or zero ideal DCG would remain in the denominator with metric zero; missing indexed positives would remain in ideal-DCG and recall denominators. Neither condition occurs here. Identical query/document IDs are excluded before ranking in every arm, although this instance has no such overlap.

The paired bootstrap draws 699 positive-link components with replacement in each of 4,000 replicates, using PCG64 seed 20261023. Canonical question IDs are shared across query/document roles; all components contain one evaluated query here. Each replicate pools its sampled queries, rather than equally weighting unequal component means; percentile endpoints use linear interpolation. The new primary difference is .080335947 and its 95% interval is [.064962238,.096283352]. The old arrays

and draws are absent, so neither a numerical cause nor a Monte Carlo explanation of the old/new interval difference is established.

The source, prediction and evaluation seals respectively have SHA256 prefixes 25066623949c5737, cb0ad6df387be6e7 and 6e25b071bee69d6d; full digests are in the evidence manifest. Before parsing already downloaded judgment bytes, the evaluator reproduces all 3,495 saved-vector rankings and score fingerprints. All 56 inactive queries have exact within-panel outputs across the four lexical completions. This current program order is verified, but does not turn a historically exposed replication into an untouched test.

An independent saved-output audit reproduces all rankings, per-query metrics and 4,000 bootstrap draws. Maximum scalar metric discrepancy is 2.22×10^{-16} and maximum draw discrepancy is 1.11×10^{-16}. A second, nonduplicative check verifies raw-input identities, ordered populations, source-block coverage, teacher/head bindings and resource totals. It does not rerun encoder inference or fit another head. The final continuation receipt records a complete new C1 with 9,808.418 conservatively charged CPU seconds for scientific execution and all audits, within the adopted 20,790-second aggregate ceiling. The new target/evaluation ledger charges 6,163.698 CPU seconds before those extra audits. These are offline execution accounts, not serving latency or directly comparable controlled hardware timings.

Completed evidence is retained locally. Its external archive availability is pending and is not implied by the current manuscript. The new C1 verifies the present execution and its positive lexical contrast; it does not recover the old completion artifacts, authenticate their chronology, or by itself establish the supporting analyses. Separate subsequent executions now verify C2/C3 (Appendix L) and C4 (Appendix M).

Appendix L. Original and newly replicated E5 prior/depth evidence

L.1 Original report and common endpoint means

The original manuscript reports the fixed identity prior and shared-readout depth comparisons below. The old numerical results, intervals, audit tolerances and chronology are preserved as reported. A separate new reconstruction now independently verifies all 20 endpoint means at the displayed precision; its newly computed intervals and provenance are separated in L.2. At 128 source queries and the zero-centered CV penalty held fixed, prior-centered ridge improves Android nDCG by .06783, with the three-comparison Bonferroni family-wise 95% interval [.05084,.08591] (98.333% per contrast) (pointwise 95%: [.05383,.08250]). The separately source-selected warm fit also exceeds unchanged identity by .06721 and intercept-only identity by .02571. At R2048 those descriptive gains are .18633 and .14364. The source-selected fit thus adds quality beyond both the unchanged and recentered shallow encoder. Absolute low-support performance remains substantially below the native deep reference.

With the common source-ridge rule held fixed, deep contextual query features exceed the deep teacher-token lookup by .11228 [.08874,.13600] under the same three-comparison family correction (pointwise 95%: [.09269,.13199]). All eight raw-mean interface cells are reported below: lookup versus contextual evaluation, two versus twelve layers, and native versus ridge readout. Native deep E5 already reaches .46987; its ridge endpoint is .47140, so the deep contextual-feature contrast does not imply a general gain from fitting an extra readout. Shallow contextual ridge reaches .17064, compared with .14560 for its lookup counterpart.

The original report states that these protocols were locally frozen before C1 label acquisition. Their source fits and predictions were produced after the shared Android labels became known, without changing the fixed arms, source folds, grid or metrics. They are preplanned supporting analyses, not a second untouched blinded confirmation. The C2 regularization center is the 384-dimensional identity map on normalized shallow features; C3 uses raw masked means at both depths, with the same source-only ridge family. The common deep-lookup endpoint exactly reuses the original C1 fit and predictions, and native deep reuses the teacher reference.

