GEOMETRY-GUIDED LAYER SELECTION FOR LANGUAGE MODEL COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Which transformer blocks should be considered for removal? We study a layer-selection criterion that compares centered activation subspaces through projector distance. An analytical construction distinguishes this population statistic from tokenwise cosine, and a rank–angle decomposition makes its selections interpretable. On DeepSeek-R1-Distill-Qwen-14B, we compare geometry, Block Influence, and positional references using matched calibration and separate single-block interventions on 144 metallurgy questions. The geometry-selected set has a mean reference-token loss change of -0.1483 nats, compared with -0.1439 for Block Influence and positive changes for the early and late positional references. The geometry–Block Influence contrast is unresolved by the workload-bootstrap interval. All 15 geometric choices are reproduced by angular changes between retained rank-one subspaces at three calibration seeds. These results characterize an interpretable selection signal and its conditional-loss profile; they do not estimate joint deletion effects or free-generation accuracy. A separate historical compression-and-adaptation study provides resource context: a 33-block geometric configuration uses 28.0% fewer parameters and achieves $1.35\times$ the observed token throughput of the 48-block original. We distinguish candidate selection from validation of the resulting compact model.

1 INTRODUCTION

Removing transformer blocks reduces model depth and parameter count, but choosing which blocks to remove requires a diagnostic of their contribution. Representation similarity offers an inexpensive surrogate for directly evaluating every candidate compressed model. The relevant question is what such a surrogate measures and how its selections relate to observed effects of deletion. We study this question through centered activation geometry and matched single-block interventions. Metallurgy provides the domain workload; the selection rule itself does not depend on that domain.

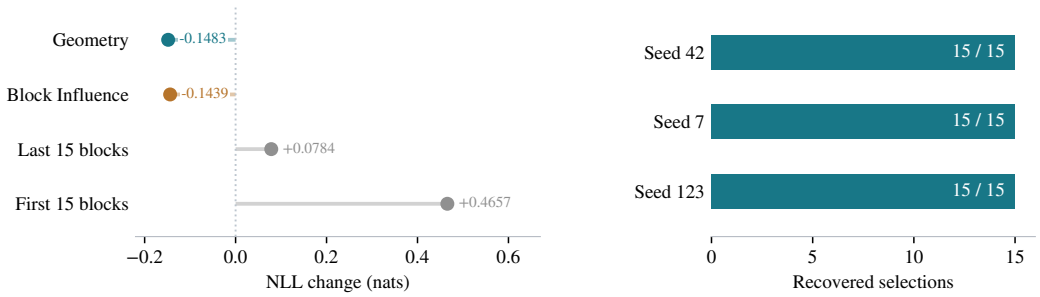
Our criterion compares principal subspaces of activation populations before and after a block. It captures changes in directions of variation across calibration tokens. This differs from the paired tokenwise cosine used by Block Influence (BI) (Men et al., 2024): two populations may share a large mean and have almost parallel token pairs, yet vary in orthogonal directions after centering. Conversely, a mean shift can change tokenwise cosine without changing centered subspaces. Projector distance therefore provides a distinct, inspectable selection statistic.

Code and numerical results are available in an anonymous repository.[1]

We test the statistic on the pinned 48-block DeepSeek-R1-Distill-Qwen-14B model (DeepSeek-AI et al., 2025). Geometry and BI share calibration tokens and are evaluated against the same matrix of 48 separate block deletions on 144 questions. The primary endpoint is mean reference-token negative log likelihood (NLL), with identical response conditioning across interventions. Both selectors have lower selected-set means than early and late positional references. Their aggregate contrast is small and its interval includes zero. They nevertheless select different blocks and yield different composition- and strength-recall profiles.

[1] https://anonymous.4open.science/r/geometry-layer-selection-artifact-B940/

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



(a) Individual-deletion loss

(b) Angular-only selection

Figure 1: Matched layer selection. (a) Mean NLL changes from separate single-block deletions, averaged over each 15-block set and 144 questions. Both activation-based selectors have lower means than the positional references; their mutual contrast has a 95% workload-bootstrap interval spanning zero. (b) Angular-only sorting recovers all 15 geometric choices at each calibration seed; every selected transition has retained rank one at both endpoints. These measurements do not represent simultaneous removal of 15 blocks or task accuracy.

The geometric selections admit a concrete explanation. At three calibration seeds, all 15 selected transitions connect retained rank-one subspaces, and sorting by the angular component alone recovers the full selection. This finding connects the rule to the observed choices and identifies which component matters in this setting. Figure 1 summarizes the intervention result and the selection mechanism.

Contributions. We provide (1) a specified activation-subspace selection criterion, with an analytical distinction from tokenwise cosine and an interpretation through classical projector geometry; (2) a matched individual-block intervention comparison against BI and positional references; and (3) a component analysis explaining the selected set and its stability across three calibration seeds. A separate historical study supplies resource measurements for compact models constructed by pruning and domain continued pretraining (CPT).

Our central claim concerns candidate selection under a conditional-loss diagnostic. Individual deletion effects need not add under joint pruning. We therefore evaluate the selection signal separately from historical compact checkpoints and their resource and answer-quality evidence.

2 Related Work

ShortGPT introduces BI as a representation-based layer-redundancy signal (Men et al., 2024); SLEB studies eliminating transformer blocks to accelerate inference (Song et al., 2024). Work examining cosine similarity and layer relevance motivates checking representation surrogates against actual deletions (Hinostroza et al., 2026). We recompute BI on the same raw block activations and calibration positions as geometry, then compare both against shared intervention measurements. The historical adapted BI checkpoint is reported separately and is not a reproduction of every ShortGPT training choice. Buchstaber's work on low-dimensional subspace representations in time-series analysis motivates the population-subspace viewpoint (Buchstaber, 1994). Our method uses centered transformer activations, without the complete Hankel reconstruction procedure or its guarantees. Projector and principal-angle identities are established mathematics. The contribution here is the specified selector, its empirical intervention profile, and an explanation of the components that determine its choices. Lightweight replacement networks (Chen et al., 2025), adaptation data selection (He et al., 2026), and low-rank updates (Hu et al., 2021) address complementary parts of compression. The historical compact configurations use LoRA CPT on retained blocks and no replacement module. Representation reconstruction need not ensure downstream quality (Shin et al., 2024), and pruning and healing can affect autoregressive loss and question-answering accuracy differently (Gromov et al., 2025). We therefore distinguish selection diagnostics, resource measurements, and answer evaluation rather than treating any one as a substitute for the others.

3 ACTIVATION-SUBSPACE SELECTION

3.1 SCORE AND SELECTION RULE

Let a transformer contain L blocks with $h_{l+1} = F_l(h_l)$. Stack calibration vectors at stage l in $X_l \in \mathbb{R}^{n\times d}$ and subtract their column mean to obtain $\widetilde{X}_l$. Let $U_l \in \mathbb{R}^{d\times r_l}$ contain orthonormal approximate principal directions retained by an energy rule, and $P_l = U_l U_l^\top$. We score block l by

$$s_l = \|P_l - P_{l+1}\|_F = \sqrt{r_l + r_{l+1} - 2\|U_l^\top U_{l+1}\|_F^2}. \tag{1}$$

The second form follows from projector idempotence and the trace identity. It avoids storing $d \times d$ matrices and accommodates unequal ranks. Given budget k, we select the k smallest scores, breaking ties by block index. Removing those blocks preserves the order of retained blocks; evaluating that joint deletion is a further experiment, distinct from the single-block assessment used here.

The score is invariant to the choice of orthonormal basis within each subspace and to a common orthogonal change of coordinates. Centering removes mean shifts, and projection discards amplitudes within retained directions. Thus, s_l describes population geometry rather than functional equivalence of a block to the identity.

3.2 RANK AND ANGULAR CONTRIBUTIONS

Let $m = \min(r_l, r_{l+1})$ and let $\theta_1, \ldots, \theta_m$ be the principal angles, whose cosines are the singular values of $U_l^\top U_{l+1}$. Equation 1 gives

$$s_l^2 = \underbrace{|r_l - r_{l+1}|}_{a_l\text{: rank change}} + \underbrace{2\sum_{j=1}^{m} \sin^2\theta_j}_{b_l\text{: angular change}} . \tag{2}$$

Nested subspaces have zero angular contribution; equal-rank subspaces have zero rank-change contribution. For rank-one unit vectors u, v, $s_l = \sqrt{2 - 2(u^\top v)^2}$. The decomposition makes the source of a small score explicit. It is not rank-normalized, and its classical identity alone supplies no guarantee about deletion loss.

Equation 2 yields a useful structural interpretation before any loss measurement. Since the rank term is a nonnegative integer, $s_l < 1$ implies equal retained ranks. In the rank-one case, the principal angle $\theta_l \in [0, \pi/2]$ satisfies $s_l = \sqrt{2}\sin\theta_l$: ordering scores is exactly ordering changes in the dominant centered direction, with the sign of the singular vector irrelevant. Thus a small score can be explained in geometric terms, rather than only by its position in an empirical ranking. This statement concerns the estimated subspaces; discarded directions and activation magnitudes remain outside the statistic.