Supporting endpoint	nDCG@10	Recall@10	MRR@10
C2 n=128: zero CV	0.002039	0.002211	0.002779
C2 n=128: prior at zero-CV penalty	0.069865	0.118255	0.067023
C2 n=128: prior CV	0.073926	0.121318	0.073140
C2 n=128: zero at prior-CV penalty	0.014532	0.027967	0.013150
C2 n=128: unchanged identity	0.006716	0.014235	0.005343
C2 n=128: intercept-only identity	0.048221	0.073135	0.049215
C2 n=2048: zero CV	0.181896	0.264804	0.181010
C2 n=2048: prior at zero-CV penalty	0.193050	0.273447	0.193099
C2 n=2048: prior CV	0.193050	0.273447	0.193099
C2 n=2048: zero at prior-CV penalty	0.181896	0.264804	0.181010
C2 n=2048: unchanged identity	0.006716	0.014235	0.005343
C2 n=2048: intercept-only identity	0.049411	0.074437	0.050148
C3 T2: native	0.000260	0.000210	0.000501
C3 T2: ridge	0.145596	0.217897	0.141210
C3 F2: native	0.006716	0.014235	0.005343
C3 F2: ridge	0.170643	0.253531	0.168130
C3 T12: native	0.003694	0.009424	0.003189
C3 T12: ridge	0.359120	0.461560	0.358614
C3 F12: native	0.469873	0.580445	0.475121
C3 F12: ridge	0.471397	0.581923	0.477100

Table L1. All 20 E5 prior/depth endpoints: each of the 60 displayed metric values agrees between the original report and the newly audited reconstruction. Duplicate high-support endpoints are retained, not deduplicated. T denotes the literal-token lookup and F the contextual full-query representation; 2/12 denote encoder depth. The C2 rows use normalized F2 features, whereas C3 uses raw means before its stated readout, so similarly named shallow fits need not coincide. No target outcome selects a penalty, endpoint or reported seed.

The reported paired positive-link bootstrap and metric conventions match the original C1 specification described in Appendix K; this is not a claim of order or draw identity with the new K.3 execution. These two supporting contrasts and the consumer contrast in Appendix M retain the original Bonferroni factor of three. The original study reports unchanged sealed hashes for all 20 fits and 40 prediction files, and independent reproduction of 20 by 699 per-query metric triples within 3.33×10^{-16} and both 4,000-draw bootstrap arrays within 4.16×10^{-17}. Those completed files and audit outputs are not in the recovered package. A metadata-only tuple/list serialization repair occurred before C2/C3 prediction generation; it did not change numerical operators, endpoint order, source fits or the original compute cap. The original study describes accompanying failure and continuation records; their completed execution logs remain unrecovered.

L.2 New reconstruction: selection, intervals and independent checks

The new execution reuses the sealed C1 source and document arrays read-only. C2 fits normalized shallow contextual features to the normalized native deep teacher, using source prefixes of 128 and 2,048 queries, the original fold IDs restricted to each prefix, the eight-penalty grid, fold-local trace scaling divided by 384, an unpenalized intercept and FP64 SVD. The identity prior is retained pretrained information, not newly learned supervision. All six endpoints per prefix are preserved: zero-centered CV, prior at the zero-CV penalty, prior-centered CV, zero at the prior-CV penalty, unchanged identity and intercept-only identity. C3 crosses raw masked means at two/twelve layers, literal-token versus contextual input, and native normalization versus independently source-selected ridge. No target metric selects any fit or endpoint. Deployment uses FP32 outputs.

Only the missing 699 shallow target-query means require new full encoder forwards, at batch eight. The simultaneously extracted deep means match C1 byte-for-byte, with identical token IDs and lengths. Documents, source data and vocabulary rows are not re-encoded. This is a model/provenance compatibility check, not an independent second forward pass confirming the shallow features. All 20 predictions are sealed before the current metric evaluation; historical target exposure still makes this a reconstructed replication, not a blind confirmation.

The new C2 equal-penalty contrast at 128 queries is +.06782575, with pointwise 95% interval [.05339512,.08306296] and family-corrected interval [.05022360,.08656369]. The new C3 contextual-minus-table deep ridge contrast is +.11227759, with pointwise 95% interval [.09222520,.13254614] and family-corrected interval [.08786663,.13642882]. Both use 4,000 PCG64 draws with seed 20261023, the exact C1 component order and draw-index digest, and the original three-comparison family including the C4 contrast now completed in Appendix M. Each family-adjusted interval therefore uses a nominal 98.333333% confidence level under the same percentile convention (nominal 95% for the three-comparison family); the family was retained at three throughout the sequential replications. These intervals condition on the fixed fitted models and task; they do not quantify model-training, seed-selection or research-selection uncertainty.

The original intervals in L.1 remain part of the record. Their difference from these new intervals has no established explanation because the old arrays and draw order remain missing; we do not assume that it is Monte Carlo variation. No reseeding, query reordering or outcome-matching stopping rule was used. At low support the retained identity helps substantially relative to zero centering, but absolute retrieval quality remains low. The deep contextual gain is not evidence that ridge generally improves native deep retrieval: native deep already reaches .469873, compared with .471397 under ridge.

The independent saved-output audit reconstructs all eight source CV families, parameters and held-out coordinate-fidelity arrays without importing the producer or rerunning the model. It checks 12,582 newly ranked query/endpoints and exact reuse of 1,398 C1 rankings. All 13,980 per-query metric triples agree within 2.23×10^{-16}, and bootstrap values within 1.12×10^{-16}. All 60 Table L1 values agree with the original six-decimal report. The audit result and final receipt passed on 27 October 2026 at 17:59:51 and 18:00:32 UTC, respectively.

The four scientific stages consume 473.85 measured CPU seconds and retain 254,393,848 output bytes. The independent audit uses 117.21 actual child CPU seconds, conservatively charged as 127.21; together with the producer supervisor reserve, cumulative accounting is 671.06 CPU seconds against the prospectively adopted 1,380-second ceiling. Before any new fit or target outcome, the prediction cap was increased from 300 to 600 seconds after counting 18 newly ranked endpoints, and

the audit allowance from 120 to 200 seconds. The prediction stage actually uses 436.63 CPU seconds. Initial freezes and the pre-outcome amendment remain retained; no cap change followed outcomes. These are local execution costs, not service-latency comparisons or a claim of historical-run cost parity.

Current source, feature, prediction and evaluation seals and the independent final receipt are retained locally. This audit authenticates the new C2/C3 evidence, not the old execution or chronology. The completed C4 full-training replication is reported in Appendix M. The stronger E5 completion panel is now separately complete (Appendix O); the response-budget replay remains Leaf-only. External archival publication of the new evidence remains pending.

Appendix M. E5 adapted features under a common consumer

M.1 Complete three-seed replication

The new C4 execution confirms that a trained consumer can conceal useful adapted features. All three fixed seeds complete 500 updates. Replacing the trained KL head with source-only ridge increases mean Android nDCG from .1191981 to .1806726: +.06147449 with nominal three-comparison family-adjusted interval [.04863923,.07538613] (pointwise 95% [.05071259,.07326414]). The mean ridge-minus-frozen-prefix difference is +.01002955, with positive per-seed differences .00904001, .00959536 and .01145328. Thus common ridge recovers more than the unchanged shallow representation's performance. The native deep reference remains stronger, as shown in Table 2.

Consumer endpoint	nDCG@10	Recall@10	MRR@10
20261023: trained KL head	0.124924	0.195782	0.119414
20261023: common ridge	0.179683	0.267638	0.175916
20261024: trained KL head	0.114878	0.192072	0.104857
20261024: common ridge	0.180238	0.265968	0.178115
20261025: trained KL head	0.117792	0.188543	0.112010
20261025: common ridge	0.182096	0.270713	0.179245
Three-seed mean: trained KL head	0.119198	0.192132	0.112094
Three-seed mean: common ridge	0.180673	0.268106	0.177759
Frozen prefix: common ridge	0.170643	0.253531	0.168130

Table M1. Newly executed C4 results for every fixed seed and both adapted consumers. All seeds finish exactly 500 updates; neither target quality nor source loss selects a seed or checkpoint. The frozen comparator exactly reuses C3 F2 ridge. The seed mean is computed per query before resampling 699 components, not by treating three seeds as 2,097 independent queries.