3.3 A WORKED DISTINCTION FROM TOKENWISE COSINE

With equiprobable $z \in \{-1, 1\}$ and orthonormal e_1, e_2, e_3, consider $x = Me_1 + ze_2$ and $y = Me_1 + ze_3$. Every paired cosine equals $M^2/(M^2+1)$ and approaches one as M grows. After centering, the two populations span orthogonal rank-one subspaces, so their projector distance remains $\sqrt{2}$. Conversely, $x = ze_1$ and $y = ze_1 + Me_2$ have cosine $1/\sqrt{1+M^2}$ approaching zero, but identical centered subspaces and zero projector distance. These constructions demonstrate the difference between the statistics, including the geometric criterion's insensitivity to translation; they do not rank their predictive utility. Appendix K illustrates the first case.

3.4 MATCHED IMPLEMENTATION

We collect raw inputs and outputs of all blocks before final normalization. Both geometry and BI use 3,000 common token positions sampled from a 9,000-token pool drawn from 100 frozen assistant-message texts (10,577 nonpadding tokens). Centering uses float32; randomized subspace estimation and distance calculations use float64. A Gaussian range finder uses 120 nominal components plus 20 oversampling directions, with no power iterations. The reduced SVD retains all available components,

up to 140. Its generator is reinitialized with the recorded seed at each stage, while token subsampling uses a separate seeded stream.

For candidate rank r, estimated energy is the sum of the first r squared singular values divided by the exact $\|\widetilde{X}_l\|_F^2$. We retain the first rank reaching 90%, or the available spectrum if that target is unattainable. At all three seeds, 42 of 49 stages attain the target and seven reach the rank cap. The main seed is 42; sensitivity seeds are 7 and 123. For the same paired, uncentered vectors, BI is $1 - \frac{1}{n}\sum_i \frac{x_{li}^\top x_{l+1,i}}{\|x_{li}\|\|x_{l+1,i}\|}$; we likewise select its k smallest values, breaking ties by block index. The matched implementation uses paired positions and explicit block hooks; the historical selector used independent stage samples and has separate provenance (Appendix H.1).

For a stage of size $n \times d$ and projection width p, the leading estimator cost is $O(ndp + (n+d)p^2)$. Comparing adjacent thin bases costs $O(dr_l r_{l+1})$. These are algorithmic costs, not measured calibration times. Appendix H.1 provides the recovered estimator pseudocode and historical implementation details.

The selector takes a frozen calibration pool, a seed, an energy threshold, an estimator width, and a requested number of candidates. First, sample one common set of token positions and collect a representation matrix at every stage. Second, center each matrix and estimate its retained basis and achieved energy. Third, compare adjacent bases using Equation 1, recording both ranks alongside the scalar score. Finally, sort the block scores with the declared tie rule and return the candidate indices. The achieved-energy flag is part of the diagnostic output: it distinguishes a threshold-selected rank from truncation at the available spectrum. Computing BI from the same token positions provides a paired alternative without changing the intervention matrix. The released implementation exposes these numerical operations separately from model loading, so they can be checked on supplied activation matrices or on the preserved score records.

The diagnostic also admits a simple selection-stability check. Write the sorted scores as $s_{(1)} \leq \cdots \leq s_{(L)}$ and define the boundary gap $\gamma_k = s_{(k+1)} - s_{(k)}$. If every score changes by at most ϵ and $2\epsilon < \gamma_k$, every originally selected score remains below every unselected score. The selected set is therefore unchanged. This deterministic criterion separates stability of membership from the possible reordering of blocks within the selected set. It does not assume that calibration randomness obeys a known perturbation bound; the observed boundary gaps provide an interpretable scale against which such changes can be compared.

4 Matched Intervention Study

4.1 Model, workload, and endpoint

We use the unadapted 48-block DeepSeek-R1-Distill-Qwen-14B checkpoint at revision `1df8507178afcc1bef68cd8c393f61a886323761`, without adapters or free generation. Geometry and BI share calibration, and all selections are evaluated against the same single-block deletion matrix. The 144-question primary workload has 48 questions in each of three strata: named-composition recall, named-strength recall, and composition-to-property. Questions are selected deterministically; named grades do not repeat across the two recall strata and exact composition-condition duplicates are excluded. Fixed English reference sentences encode existing numerical values and intervals. Fourteen historical QA pairs form a separate secondary population. These questions come from the previously studied workload, not an independently held-out benchmark; training overlap is unresolved.

For each question, we teacher-force the reference tokens and average their NLL. The chat prompt ends with an open `<think>` marker, followed immediately by the reference sentence. The endpoint therefore measures compatibility with fixed wording inside a reasoning segment. It may respond to format or reasoning-policy changes as well as knowledge. Negative deletion changes mean higher likelihood of those tokens, not necessarily better generated answers. This conditioning is identical across the full model and all individual deletions.

Let d_{lq} be deletion-minus-full mean token NLL for block l and question q. For the k lowest-scoring blocks $S_k(m)$ under method m, define

$$R_k(m) = \frac{1}{144k} \sum_{l \in S_k(m)} \sum_{q=1}^{144} d_{lq}. \tag{3}$$

Each summand comes from a different one-block intervention. In particular, R_{15} is not the loss of a model with 15 blocks removed simultaneously. Early blocks 0–14 and late blocks 33–47 provide positional references.

4.2 Protocol and uncertainty

The predeclared primary contrast is $R_{15}(\text{geometry}) - R_{15}(\text{BI})$. Its percentile interval uses 2,000 stratified bootstrap resamples of the 144 declared question groups, with seed 2026. A favorable comparative result requires a negative contrast and an interval wholly below zero. This interval measures workload sensitivity conditional on the model and calibration. Calibration seeds are not independent model or training replications. Secondary checks cover $k = 1, 5, 10$, strata, score–loss association, and exclusion of block 0; no seed, budget, or question is selected after inspecting these outcomes.

For question q, first average its deletion changes within each selected set, then subtract the BI average from the geometric average. The bootstrap resamples these paired question-level contrasts within the three strata. It therefore preserves the fact that both selectors are evaluated on the same questions and that one full-model baseline is reused across all deletions. The 48 interventions are fixed candidates, not 48 independent experimental populations. Treating all block–question cells as independent would change the uncertainty calculation and discard the paired design. Likewise, the 14 historical questions do not enter the primary bootstrap, preventing the secondary population from altering the declared question mixture.

Each question contributes equally after its reference-token losses are averaged. Equal stratum sizes make the pooled result the mean of the three stratum means. Reporting all three exposes cancellation between composition and strength recall. Released records retain individual losses and token counts for checking this aggregation.

The full model and 48 separate deletions produce 7,742 question-level records across the 144 primary and 14 secondary questions. Restoring the full model afterward reproduces all 158 baseline losses exactly at recorded precision, within the fixed tolerance of 10^{-4} nats. The completed run takes 1,563 seconds on an Apple M5 Max with 128 GiB memory. Snapshot 005 preserves inputs, runtime, spectra, losses, bootstrap draws, and restoration checks; replay details appear in Appendix D.

5 Results

5.1 Selected-set intervention profiles

The full model's mean primary NLL is 3.5030 nats per reference token. Geometry and BI select sets with mean individual-deletion changes of -0.1483 and -0.1439, respectively, compared with $+0.0784$ for the last 15 blocks and $+0.4657$ for the first 15 (Table 1). Thus, both activation-based selections have lower conditional-loss means than these positional references. The primary geometry-minus-BI contrast is -0.00434 nats, with a 95% workload-bootstrap interval of $[-0.00970, 0.00064]$. This does not resolve an aggregate advantage or establish equivalence.

The contrast varies by stratum: $+0.04163$ for composition recall, -0.05017 for strength recall, and -0.00449 for composition-to-property. The historical secondary contrast is $+0.02960$, favoring BI. The selectors consequently identify different conditional-loss profiles rather than a uniform winner across question types or populations.

Each method preserves its own 15-block set across seeds 42, 7, and 123, so its R_{15} and primary interval are unchanged. Geometry and BI share six blocks. At seed 42, score–loss Spearman correlations over all 48 blocks are 0.3637 for geometry and 0.4420 for BI; excluding block 0 gives 0.3222 and 0.4056. BI has the stronger global rank association here. The first-15 reference is strongly

Table 1: Mean NLL changes from separate single-block deletions, primary seed 42. Each set contains 15 blocks; each primary stratum contains 48 questions. The 14 historical questions are a separate secondary population. Lower values mean greater reference-token likelihood.

Selection	Primary	Composition recall	Strength recall	Composition to property	Historical
Geometry	-0.1483	-0.1626	-0.1774	-0.1049	0.0015
BI	-0.1439	-0.2042	-0.1272	-0.1004	-0.0281
Last 15	0.0784	0.1112	0.0775	0.0466	0.1426
First 15	0.4657	0.4673	0.4399	0.4898	0.7339

Table 2: Candidate-budget sensitivity: mean individual-deletion NLL change on the 144 primary questions. Entries are selected-set diagnostics, not losses after joint deletion. Seed 42 is the primary analysis; seeds 7 and 123 give identical entries to each other.

Selector and calibration seed	$k = 1$	$k = 5$	$k = 10$	$k = 15$
Geometry, all three seeds	-0.2888	-0.0764	-0.1190	-0.1483
BI, seed 42	0.0853	-0.0897	-0.1506	-0.1439
BI, seeds 7 and 123	0.0853	-0.0897	-0.1252	-0.1439

affected by block 0. Removing that observation from its already selected set leaves a 14-block mean of -0.0352, still above both diagnostic-set means. The smaller-budget checks below examine whether the comparison depends on the chosen selection size.