The update rule retains the original design: FP32 eager CPU, frozen embeddings and blocks 2–11, trainable blocks 0/1 plus the affine head, source R2048 and the full 5,183-document SciFact teacher index. AdamW uses block/head learning rates $2 \times 10^{-5}/10^{-3}$, weight decay 10^{-4}, betas (.9,.999), epsilon 10^{-8}, batch 32 and global gradient clipping at one. Dropout and scheduling are disabled. The objective is literal batch-mean KL(teacher‖student) at temperature .05, without a squared-temperature prefactor. This execution explicitly fixes PCG64 epoch permutations and all-parameter group decay; the old run's exact permutation and decay grouping were not recovered.

Seed 20261023 resumes the ten-update qualification checkpoint with unchanged optimizer moments, trainable tensors, CPU/NumPy RNG states, epoch permutations and next-batch position, adding

490 updates. The other seeds each execute 500. No restart, loss-selected stopping or target-based checkpoint selection occurs. At update 500, source and target features are extracted from each fixed prefix; source-only common ridge uses the original R2048 folds, eight penalties and FP64 SVD before FP32 deployment. The six new endpoints share the complete 22,998-document target index. This consumer intervention changes the complete fitting rule, rather than isolating the training loss as its sole cause.

M.2 Current verification and statistical record

The independent saved-output audit checks all three final step-500 counters, the exact step-10 continuation, trainable/frozen-state records and source/teacher/index bindings. It independently recomputes three source-CV families and the selected consumers, verifies 4,194 new rankings and exact reuse of 699 frozen-comparator rankings, and reproduces all metric arrays and the seed-mean bootstrap. It does not repeat training or feature forwards, nor independently reconstruct every retired intermediate tensor. These checks establish current trained-state provenance and saved-output correctness without claiming a second training trajectory.

The bootstrap uses the same 699 positive-link components, 4,000 PCG64 draws, seed 20261023 and sample-index digest as C1–C3. The nominal family factor remains three: each adjusted interval uses a 98.333333% confidence level for the nominal 95% supporting family. Uncertainty conditions on the fixed three trained seeds and query population; the seed mean is not an estimate of arbitrary training-seed variation. All three seed outcomes are retained. The final independent receipt passes on 27 October 2026 at 18:36:55 UTC.

Scientific execution consumes 985.544254 CPU seconds; the audit uses 45.125178 actual child CPU seconds, charged as 55.125178. Including the 130-second continuation reserve, conservative charging is 1,170.669432 against the adopted 2,400-second ceiling. The prior qualification is charged separately as 30.152468, giving 1,200.821900 all-inclusive charged CPU seconds. Retained full-run outputs occupy 335,971,481 bytes; continuation checkpoint writes total 2,674,561,180 bytes, or 2,719,132,532 including qualification. These are offline execution accounts, not serving latency.

The new replication agrees with the original qualitative consumer finding but not its exact values. New KL/ridge means are .119198/.180673, compared with the originally reported .113642/.181331. The difference's cause is not established and is not attributed to a guessed implementation detail. The complete old panel and its reported provenance are retained once below; the current independently verified panel supplies the main scientific estimates.

M.3 Original reported panel and provenance

The original E5 analysis reports useful adapted features that the trained KL head does not fully exploit. The results, training checks and resource account below retain that report; the completed checkpoints, predictions and audit logs have not been recovered or independently recomputed here. Across all three fixed 500-update seeds, replacing that head with source-only ridge improves Android nDCG from .11364 to .18133: +.06769 [.05352,.08282] under the three-comparison Bonferroni family-wise 95% interval (98.333% per contrast; pointwise 95%: [.05608,.07977]). Common ridge also descriptively exceeds the frozen-prefix ridge baseline of .17064 by .01069; each seed has a positive difference (.01083,.01144,.00980). The gain thus extends beyond recovering the unchanged shallow representation's performance. The frozen native deep reference remains much stronger (Table 2).

Consumer endpoint	nDCG@10	Recall@10	MRR@10
20261023: trained KL head	0.113013	0.184408	0.105586
20261023: common ridge	0.181470	0.268567	0.178876
20261024: trained KL head	0.108237	0.182450	0.098037
20261024: common ridge	0.182084	0.269187	0.179914
20261025: trained KL head	0.119676	0.193440	0.111840
20261025: common ridge	0.180440	0.266099	0.178672
Three-seed mean: trained KL head	0.113642	0.186766	0.105154
Three-seed mean: common ridge	0.181331	0.267951	0.179154
Frozen prefix: common ridge	0.170643	0.253531	0.168130

Table M2. Reported results for all fixed C4 seeds and both adapted consumers, with their unchanged frozen-prefix comparator; the completed execution artifacts remain unrecovered. Seeds 20261023/24/25 each complete 500 updates; neither target quality nor source loss selects a seed or checkpoint. The baseline exactly reuses C3 F2 ridge. The mean is taken across the three seed-specific per-query metrics before query-component resampling, not by treating three seeds as 2,097 independent queries.