5.2 Sensitivity to candidate budget

Table 2 reports the predeclared secondary budgets. Geometry's first candidate has a negative mean loss change, whereas BI's first candidate has a positive mean. The ordering reverses at five and ten candidates before the means become close at fifteen. Increasing k does not monotonically improve either selected-set mean: the score orders a representation statistic, rather than the measured deletion loss. These results make the intended use of the selector concrete. A candidate budget controls which part of its ranking is tested, so a comparison at one budget should not be extrapolated to all compression ratios.

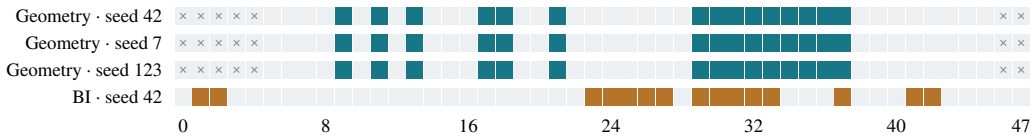
Geometry's reported means are unchanged across calibration seeds at all four budgets. BI changes at $k = 10$ while retaining its $k = 15$ set, illustrating that stable large-set membership can coexist with movement inside the ranking. The full per-question loss matrix is shared by all these comparisons; changing a seed recomputes candidate selection, not the underlying deletion experiments. The table is a sensitivity analysis of the same evidence, with the original $k = 15$ contrast remaining primary.

5.3 Angular variation explains the geometric choices

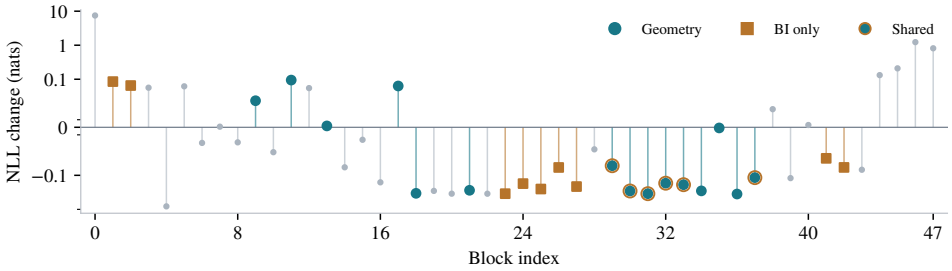
We retrospectively decompose the saved squared scores into $a_l = |r_l - r_{l+1}|$ and $b_l = s_l^2 - a_l$ using Equation 2. The angular term is recovered algebraically from recorded distances and ranks, not independently remeasured from saved singular vectors. Snapshot 008 records this component analysis separately from the predeclared comparison and checks ranks against the saved spectra and energy rule.

All 15 selected transitions have retained rank one at both endpoints and an attained energy target, at every seed. Sorting by b_l alone reproduces the complete geometry set at all three seeds. Rank change alone ties 45 blocks at seed 42 and 44 at the other seeds, so it cannot uniquely recover the selection. The retained ranks describe estimated centered calibration populations, not the model's hidden width or full representational capacity. The selected region is ordered by changes in the dominant direction of population variation.

At seed 42, rank changes occur only at blocks 4, 45, and 46 and contribute 33.86% of the sum of squared scores over all blocks, but zero within the selected set. Other seeds change ranks near blocks 44–46 without changing the choices. None of the chosen transitions touches a capped stage. This establishes an explanation of the recorded selection, without demonstrating that a variable-rank penalty improves it or that a higher-rank estimator would preserve every ordering. Figure 2 aligns



(a) Selected blocks



(b) Individual-deletion effects

Figure 2: Layer selections and individual-deletion effects. (a) Geometry selects the same 15 blocks across three calibration seeds; each chosen transition connects retained rank-one subspaces. Crosses mark transitions touching an energy-capped stage. (b) All 48 mean individual-deletion NLL changes, on a symmetric-log scale with linear threshold 0.05 nats. The outlying block 0 is retained. Geometry and BI share six blocks and each selects nine additional blocks. Each point represents a separate deletion, not a jointly compressed model.

Table 3: Shared and method-specific candidates at seed 42. Each entry averages separate block-deletion effects. The group sizes partition the union of the two 15-block selections; the question strata are the same as in Table 1.

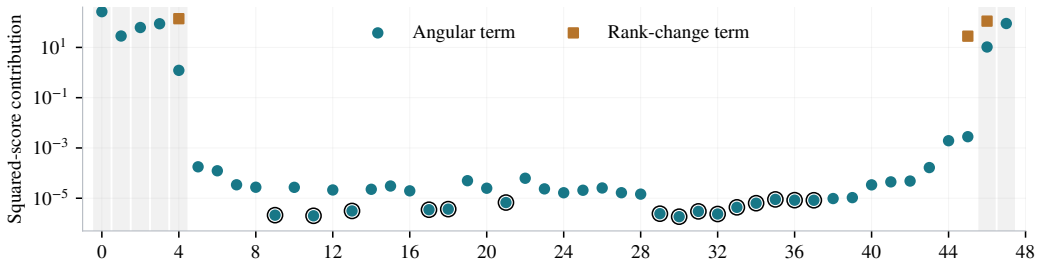
Candidate group	Blocks	Primary	Composition recall	Strength recall	Composition to property
Shared	6	−0.1939	−0.2670	−0.1959	−0.1187
Geometry only	9	−0.1179	−0.0930	−0.1650	−0.0956
BI only	9	−0.1106	−0.1624	−0.0814	−0.0882

the selection with direct interventions. Figure 3 shows the score decomposition and corresponding deletion effects at seed 42.

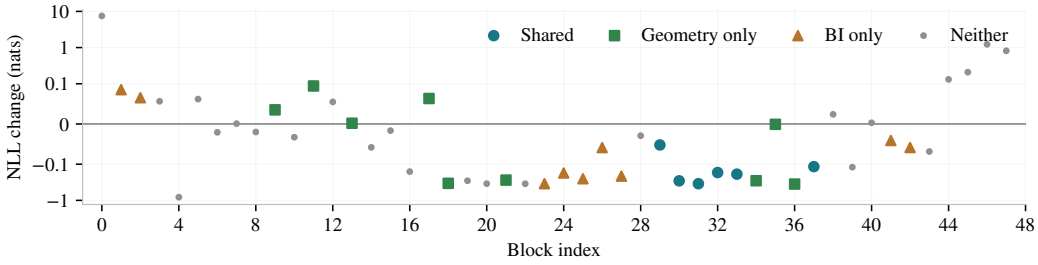
A close aggregate contrast does not imply that the methods identify the same region of the network. Their six shared blocks are 29–33 and 37. Geometry's nine additional choices are 9, 11, 13, 17, 18, 21, and 34–36; BI instead adds 1, 2, 23–27, 41, and 42. Table 3 separates their shared core from these two alternatives. The shared core has a more negative mean than either method-specific group, so it contributes substantially to both aggregate results.

The shared blocks cancel exactly in the geometry–BI contrast. Multiplying the difference between the method-specific primary means by $9/15$ recovers -0.00434 nats. The same decomposition applies within each stratum: the geometry-only group has lower loss on strength recall, while the BI-only group has lower loss on composition recall. This explains how appreciably different layer choices can yield similar aggregate means. It also identifies the nine-for-nine substitution that a subsequent comparison of combined deletions would need to test, without attributing the shared core's effect to one selector alone.

Replaying the saved geometric scores gives gaps between the 15th and 16th candidates of 1.2002×10^{-4}, 1.5450×10^{-4}, and 1.1297×10^{-4} at seeds 42, 7, and 123, respectively. The largest selected score is between 0.002988 and 0.003002. All selected transitions consequently fall far below the unit threshold for a rank change, consistent with the recorded rank-one explanation. The stability criterion in Section 3.4 certifies unchanged membership for uniform score perturbations smaller than



(a) Angular and rank-change contributions



(b) Individual-deletion effects by selection

Figure 3: Geometric score components and deletion effects at seed 42. (a) Angular and rank-change contributions to the squared score on a logarithmic scale; zero terms are omitted. Rings mark geometry-selected blocks, and shading marks transitions touching an energy-capped stage. (b) Mean NLL changes from separate single-block deletions on 144 questions, grouped by selector membership. The vertical scale is symmetric-logarithmic with linear threshold 0.1 nats. This is a retrospective decomposition of the recorded measurements.

half the corresponding gap. These are retrospective margins computed from the released scores, not additional calibration trials or probabilistic error bounds. Together with identical membership across seeds, they make both the observed stability and its numerical scale inspectable.

Angular variation is a weaker fine-grained loss ranking within rank-one transitions. At seed 42, its Spearman correlation is 0.1373 over 40 such transitions, compared with 0.3637 over all 48 blocks. At the other seeds it is 0.0696 over 39 transitions. The component analysis thus explains which information determines selection while bounding the interpretation of that information as a general deletion-sensitivity predictor.

5.4 Resource context from historical compact models

A separate historical branch removes 15 blocks together and applies domain CPT. It evaluates the original 48-block model and three 33-block configurations using geometry, BI, or contiguous selection. All compact models use their step-61,201 adapters; the original has no additional CPT. This branch's geometric set differs from the matched set by one block (39 in place of 36), and its calibration uses different sampling conventions. Its results therefore describe constructed checkpoints, not a joint-deletion validation of the matched selector.

All four configurations share 1,118 questions, base revision, tokenizer, bfloat16 precision, batch size one, and greedy generation capped at 300 new tokens. Token throughput divides output tokens by summed device-synchronized generation-call time, including prefill and excluding loading and adaptation. Sequential runs use the same shared Apple Silicon machine. Table 4 reports fewer parameters and higher observed token throughput for each compact checkpoint, with no repeated-run timing interval.

The original prefix rubric gives higher aggregate scores to compact checkpoints, but its question-echo control exposes a failure to distinguish some unanswered inputs. A retrospective, category-balanced semantic reading of 30 questions does not recover a compact-model advantage; its 300-token cap

Table 4: Historical resource measurements under the common 300-token protocol. All compact configurations receive domain CPT; the full original does not. These are token rates, not rates of correct or completed answers.

Configuration	Blocks	Parameters (B)	Tokens/s	Relative rate
Original	48	14.770	11.92	$1.00\times$
Geometry + CPT	33	10.641	16.03	$1.35\times$
BI + CPT	33	10.641	17.27	$1.45\times$
Contiguous + CPT	33	10.641	18.40	$1.54\times$

and single AI evaluator limit generalization. Separately, a six-question 2,048-token pilot records EOS with a final-answer boundary for five original responses and none of the compact responses. Some capped texts contain direct answers, so these counts measure termination rather than factual correctness. We retain rubric controls, semantic reading, and interface evidence in Appendices I, J, and E. Together, these observations support resource savings without establishing preserved answer quality or isolating pruning from adaptation.

6 Discussion and Scope

The method offers a population-level alternative to tokenwise cosine whose selected set can be traced to specific angular changes. Its intervention means are lower than the tested positional references, and its distinction from BI is visible in both layer membership and workload strata. The current evidence supports a characterized selection signal; it does not establish a universally stronger ranking. The weak within-rank-one correlation is particularly relevant when interpreting the angular mechanism.

Single-block losses do not capture interactions among deletions. Consequently, the stable selection and favorable positional comparison justify studying these candidates, but do not certify simultaneous pruning or answer preservation. The historical resource study demonstrates a feasible smaller checkpoint; its adaptation and interface outcomes must be assessed separately. A deployment decision would require direct answer-quality evaluation of the intended compact checkpoint and decoding protocol. Increasing the generation budget alone does not supply that validation.

The matched study covers one model family, one domain, one calibration pool, and one response-conditioning protocol. Bootstrap intervals quantify workload sensitivity, not uncertainty across independently trained models. Reference-token loss inside an open reasoning segment can change with response policy, and workload records have unresolved training overlap. Seven activation stages hit the rank cap. The post-hoc decomposition does not isolate the causal value of centering or the benefit of variable rank over simpler estimators. The historical adaptation corpus and calibration revision are not fully reconstructable; its question exposure and heuristic scoring are disclosed in the appendices. These boundaries identify the additional evidence needed for a broader compression claim, while leaving the reported selection behavior and matched intervention measurements directly inspectable.

7 Conclusion

Centered activation subspaces provide an interpretable criterion for selecting candidate layers for compression. Matched single-block interventions show lower mean conditional reference-token loss for geometry and BI selections than for early and late positional references, with no resolved aggregate geometry–BI advantage. Angular changes between retained rank-one subspaces recover all 15 geometric choices across three calibration seeds, explaining the observed selection. A separate historical compact checkpoint supplies a resource result of 28.0% fewer parameters and $1.35\times$ observed token throughput. The contribution is a specified and empirically characterized layer-selection signal, with candidate diagnosis kept distinct from validation of a jointly compressed model.

AI USE STATEMENT

Generative AI assistants were used in literature search, data inspection and computational audits. Reported numerical results come from executed experiments and retained output artifacts. The human authors are responsible for the final submitted text, claims, references, and artifacts.

REPRODUCIBILITY STATEMENT

The selection rule and analytical construction are specified in the main text, and Section 4 describes the matched intervention protocol; historical scoring and configurations are documented in Appendix H. Appendix C indexes the eleven frozen experiment snapshots, checksums, and replay tooling; Appendices D, E, and F provide diagnostic, completion, and component-analysis details. The anonymous repository linked in the introduction provides the selector implementation, matched-loss and component-analysis code, 7,742 per-question loss records, calibration scores for three seeds, numerical summaries, and a hash-checked analysis replay script. The released numerical analyses reproduce the central matched and component results. Model weights, calibration activations, question/reference text, and the training corpus are not included; the release does not reproduce historical training or fresh model inference.

REFERENCES

V. M. Buchstaber. Time Series Analysis and Grassmannians. In *Applied Problems of Radon Transform*, volume 162 of *American Mathematical Society Translations, Series 2*, pp. 1–17. American Mathematical Society, 1994. URL https://homepage.mi-ras.ru/~buchstab/download/AMSTransl2.1994.162.1.pdf.

Xiaodong Chen, Yuxuan Hu, Jing Zhang, Yanling Wang, Cuiping Li, and Hong Chen. Streamlining Redundant Layers to Compress Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4b00a351b41358965613c118e87dc28c-Abstract-Conference.html.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948, 2025. URL https://arxiv.org/abs/2501.12948.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The Unreasonable Ineffectiveness of the Deeper Layers. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2403.17887v2.

Bowei He, Lihao Yin, Huiling Zhen, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. PASER: Post-Training Data Selection for Efficient Pruned Large Language Model Recovery. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/180d2acb13633fe78688d0d2347c731f-Abstract-Conference.html.

Cristian Hinostroza, Rodrigo Toro Icarte, Christ Devia, Andres Carvallo, Eugenio Herrera-Berg, Denis Parra, and Jorge Silva. Rethinking Layer Relevance in Large Language Models Beyond Cosine Similarity. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4aa5e0b33e6f774c0d34ab2e1bc3bced-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in Large Language Models are More Redundant Than You Expect. arXiv preprint arXiv:2403.03853, 2024. URL https://arxiv.org/abs/2403.03853.

Sungbin Shin, Wonpyo Park, Jaeho Lee, and Namhoon Lee. Rethinking Pruning Large Language Models: Benefits and Pitfalls of Reconstruction Error Minimization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2406.15524.

Jiwon Song, Kyungseok Oh, Taesu Kim, Hyungjun Kim, Yulhwa Kim, and Jae-Joon Kim. SLEB: Streamlining LLMs through Redundancy Verification and Elimination of Transformer Blocks. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.09025.

A Historical Structural Diagnosis

The historical single-block probe removes one block at a time and measures the change in gold-answer negative log likelihood relative to the full model. The probe uses a separate 60-text calibration routine, rather than the main selector's 100 assistant-message texts. The recorded Spearman correlation between geometric score and intervention damage is approximately 0.325. The mean recorded damage is -0.0124 for the 15 blocks with the lowest geometric scores in the recovered probe, compared with 0.7547 for the 15 blocks with the highest geometric scores; the predefined geometric removal set has mean damage 0.0049. These are exploratory diagnostic measurements under the probe's own protocol, not additional observations from the 1,118-question final comparison. Single-block effects do not determine the effect of deleting a set simultaneously. The recovered helper defines 14 fixed QA pairs and a domain system prompt followed by the question in a chat template. Gold answers are teacher-forced; token losses are averaged within each answer and then equally across the 14 answers. The calibration helper selects up to 60 qualifying corpus chunks at stride 401, truncates them to 256 tokens, and requires at least 16 tokens after tokenization. These sources are preserved in snapshot 004 as recovered current files, without asserting an exact historical executable identity.

B Historical Learned-Bridge Evidence

The bridge branch fits a residual feed-forward mapping to activation pairs across a removed span and explores blending with an identity path. The recorded calibration comprises 131,047 tokens at hidden width 5,120. A subsequent short LoRA stage uses rank 16 and scaling parameter 32, distinct from the rank-32 long-CPT branch used in the historical compact-model experiment.

For the recorded blend setting 0.5, the NLL trajectory is 3.4552 initially, 2.2129 at step 300, and 2.0484 at step 3,000. The recorded setting 0.0 gives 6.7779, 2.4066, and 2.1435 at the same stages. This supports faster early alignment and a lower final NLL in that comparison. A separate small free-generation evaluation records 0.4667 versus 0.6000, respectively; improvement on the NLL trajectory does not imply improvement on every endpoint. These exploratory results motivate treating the bridge and its training objective as design choices. They do not supply the historical comparison's 4,472-response comparison or establish superiority over its no-bridge compact models.

Subsequent paired replication. A new seeded N15 experiment preserves separate 300-step adapters for the zero-blend control and the 0.5-blend bridge, using identical initial LoRA weights and 300 training batches. Their post-training gold NLL values are 2.3950 and 2.1679; historical prefix-rubric scores are 0.30 and 0.35. Neither configuration completes any of six diagnostic responses within 2,048 tokens. A separate check on the first of those questions extends each generation to 32,768 tokens: both reach the cap without EOS or a closing reasoning marker and persist in repetition. The first 2,048 tokens match the shorter run exactly. These are new replication and budget-diagnostic results, not recovered historical weights, an accuracy estimate, or an evaluation of the main geometry-selected checkpoints. Snapshots 010 and 011 preserve the complete records; the latter uses per-token MPS cache release to complete execution without the memory failures encountered in earlier attempts.