Each run initializes its affine head from the same frozen-F2 source ridge fit and adapts encoder blocks 0/1 plus that head. Embeddings and blocks 2–11 remain frozen, as do the native 12-layer teacher queries and complete 5,183-document SciFact index. AdamW uses block/head learning rates $2 \times 10^{-5}/10^{-3}$, weight decay 10^{-4}, betas (.9,.999), epsilon 10^{-8}, batch 32, gradient clipping 1, and no dropout or scheduler. Its full-index objective is literal batch-mean KL(teacher‖student) at temperature .05, without an additional squared-temperature prefactor. After update 500, adapted features are frozen and the common ridge head is fitted using only R2048, the original source folds and penalty grid. Query inference uses the same raw masked-mean interface and fixed document index as C3.

The original report states that the protocol fixed the two-layer adaptation, 500 updates, three seeds, complete-index KL and common-readout comparison before C1 labels. The explicit exclusion of embeddings from trainable parameters, installed AdamW implementation defaults and literal no-squared-temperature KL convention were documented operational clarifications after C1 unblinding; they were not retroactively present in the initial freeze. The report describes exact resumption of the qualification's first ten seed 20261023 updates, including optimizer, RNG and permutation state, an unchanged numerical update body, and verification of 500 optimizer steps in all three final checkpoints. These are preplanned supporting analyses with disclosed implementation chronology, not untouched blinded replications.

The original reported paired bootstrap uses 699 positive-link components, 4,000 PCG64 replicates and linear percentile interpolation, following the original C1 conventions described in Appendix K. This does not establish component-order or draw identity with the new K.3 execution. Its interval conditions on the fixed three-seed fits, model and task; it does not estimate variation across arbitrary training seeds. The original study reports independent reproduction of all per-query metric triples within 1.11×10^{-16} and bootstrap draws within 2.78×10^{-17}. It also reports finite gradients, nonzero aggregate block/head gradients and changed trainable-parameter hashes at every training step, with unchanged frozen-parameter and teacher/index hashes. These are original audit claims, not checks repeated on recovered checkpoints here. The readout intervention changes fitting objective and solver together, so it identifies a useful complete consumer replacement rather than a pure loss-function effect.

The original resource account reports 734.79 CPU seconds for the three training workers; all continuation workers consume 972.86, plus 2.23 for independent audit. The 1,113.86 allocation charge includes reserves, not measured computation alone; qualification is separate. Latest checkpoints write 66.27 GB cumulatively but retain 133.56 MB. Neither cumulative writes nor offline experiment time is serving cost.

Appendix N. Source-only completion and fixed-head response budgets

N.1 Candidate selection and retained information

These new Leaf experiments retain the original model revision, vocabulary, fitting support, R2048 head, ordered query populations and complete indices. The raw [30,522,384] embedding tensor is extracted from the original official model artifact after verifying its complete 90,272,656-byte SHA256 against the historical identity. No new model inference supplies the completion search or budget replay.

The five type folds sort source IDs by a fixed SHA256 rule and assign sorted positions modulo five. Every fold has 911 validation and 3,644 training types. Prediction error is averaged equally over the 4,555 out-of-fold types and 384 raw teacher coordinates. Source E512 retrieval is diagnostic, not a selection criterion. Validation teacher rows are withheld from the learner; outside-support teacher rows and natural-task outcomes cannot choose a model. Final fitted tables restore all $H[S]$ rows only after selection, not during token-type validation.