C Artifact and Reproduction Map

Released artifact. The anonymous repository linked in the introduction contains the selection and analysis code and numerical results from snapshots 005 and 008. Download the repository, install `requirements.txt`, and run `python3 verify.py` to check hashes and replay both analyses. The archive inventory below also includes locally retained materials outside this public release.

The local research archive contains eleven frozen evidence snapshots, plus separately versioned supplementary analyses. The manuscript ZIP contains sources and figures, not these complete snapshots. Snapshots 001–008 support the historical comparison and matched diagnostics; snapshots 009–011 document separate learned-bridge follow-ups:

- `experiments/exp-001-clean-final-comparison/`: all four sets of generated responses; the 1,118 questions; configuration, loading, and completion records; evaluation

code; configuration files; environment lock; category analyses; and SHA-256 hashes. This is the source for Table 4 and Appendix I.

- `experiments/exp-002-historical-method-evidence/`: structural and bridge experiment passports, selected historical source and result files, the Buchstaber source lock, and historical revision qualifications. This supports the historical method details, Supplementary table S7, and Appendices A–B.
- `experiments/exp-003-workload-construction/`: recovered generator, merger, scoring helper, input question sets, and SteelBench source table. Replaying the construction matches the frozen questions record for record. Local preservation does not assert public redistribution rights.
- `experiments/exp-004-probe-protocol/`: recovered helper sources, 14 QA pairs, and protocol notes for the historical diagnostic. The exact historical executable revision remains unverified.
- `experiments/exp-005-matched-layer-diagnostic/`: matched inputs, execution and analysis source, 7,742 losses, calibration scores and spectra, bootstrap draws, restoration checks, environment lock, and SHA-256 manifest. Large raw activation arrays have local pointers and hashes; durable public storage remains a release task. This supports Table 1 and Appendix D.
- `experiments/exp-006-completed-answer-pilot/`: 24 generations with a 2,048-token cap, exact token records, annotation controls and both versions of calibration, model-blinded annotations, configuration identities, and replay code.
- `experiments/exp-007-final-format-control/`: 24 fresh paired generations with a common 512-token cap, locked prompts, stop records, structural analysis, token verification, environment, and non-mutating replay. These two snapshots support Supplementary table S1 and Appendix E.
- `experiments/exp-008-geometry-components/`: retrospective decomposition of all 48 matched geometric scores at all three seeds; copied source losses and spectra, per-block components, group profiles, numerical tests, plot, and exact analysis replay. No model execution is added. This supports Section 5.3 and Appendix F.
- `experiments/exp-009-n10-bridge-pilot/`: separate N10 bridge-only pilot, with full, pruned, and bridge configurations on six questions; this does not test the N15 short-adaptation branch.
- `experiments/exp-010-n15-300-paired/`: new paired 300-step N15 replication, checkpoint identities, training records, likelihood diagnostics, and generated responses.
- `experiments/exp-011-n15-long-budget/`: one-question paired 32,768-token diagnostic, raw token IDs, exact-prefix checks, and execution-memory records. These follow-ups are separate from the main 4,472-response comparison.
- `revisions/eos-serialization-audit/`: hashed preprocessing-source copies, AST-extracted function replay, tokenizer checks, and audit results. This is a source-level reconstruction, not a new training experiment.
- `revisions/target-evaluation-v1/` and `target-evaluation-v2/`: frozen specifications, parsers, constructed controls, timestamped hash locks, arm-level results, and all extracted spans. V2 was developed after V1 results were known; this chronology is preserved.

The loading records name a field `effective_generation_config`, but it contains model defaults before call-time overrides. The executed protocol is determined by the explicit arguments in `code/run_eval.py` and the frozen configuration: in particular, greedy decoding overrides the stored sampling default. The loading record is preserved unchanged.

The first snapshot identifies the code capture commit and records that the runtime hashes match the completed runs. Its capture commit is distinct from the historical training commit. The second snapshot explicitly marks unresolved executable revisions as approximate; preserving a script today does not establish its exact historical execution environment.

The following command, run from the paper directory, verifies the snapshot hash manifests and rescores all 4,472 responses using Python's standard library:

```
python3 src/rescore.py
python3 src/verify_workload.py
python3 src/analyze_diagnostic.py
python3 src/check_target_score_v2.py
python3 src/run_target_score_v1.py
python3 src/run_target_score_v2.py
```

Reproduction target	Packaged evidence	Remaining access or provenance
Rescore stored outputs	Questions, responses, frozen scorer, hash checks	None for local package
Regenerate outputs	Pinned base revision, adapter identities, layer sets, evaluation code	Base and adapter assets; compatible runtime and launch setup
Reselect and retrain	Historical scripts, settings, selected result records	Exact calibration bytes, corpus access, historical splits and dependencies

Independent rescoring requires the frozen responses, question references, and scorer, without loading a model. Regenerating responses additionally requires the pinned base checkpoint, the corresponding adapter and removal metadata, and a compatible inference environment. Retraining additionally requires the knowledge corpus and historical preprocessing decisions. The draft does not claim that all model weights and corpus rights have already been cleared for public distribution. Publication preparation must identify which artifacts can be distributed and provide access instructions for the remainder.

The package preserves the distinction between a runnable evaluation specification and a completed public release. A public reproduction repository can expose the code, environment, layer sets, identities, hash manifests, and permitted result artifacts while documenting access to the model and corpus. No public repository address is asserted in this draft.

D Matched Diagnostic Sensitivity and Replay

Secondary budget checks, primary seed 42. Mean individual-deletion NLL changes; budgets are not simultaneous deletion experiments.

Selection	k=1	k=5	k=10
Geometry	−0.2888	−0.0764	−0.1190
BI	0.0853	−0.0897	−0.1506
Last blocks	0.8129	0.4635	0.2124
First blocks	7.4773	1.3735	0.6925

At seeds 7 and 123, BI's R10 is −0.1252; all other entries in this table are unchanged. Geometry's rank correlations at seeds 7 and 123 are 0.3646 and 0.3642; BI's are 0.4439 and 0.4426. All primary R15 results are unchanged across these seeds because each method's selected set is unchanged. These sensitivity seeds use the same calibration pool and deletion matrix, so they are not independent replications.

The geometric set is {9, 11, 13, 17, 18, 21, 29, 30, 31, 32, 33, 34, 35, 36, 37}; BI selects {1, 2, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 37, 41, 42}. These sets belong to the matched experiment, not a new evaluation of jointly pruned or adapted checkpoints.

To reproduce the matched analysis, copy snapshot 005 to a writable scratch directory and run its `reproduce.sh` with NumPy installed. The frozen results contain the 48-by-158 damage matrix, all per-question source losses, and bootstrap indices. The reproduction command writes derived files, so

it should not be run directly in the immutable original. An initial attempt stopped at an instrumentation keyword collision after calibration and before the historical compact-model comparison; its failure record is retained. The completed run uses a fresh output directory and the same frozen inputs and metric decisions.

E Response Completion and Serialization Evidence

E.1 Answer content is not an EOS event

The boundary parser first checks whether a reasoning segment was opened in the prompt. If no closing marker appears in the output, it returns `candidate_final=None`. This implements an explicit chat-format criterion; it does not inspect the output for a semantically complete answer. The same historical checkpoints and removal sets are used in the 300-token comparison and the six-question follow-up. For all 24 paired records, the earlier response text is an exact prefix of the longer response. The follow-up therefore changes the budget and endpoint, rather than demonstrating a newly broken model.

For example, the geometry checkpoint's response to item `v2_0113` begins:

> Answer: The nominal carbon content for steel grade 1213 is approximately 0.13 wt%.

The stored reference is 0.13 wt%. The same checkpoint begins item `v2_0158` with a carbon value of 0.15% for grade 302, and item `v2_0177` with 12% Ni for grade 308; both agree with their stored references. These outputs then continue with other questions, without closing the prompted reasoning segment. These retrospectively selected examples establish the existence of direct reference-matching answers in the saved output, not an accuracy rate or independent verification of the source values. Other outputs contain errors, contradictions, and repetition.

The pilot's six questions are the first two entries per scenario in the frozen diagnostic ordering. They are not a representative sample of the 1,118-question workload. All six reference audits record ambiguities in grade/condition naming or in inferring properties from composition and heat treatment. These ambiguities do not invalidate recorded token stops; they prevent turning stop counts into a factual-quality ranking. The raw-text CPT preparation and open-reasoning prompt also differ in format. The paired closing-prefix control tests one local change to that prompt, not restoration of instruction following or a causal separation of pruning and CPT.

E.2 Completed-answer and format-control protocols

We retain the original 300-token comparison and conduct two separate follow-ups on the same pinned base model, historical removal sets, and step-61,201 compact adapters. The first freezes six questions, two from each numerical scenario, before generation. Each configuration receives the same questions and a maximum of 2,048 new tokens, with a 300-second per-answer and 90-minute total budget. A recorded EOS and a nonempty final answer are distinguished from exhausting the token cap. The prospective completion gate requires at least five of six completed answers in every configuration before expanding evaluation. Raw token IDs, exact prompts, stopping reasons, and timings are preserved in snapshot 006.

Two separate Codex-agent annotation passes inspect model-blinded records after common calibration. They agree on the completion labels but are neither human experts nor independent validations of the reference set. Calibration rules were revised after two shared disagreements on 12 constructed controls, before inspecting natural outputs; both passes then agree with all 12 revised controls. The earlier rules and results remain preserved. Numerical reference agreement and semantic assessment are recorded separately from completion (Appendix E).

The second follow-up fixes the first question from each pilot scenario and generates fresh paired open/closed-prefix outputs for all four configurations: 24 calls in total. The closed mode only appends a literal closing reasoning marker and two newlines to the unchanged prompt. Both modes use 512 new tokens, greedy decoding, repetition penalty 1.1, and the same precision, device, and weights. No answers or extra instructions are supplied. The principal descriptive endpoint is the

within-configuration change in EOS with nonempty structurally final text, out of three questions. The supplied closing marker is not counted as a transition generated by the model. The protocol has a 90-second per-answer and 30-minute total budget, with a 30-second watchdog grace; no budget extension or adaptive prompt search is allowed. Snapshot 007 preserves the pre-generation protocol, alternating mode order, results, and executable analysis.

These are sequential diagnostic follow-ups on a known workload, not an independently held-out validation study. Their different token budgets are reported separately. Only the second experiment supplies a symmetric paired test of the closing-prefix intervention; neither isolates CPT from pruning.

E.3 Chat-turn termination and a paired prefix control

The longer-budget follow-up measures compliance with the specified chat-turn termination rule. In the six-question pilot, the original model produces five completed final answers with EOS; every compact configuration reaches the 2,048-token cap on all six questions (Supplementary table S1). None of the 18 compact outputs closes the reasoning segment opened by the prompt. Both calibrated annotation passes agree on completion for all 24 outputs. The pilot therefore fails its criterion of at least five completed answers per configuration, and the larger evaluation does not proceed.

Supplementary table S1. EOS-and-boundary pilot and fresh paired prefix control. Entries count EOS with a nonempty final answer under the boundary parser, not factual accuracy or the presence of answer content. Pilot completion also received two calibrated model-assisted annotation passes. The format control uses a structural endpoint and a single editorial inspection. Its open and closed modes both use 512 new tokens; the earlier pilot is not substituted as the open control.

Configuration	Pilot: 2,048 tokens (n=6)	Open: 512 tokens (n=3)	Closed: 512 tokens (n=3)
Original, 48 blocks	5/6	1/3	3/3
Contiguous, 33 + CPT	0/6	0/3	0/3
Geometry, 33 + CPT	0/6	0/3	0/3
BI, 33 + CPT	0/6	0/3	0/3

Closing the supplied reasoning prefix increases completion from one to three questions for the original model, but leaves all three compact configurations at zero. All 18 compact control outputs reach 512 tokens, and none generates a new opening reasoning marker. Their continuations include repeated assertions, additional self-generated questions, and document-like passages. Thus, supplying the closing marker alone is insufficient for completion on these questions and checkpoints within this budget. This result neither establishes that a longer or different protocol would fail nor identifies compression or CPT as the cause.

Completion also does not certify factual correctness. In the pilot, one original-model final answer agrees numerically with its reference within the specified tolerance but contains an inconsistent explanation of percentage units. The other four completed answers involve insufficient or ambiguous context and references; these cannot be counted automatically as four factual errors. Capped outputs are not substituted for completed answers in that assessment. Appendix E documents the calibration, representative control outputs, and serialization audit. The pilot takes 3,585.756 seconds and produces 42,448 tokens; the format control takes 791.021 seconds and produces 11,022 tokens. Both finish all planned calls without errors or timeouts.

E.4 Pilot annotation and controls

The six pilot questions are fixed before generation and span named composition, named strength, and composition-to-property queries. Their references are recovered from 14 source rows, whose independent physical validation is not established. All four arms use the same inputs and generation settings. The original model has five EOS stops and one cap stop; each compact arm has six cap stops. The 19 capped texts have no completed-answer assessment substituted from their reasoning fragments.

The two model-assisted passes agree on all natural-output completion labels after shared calibration. In the initial 12 controls they share two disagreements with expected labels. Revised rules specify

that an incomplete question echo receives no final-answer score, and that a completed refusal without the reference assertion is a numerical nonmatch without necessarily being factually wrong. One expected control label changes. Both passes subsequently achieve 12/12 on the revised controls. This repeated calibration is disclosed rather than presented as independent evaluator validation. All control versions, individual labels, and exact supporting spans are retained in snapshot 006.

One completed original response satisfies numerical reference tolerance but equates 18 thousandths of a percent with 0.18%, an arithmetic inconsistency in its explanation. Four other completed responses involve insufficient or ambiguous context/reference specification. Consequently, neither the five completions nor their numerical agreement counts establish a factual-accuracy estimate.

E.5 Paired format control and observable continuation

The paired questions are v2_0089, v2_0487, and v2_0615, selected as the first pilot item in each scenario. Modes differ only by the supplied closing marker; all 24 generations are fresh. For the original model, v2_0487 completes in both modes, while the other two complete only with the closed prefix. Original outputs total 1,375 tokens in open mode and 431 in closed mode. Each compact configuration produces 1,536 tokens per mode, exhausting all three per-answer caps. No output reopens reasoning, so the compact result cannot be explained by a generated reopening in this control.

Editorial inspection records several distinct continuations. The geometry-selected model answers the closed composition query with approximately 0.95 wt% and then generates further questions about other elements and steel grades. Its closed composition-to-property output gives 934 MPa and 690 MPa before continuing with another question and explanation. The contiguous closed strength output gives approximately 125 MPa and repeats a specification-based explanation. The BI closed composition-to-property output repeats 345 MPa and 290 MPa until truncation. These are quoted model assertions, not validated physical values; they differ from saved references and illustrate why a recognizable answer fragment cannot certify a correct completed response. This inspection is a single editorial pass, not a repetition of the pilot's two-pass annotation.

The prefix intervention is motivated by the observed pilot and source audit, whose outcomes are already known. The three-question control supports a bounded comparison of interfaces, not population inference or prompt-selection claims. No alternative suffix, checkpoint, or token budget is selected afterward to replace its results.

E.6 Source-level serialization audit

All 24 pilot raw outputs decode exactly from their stored token IDs, and all prompts re-encode exactly to their recorded input IDs. The same checks pass for all 24 format-control outputs. The effective EOS ID is 151643 in all arms, and no configured suppression or minimum-generation length explains the contrast. Each rendered prompt has one beginning-of-sequence token; adding tokenizer special tokens does not duplicate it. The prompt itself opens the reasoning segment, which is therefore not expected among newly generated tokens.

Recovered MPS and CUDA preprocessing functions encode ordinary documents without special tokens, split them into chunks of at most 512 tokens, and discard chunks shorter than 64 tokens. Appended question texts also receive no EOS from this tokenizer. Padding uses the EOS token ID but sets the corresponding labels to -100. AST-extracted functions executed on synthetic ordinary text produce zero EOS training targets; inserting one literal EOS into content produces one target. Thus, padding is masked correctly, and actual EOS content is not removed merely because it shares the padding ID. The preparation functions agree across preserved and recovered source copies, but the exact executable revision of each historical training run remains unresolved.

These checks establish that the recovered procedure does not explicitly append end-of-document or assistant-turn markers to ordinary content. They do not establish that the entire historical corpus contained no such tokens. Continued pretraining on raw text is compatible with domain adaptation; retaining a desired response interface after that adaptation requires its own evaluation. The paired control tests one simple interface intervention, without separating the contributions of pruning, CPT, or their interaction.

The audit also identifies the open-reasoning conditioning of the matched NLL endpoint in Section 4.1. All its numerical records and bootstrap results remain unchanged. The additional qualification concerns what that likelihood measures, not its arithmetic reproducibility. In particular, a reduction in this loss cannot alone distinguish improved reference knowledge from a change in format preference.

F Rank and Angular Contributions

The identity $s_l^2 = a_l + b_l$ separates the absolute rank difference from the nonnegative angular residual. For every recorded block and seed, the analysis checks $0 \le b_l \le 2\min(r_l, r_{l+1})$ within numerical tolerance. It recomputes the chosen rank and achieved energy from the stored singular spectrum and total centered variance. The saved spectra verify the implemented energy decision; without re-estimating the bases, they do not test estimator bias. Group deletion means are recomputed from the original question-level losses and reproduce the existing R15 results. Snapshot 008 packages all 144 seed–block rows and exact standard-library replay. The angular residual does not recover individual principal angles.

Calibration seed	Rank-changing blocks	Rank-one transitions	Selected rank-one transitions	Angular-only overlap	Within-rank-one correlation
42	3	40	15/15	15/15	0.1373
7	4	39	15/15	15/15	0.0696
123	4	39	15/15	15/15	0.0696

At seed 42, stages 0–4 and 47–48 are capped at rank 140, stages 5–45 have retained rank one, and stage 46 has rank 29. Seeds 7 and 123 instead give stage 45 ranks 11 and 4, and stage 46 ranks 90 and 44. The cap-affected transitions remain blocks 0–4, 46, and 47. The chosen geometric set is unchanged and contains no cap-affected transition. Geometry and BI share blocks {29, 30, 31, 32, 33, 37}; geometry alone chooses {9, 11, 13, 17, 18, 21, 34, 35, 36}, and BI alone chooses {1, 2, 23, 24, 25, 26, 27, 41, 42}.