Affine/ridge uses a training-mean intercept and FP64 SVD. Its coefficient grid is

$$\{0, 10^{-6}, 10^{-4}, 10^{-2}, 1, 100\}.$$

The penalty is the coefficient times $\mathrm{tr}(X_c^\top X_c)/384$; coefficient zero uses the historical numerical SVD cutoff. Nyström RBF kernel ridge normalizes each raw input row, chooses 512 anchors by a fixed hash among training types, and uses bandwidth multipliers $\{.5, 1, 2\}$ times the square root of the median positive anchor-pair squared distance. It retains anchor Gram eigenvalues above 10^{-12} times the largest and fits a centered ridge readout with coefficients $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 100\}$. Its penalty is $\lambda = \alpha\,\mathrm{tr}(\Phi_c^\top \Phi_c)/r$, where r is the retained feature dimension. Means, anchors, bandwidth estimates and decompositions are fit on training types only.

Neighbor models use raw-embedding cosine with exact ties resolved by token ID. The fixed neighbor counts are $\{1, 4, 8, 16, 32\}$ and the weights are uniform or inverse cosine distance with denominator at least 10^{-6}. The duplicate one-neighbor rule is removed, leaving nine configurations. Altogether there are 30 candidates and 150 candidate-fold scores, not 30 target searches. Each family winner and the single cross-family winner minimize source OOF MSE, with fixed family/grid tie rules. All candidates and scores are retained in the source candidate ledger.

Selected family	Fixed selected parameters	Source OOF MSE
Affine/ridge	coefficient .01	.000047214892
Nyström RBF	512 anchors, bandwidth 1, coefficient 10^{-6}	.000045296582
Weighted neighbors	8 neighbors, inverse-distance weights	.000067086887

Table N1. Source-selected members of the three completion families. The newly evaluated unregularized affine and uniform eight-neighbor candidates have MSE .000047439588 and .000069374176. These type-validation scores are new evidence, not historical retrieval-validation values.

The new table pool accumulates FP32 rows in FP64, casts once to FP32, and uses the unchanged FP32 affine/scoring lane. Historical affine and eight-neighbor compiled-table hashes reproduce exactly, including all 25,967 saved neighbor-ID rows. All six new arms have exact R2048 features and inactive outputs within that lane. The original alignment CV/head files remain unchanged. All 1,808 full-baseline ordered top-ten rows match, but newly pooled natural-query fingerprints match 0/1,296 historical fingerprints; replaying saved original vectors matches 1,296/1,296. The latter must not be presented as byte parity of newly pooled vectors. There were no historical full-score arrays to test for archived score-array identity. Full, historical-affine and historical-eight-neighbor per-query metric triples reproduce exactly on both natural populations.

The two primary full-minus-champion intervals use 4,000 paired PCG64 replicates, seed 20261019 reset for each task, the original 134 ArguAna connected groups with query-weighted pooled means, and the 323 NFCorpus query units. Their individual coverage is 97.5% for a nominal 95% two-task Bonferroni family. Other completion contrasts are descriptive. The uncertainty is conditional on the fixed source selection; it does not integrate candidate-selection or arbitrary training-seed variation.

N.2 Every fixed response-budget endpoint

The original 4,555 source teacher rows and 2,048 full-query alignment targets remain common resources. An additional budget counts distinct revealed rows outside S. The source-error proxy is the average champion OOF squared row error of the eight nearest source types in raw-embedding cosine; unseen types are sorted by decreasing proxy and then token ID. The three random orders are separate PCG64 permutations at the seeds below. Acquisition uses no outside-support teacher error or target token frequency. All orders are sealed before budget predictions; the same champion and head are used throughout.

Acquisition order	Additional rows	ArguAna nDCG@10	NFCorpus nDCG@10
Shared zero	0	0.456580	0.290709
Source-error proxy	256	0.456900	0.290709
Source-error proxy	1,024	0.458908	0.290807
Source-error proxy	4,096	0.470762	0.292109
Source-error proxy	16,384	0.485380	0.296674
Random 20261027	256	0.457443	0.291816
Random 20261027	1,024	0.458056	0.292963
Random 20261027	4,096	0.463579	0.297778
Random 20261027	16,384	0.489336	0.305889
Random 20261028	256	0.458054	0.290843
Random 20261028	1,024	0.462662	0.290469
Random 20261028	4,096	0.466343	0.291209
Random 20261028	16,384	0.492927	0.299416

Acquisition order	Additional rows	ArguAna nDCG@10	NFCorpus nDCG@10
Random 20261029	256	0.457341	0.290638
Random 20261029	1,024	0.458700	0.291668
Random 20261029	4,096	0.461947	0.292716
Random 20261029	16,384	0.486789	0.301213
Shared full	25,967	0.508233	0.304053

Table N2. All 18 distinct budget endpoints, including every fixed random seed. Zero and full endpoints are shared, not independent seed replications. Lines joining budgets in the figure are visual guides rather than a fitted monotone law. The original full table has 30,522 exact rows.