The rank-change-only cutoff is zero, with 45 tied candidates at seed 42 and 44 at the other seeds. Depending on arbitrary tie breaking, a 15-block rank-only choice can overlap the geometric set in anywhere from zero to all 15 blocks. We therefore do not report a single rank-only performance number. Angular-only ranking preserves the chosen set, but this is removal of a term from saved scores, not a new calibration or adaptation experiment. It does not establish the superiority of a fixed-rank estimator.

The rank term accounts for 33.86%, 32.54%, and 33.67% of the pooled squared-score sum for seeds 42, 7, and 123, respectively. Pooled fractions emphasize large scores outside the selected set; they are not proportions of explained deletion loss. Full-score correlations with primary damage are 0.3637, 0.3646, and 0.3642. Angular-term correlations over all 48 transitions are 0.3841, 0.3896, and 0.3843; rank-term correlations are 0.1097, 0.1971, and 0.1976. Ties receive average ranks. These descriptive statistics share one intervention matrix and carry no independent-sample significance claim.

Supplementary table S2. Shared and differing selections in the matched diagnostic. Means average separate single-block deletion effects, in nats per reference token, within each group. These groups are identical across the three calibration seeds. They do not describe jointly deleting each group. All losses retain the open-reasoning conditioning of Section 4.1.

Block group	Count	Primary mean	Composition recall	Strength recall
Shared	6	−0.19388	−0.26703	−0.19586
Geometry only	9	−0.11787	−0.09297	−0.16502
BI only	9	−0.11064	−0.16236	−0.08140

G SUPPORTING DIAGNOSTIC FIGURES

These retained figures document the historical removal sets, prefix-rubric operating points, and score–damage association. Score decomposition appears in Figure 3 in the main text. These supporting plots complement the main figures by exposing the full diagnostic and historical context. The prefix-rubric ordinate below is not a validated answer-quality measure.

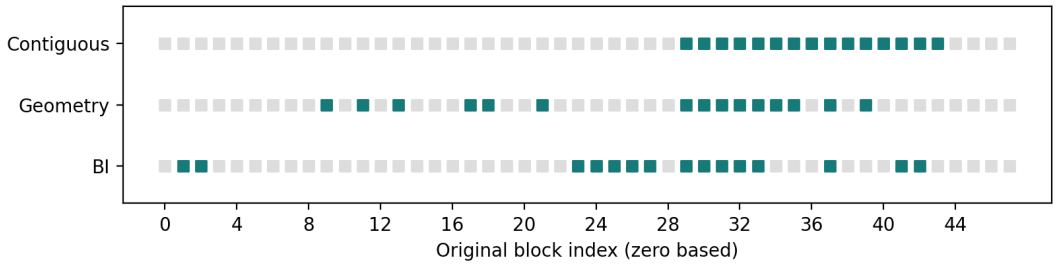


Figure 4: Removal sets at a common budget of 15 blocks. Colored squares indicate removed blocks.

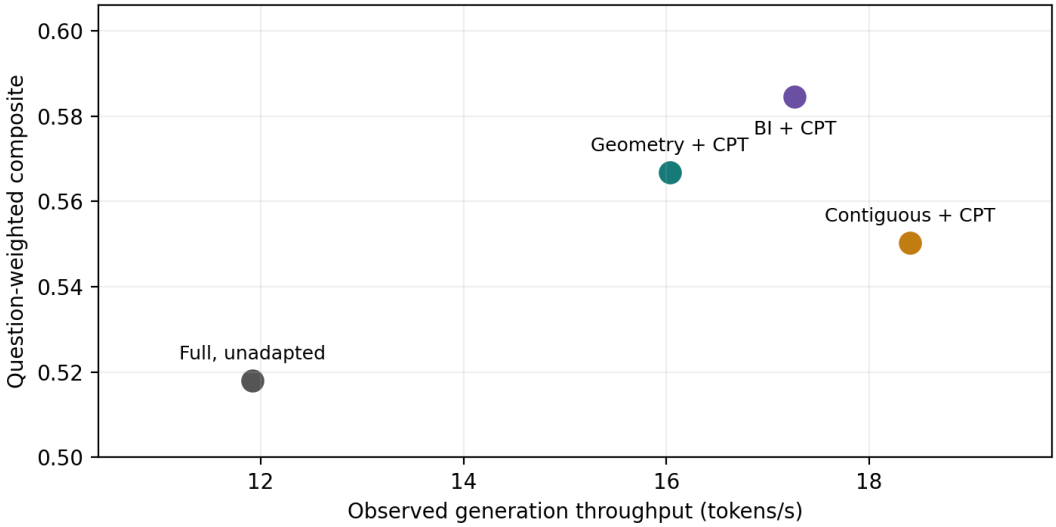


Figure 5: Measured operating points under the common generation budget. All compact models have 33 blocks; the full model has 48.

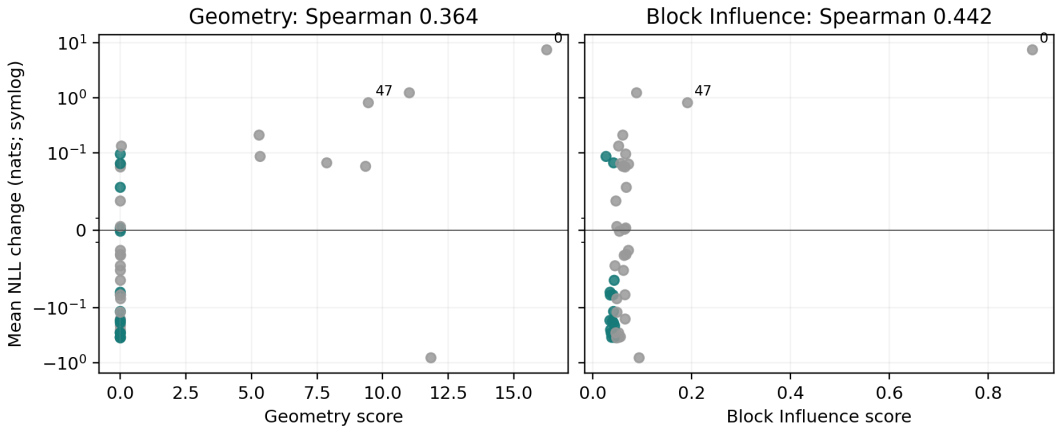


Figure 6: Matched diagnostic scores versus mean individual-deletion NLL change on 144 questions. Green points mark each criterion's selected 15 blocks; gray points mark the remainder. The vertical axis uses a symmetric logarithmic scale. Block labels 0 and 47 expose boundary interventions.

H HISTORICAL COMPACT-MODEL STUDY

The following sections preserve the separate historical compression-and-adaptation protocol. Their geometric removal set and calibration are distinct from the matched intervention study.

H.1 IMPLEMENTED ESTIMATOR AND SELECTION

The historical selector uses 100 assistant-message texts from the training source, a maximum sequence length of 256, and batch size two. It collects unpadded vectors at all hidden-state stages, caps each stage's pool at 9,000 vectors, and subsamples at most 3,000 vectors for subspace estimation. Stage samples are drawn independently. Observed distances therefore combine population change with finite-sample and randomized approximation effects. Vectors are centered in float32 and then cast to float64 for the randomized decomposition and subspace-score computation.

A randomized range finder uses a nominal component count of 120 and 20 oversampling directions. The implementation retains the entire reduced singular-value decomposition, up to 140 components, rather than truncating back to 120. For candidate rank r, the retained energy is the cumulative squared estimated singular values divided by the exact squared Frobenius norm of the centered matrix. The implementation selects the first rank reaching the requested threshold, clipped to the available estimated spectrum. Consequently, the nominal energy threshold need not be reached when the spectrum truncation is limiting. The historical setting uses 90%; sensitivity checks also use 85% and 95%.

Blocks are ordered by increasing score, with block index resolving ties, and the lowest 15 are removed. The recorded main removal set is

$$\{9, 11, 13, 17, 18, 21, 29, 30, 31, 32, 33, 34, 35, 37, 39\}.$$

Indices are zero based. The historical checks retain the same set across the three energy thresholds; runs with seeds 42, 7, and 123 share 14 of the 15 selected blocks. These checks characterize this calibration setting rather than a universal stability guarantee. The source code and historical records are preserved separately from the final evaluation snapshot because the exact historical calibration file revision is unresolved.

The historical single-block probe provides a complementary intervention check: its geometric scores have a Spearman correlation of approximately 0.325 with measured NLL damage, and the low-score group has lower average damage than the high-score group (Appendix A). This association supports using the diagnostic to generate removal candidates. The adapted checkpoints and their final category profile are separate empirical observations; the probe does not establish that geometry caused that profile. Retained ranks and achieved energies were not preserved for the threshold sweep, so identical selected sets do not establish that every energy target was attained.