Additional rows	ArguAna proxy minus random mean	NFCorpus proxy minus random mean
256	-0.000713 [-0.002354, 0.000986]	-0.000390 [-0.000917, 0.000084]
1,024	-0.000898 [-0.004552, 0.003060]	-0.000893 [-0.001765, -0.000027]
4,096	0.006806 [-0.000183, 0.013499]	-0.001792 [-0.005208, 0.001605]
16,384	-0.004304 [-0.012423, 0.003707]	-0.005499 [-0.009902, -0.001045]

Table N3. Paired descriptive 95% intervals conditional on the three fixed random orders, using the same query/group resampling units as N.1. These are not simultaneous tests and do not quantify general variation over acquisition seeds. No row or seed is promoted into a new acquisition policy after inspecting these results.

The NFCorpus proxy covers 95.79% of ordered token occurrences at budget 16,384, compared with 96.70% for the fixed random mean. It fully covers 172 queries, whereas the three random orders cover 211, 187 and 193. These post-seal descriptors include formatting tokens. Lower coverage is associated with weaker quality here; this is not a causal explanation or a target-frequency acquisition rule. Every fully covered query matches the full-table vector, score fingerprint and ordered top ten exactly. At zero budget, this includes 77 NFCorpus and 93 source-E512 queries. ArguAna has no inactive queries, so it supplies no empirical zero-budget inactive control.

All 18 endpoints retain byte-identical R2048 features. The zero and full endpoints reproduce all 1,808 first-panel metric triples exactly. Every row replacement reduces or preserves that row's squared error relative to the actual teacher table, but this does not impose monotonicity on a ranking metric downstream of the fixed head. NFCorpus seed 20261028 also declines from .290843 at budget 256 to .290469 at 1,024; seed 20261029 declines from .290709 at zero to .290638 at 256. These outcomes remain in the reported curves.

N.3 Audit coverage and offline cost

The first-panel independent saved-output audit checks all 150 candidate-fold scores and source choices, selected OOF errors, recursive seals, compiled-table hashes, fitting/inactive invariance, and all six arms' metrics and bootstrap. It passes 40,115 assertions. Independently implemented scalar metrics differ by at most 4.44×10^{-16} and alternate query-weighted bootstrap arithmetic by at most 2.89×10^{-15}. The auditor does not refit source learners, retokenize, rerun encoder inference or regenerate full score matrices. Source reconstruction and score-replay computations are verified through frozen code and receipts rather than independently repeated.

The budget audit passes 57,763 assertions, including all four acquisition orders, nested prefixes, all 18 reconstructed table hashes, common endpoint identities, coverage/inactive controls, metrics

and all 324 descriptive interval records. It recomputes the proxy from the fixed source OOF errors and verifies the three PCG64 permutations. It does not fit a new completion or evaluate another acquisition strategy.

The first panel uses 68.106 process CPU seconds and peaks at 791.5 MiB RSS; its independent audit adds 1.425 CPU seconds. The budget replay uses 61.210 process CPU seconds and peaks at 515.9 MiB; its independent audit adds approximately 3.21 CPU seconds. These are new offline processing costs. Cached teacher responses are prior information, not free historical acquisition, and no serving latency or quality/cost frontier follows from these experiments. The two studies remain historically exposed fixed-setting comparisons, with negative and positive outcomes preserved.

Appendix O. Reproducibility record and E5 stronger completion

O.1 Exposure and scope of the current evidence

Exposure is experiment-specific. ArguAna973 excludes overlapping connected debate groups and effective inputs from the earlier official population, while retaining an indexed domain and a declared canonical-self rule. It is historical for the final lexical and depth interventions. NFCorpus323 was prospective for the initial correction/policy comparison, but historical for those later interventions. The original E5/Android report describes fixing the model, task, source folds and primary lexical contrast before target execution, and sealing predictions before judgments. It describes the prior, depth and consumer protocols as designed before that unblinding but executed afterward; they are supporting analyses, not additional untouched confirmations. Appendix K separates that reported chronology from the new historically exposed C1 replication. Appendix L now supplies a separate, independently audited C2/C3 replication; Appendix M supplies the complete new C4 replication and preserves its original reported record separately. Appendix A gives the Leaf exposure record. The new Leaf completion search and teacher-budget replay are also retrospective; their source-only selection and prediction seals do not make the previously seen tasks blind.

The broader E5 completion search added in this revision does not alter those exposure classifications. Current evidence cutoff is 27 October 2026, 18:36:55 UTC. New C1, C2/C3, C4 and stronger-completion outputs have independently audited receipts. All three C4 seeds complete 500 updates; current results replace the historical-only consumer evidence in the main text.

O.2 Numerical replay and original records

The new execution uses FP64 accumulation of FP32 table rows followed by one FP32 cast, rather than a byte-identical replay of historical Torch pooling. Within-panel source features and all 77 inactive NFCorpus queries are exact across the six arms. All 1,808 original full-baseline top-ten rows reproduce, and the original full/affine/eight-neighbor metric arrays agree exactly. Newly pooled natural vectors do not reproduce the old fingerprints; replaying the saved original vectors does. Appendix N gives this numerical boundary and the source-selection details.

The original E5 C1 interval [.06528,.09580] and new interval [.06496,.09628] differ although means agree at reported precision. Old prediction arrays and draw order remain missing, so the cause is not established. The current runner checks all 3,495 saved-vector rankings before parsing judgments; judgment bytes had already been acquired and the outcomes were historically exposed. Appendices K–L retain full original/new chronology, C1 and C2/C3 intervals, current seals and accounting. Recovery of old artifacts is distinct from executing and verifying new experiments.

O.3 E5 stronger-completion design and independent checks

The panel transports the same finite 30-configuration family used in the Leaf experiment: six affine/ridge, fifteen Nyström and nine weighted-neighbor candidates. All 150 candidate-fold outcomes are retained. Five validation folds contain 911, 911, 911, 910 and 910 types; aggregate source loss weights all 4,553 types equally. Only H[S] supplies teacher labels during selection; final maps are sealed before the compiler opens the complete teacher table. Source-selected affine alpha .01 has OOF MSE .0160560523357; selected Nyström bandwidth .5 and alpha 10^{-6} has .0149852472340; weighted eight-neighbor power one has .0155594241327. The unregularized affine candidate has .0161501705152. The champion is selected on source loss and remains primary even though the historical affine endpoint has higher Android relevance.

All six compiled tables restore the 4,553 source teacher rows and preserve every R2048 pooled feature and raw/normalized prediction bitwise under the C1 head. The numerical lane retains C1 Torch FP32 pooling and its 384-dimensional consumer, rather than the Leaf FP64 accumulation/padded-width lane. The three historical completions reproduce C1 vectors bitwise and reuse 2,097 already-audited C1 rankings, scores and fingerprints. The three selected completions receive 2,097 new full-index rankings. All 56 inactive queries match full across vector, score, top-ten and metric channels.

The primary full-minus-champion contrast is .087932614150 with nominal conditional 95% percentile interval [.072636197347,.104946910513]. Four thousand PCG64 replicates retain the C1 component order, seed 20261023 and sample-index digest. The three historical metric and bootstrap arrays are byte-identical to C1. This one-task extension does not change the original C2/C3/C4 three-comparison family. The intervals condition on the selected models and query population, not arbitrary selection, training or acquisition-seed variation.

The independent audit checks all 30 source aggregates and five-fold diagnostics, independently recomputes the 15 selected-family OOF fits and three final maps, and verifies all 2,097 new rankings. It does not independently refit all 150 candidate-fold models. Maximum selected OOF discrepancy is 5.00×10^{-16}; maximum metric and bootstrap discrepancies are 2.23×10^{-16} and 1.12×10^{-16}. The final receipt passes on 27 October 2026 at 18:14:05 UTC. Five scientific stages use 142.634962 child CPU seconds; audit child CPU is 62.583005, conservatively charged as 67.583005. Including the 30-second prior qualification/supervisor reserve, total conservative charging is 240.217967 against the adopted 600-second cap. Retained scientific outputs occupy 298,032,907 bytes. No new teacher inference, target acquisition or external publication is performed by this extension.