Historical algorithm. Geometry-guided compression and adaptation. The following pseudocode specifies the recovered estimator, including its retained oversampling directions. Here n_l is the number of vectors after subsampling and $p_l = \min(140, n_l, d)$. The SVD generator is reinitialized with the recorded seed for each stage, so equally shaped Gaussian projection matrices coincide across stages; stage subsampling advances a separate seeded random stream.

```
Inputs: model F, calibration C, knowledge corpus D, deletion budget k=15
For every hidden-state stage l:
    collect unpadded token vectors; cap pool at 9000
    subsample at most 3000 vectors -> E_l (float32)
    A_l <- float64(E_l - column_mean(E_l))
    draw Gaussian Omega_l with d rows and p_l columns
    Q_l <- thin_QR(A_l Omega_l)
    (_, sigma_l, V_l^T) <- reduced_SVD(Q_l^T A_l)
    r_l <- first r with sum(sigma_l[1:r]^2) / ||A_l||_F^2 >= 0.90
    if no such r exists, r_l <- number of returned singular values
    U_l <- first r_l columns of V_l
For each block l:
    s_l <- sqrt(max(0, r_l+r_(l+1)-2||U_l^T U_(l+1)||_F^2))
R <- first k block indices sorted by (s_l, l)
```

```
Delete blocks in R; preserve retained order and reindex attention layers
Fit LoRA updates on D; merge updates for inference
Return adapted compact model and recorded selection/evaluation metadata
```

The historical selector reads the model's `output_hidden_states` tuple. Whether its last entry includes final normalization is unresolved for the historical library revision; the last score should therefore be understood as a returned-stage transition rather than a verified isolated final-block effect. The preserved removal set is unchanged.

The historical implementation uses a denominator floor of 10^{-12} for the energy fraction and retains at least one direction in a degenerate zero-variance case. Such a case has no identifiable principal direction; the reported diagnostic is intended for nondegenerate activation samples. The randomized estimator has no power iterations and returns all oversampled components. For stage size $n_l \times d$ and projection width $p_l \ll \min(n_l, d)$, its leading dense-matrix cost is $O(n_l d p_l + (n_l + d)p_l^2)$. Computing a neighboring score from the thin bases costs $O(d r_l r_{l+1})$ and avoids explicitly storing projectors. These are algorithmic costs, not measured calibration runtimes.

H.2 Domain adaptation

After deletion, the historical branch performs causal-language-model continued pretraining on the target corpus using LoRA updates (Hu et al., 2021). The retained model weights are frozen while rank-32 updates, scaling parameter 64, and dropout 0.05 are applied to attention and MLP projections. The final adapters are merged before evaluation. This avoids adding a separate bridge module to the deployed configurations in the historical compact-model comparison.

The recorded corpus comprises approximately 498 million tokens from 16,980 files. Training code uses 512-token chunks, excludes chunks shorter than 64 tokens, and reserves up to 500 chunks for validation. The recorded schedule uses a learning rate of 5×10^{-5}, cosine decay, 200 warmup steps, and effective batch size 16. All three evaluated adapters correspond to step 61,201, recorded as the end of epoch one. These details describe the recovered training procedure; the frozen final evaluation does not establish byte-identical historical training splits across branches.

The adaptation objective uses the available domain knowledge. An initial set of 235 evaluation questions was appended to the training material without answers. This exposure is disclosed because the evaluation cannot be described as an independently held-out generalization test. It remains a common comparison of the four specified configurations on a fixed workload.

H.3 Historical exploration of learned interfaces

Historical design exploration considered a learned interface across a removed span alongside adaptation of the retained model. Historical experiments explore a residual feed-forward bridge fitted to activation pairs, followed by short LoRA adaptation of the surrounding model. This is a separate experimental branch, with different training budgets and evaluation records. Appendix B reports its supported findings. The main three adapted configurations require no such module, and their reported inference results are not attributed to bridge training.

H.4 Configurations and common evaluation

We evaluate DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI et al., 2025) and three compressed configurations. The contiguous reference removes blocks 29 through 43. The geometric configuration uses Section H.1's removal set. The historical BI configuration removes blocks $\{1, 2, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 37, 41, 42\}$. A historical directory name, `uniform`, refers to this BI configuration; it does not denote evenly spaced deletion.

All four evaluations use base-model revision `1df8507178afcc1bef68cd8c393f61a886323761`, a common tokenizer and 1,118 questions, bfloat16 precision, MPS with SDPA attention, batch size one, and greedy generation with a maximum of 300 new tokens. The recorded repetition penalty is 1.1, temperature is 1.0, and seed is 42. Temperature does not introduce sampling under greedy decoding. Runs are sequential on the same shared Apple Silicon machine. Preserved runtime records report 128 GiB physical memory, 18 logical CPUs, six PyTorch threads, macOS 26.5.1 arm64,

Python 3.12.13, PyTorch 2.12.0, and Transformers 5.8.1. The exact chip model is not established by these records.

The full model receives no additional domain adaptation. Thus, the comparison measures the complete smaller-and-adapted deployment configuration against the original model. It does not isolate the causal contribution of pruning from that of domain training. The three compressed configurations provide same-depth references, with the historical training-provenance qualification described above.

The workload contains 235 records tagged `v1_curated` and 883 tagged `v2_steelbench`; 1,088 questions are in English and 30 in Russian. Records include topic, prohibited-phrase, and numerical references where applicable. The recovered expansion generator constructs 409 composition-recall, 140 strength-recall, and 300 composition-to-property items from SteelBench records, plus 34 scripted refusal traps. It restricts numerical references to composition in wt%, strength in MPa, and temperatures in degrees Celsius, excluding mixed-scale hardness. For grouped strength values, it uses a median point when the relative range is at most 15% and a min–max interval otherwise; named-composition items require relative spread at most 10%. The final builder resolves 20 colliding identifiers, moves non-numerical grade names into topic references in 20 items, and adds refusal-marker rubrics to 25 initially empty trap items.

Replaying these preserved scripts with the recovered source table produces the same 883 expansion records and the same 1,118 merged records, including reference fields, as the frozen evaluation input. This verifies construction consistency, not independent factual validation of the upstream records. A complete human annotation-validation history and an independent sampling rationale for the mixture are not established. The 883-record source tag does not by itself establish held-out status. Question-weighted results describe this fixed mixture.

H.5 Measurements and scoring

Generation throughput is total generated tokens divided by summed device-synchronized generation-call durations. It includes work performed within generation, including prefill, and excludes loading, adaptation, and output-file writing. This is an observed batch-one throughput measurement, not an isolated decode-only kernel benchmark. No replicated timing confidence interval, energy measurement, or minimum-memory-device demonstration is reported.

Each question's composite averages the scoring components present in its reference. Topic matching checks whether any whitespace-delimited word from an expected topic occurs as a substring in the generated text. Prohibited phrases incur literal-match penalties of 0.5 each, floored at zero. Numerical matching extracts numbers and checks reference point or interval criteria: point error is divided by $\max(|y|, 1)$ and must be below 0.15; interval bounds receive the implemented 0.98/1.02 multipliers. The exact scorer is included with the snapshot.

This score measures agreement with the fixed rubric. A topic substring or numerically matching value need not establish a correct factual proposition. The evaluator scores the entire generated prefix, including any reasoning text, rather than requiring a completed final-answer segment. Stored per-question composites are rounded to three decimals before aggregate reporting. Component analyses and paired comparisons additionally recompute the unrounded rubric values.

H.6 Evidence preservation

The final comparison contains 4,472 responses: 1,118 for each configuration. The immutable evaluation snapshot includes responses, questions, configuration identities, runtime code, environment lock, aggregate results, and completion checks, protected by a SHA-256 manifest. A separate snapshot preserves historical method evidence and labels unresolved executable revisions explicitly. This separation prevents exploratory measurements from being presented as additional runs of the final protocol. Appendix C maps the evidence to the manuscript.

H.7 Model size and inference throughput

We evaluate the full 48-layer DeepSeek-R1-Distill-Qwen-14B model alongside three compressed, domain-adapted variants obtained using contiguous, Grassmannian, and Block-Influence (BI) layer selection. Each compressed variant removes 15 layers and uses its final continued-pretraining

checkpoint at step 61,201; the full model is evaluated without additional domain training. All four configurations use the same 1,118 questions, base-model revision, tokenizer, and generation settings. Evaluation uses bfloat16 precision, batch size one, and greedy generation capped at 300 new tokens on the same Apple Silicon machine. We measure generation throughput as the total number of generated tokens divided by the summed, device-synchronized generation time, excluding model loading and adaptation.

Compression reduces depth from 48 to 33 layers (31.25%) and parameter count from 14.77 to 10.64 billion (27.96%). The full model generates 11.92 tokens per second. The contiguous, Grassmannian, and BI variants achieve 18.40, 16.03, and 17.27 tokens per second, respectively, corresponding to throughput ratios of 1.54, 1.35, and 1.45 relative to the full model. Thus, all three compressed configurations provide higher observed generation throughput while using fewer parameters. The Grassmannian-selected model achieves this reduction with a layer set determined by changes in activation subspaces. These measurements establish the size and generation-speed benefits in the evaluated setting; the runs were conducted sequentially on a shared machine.

Supplementary table S3. Model size, aggregate domain score, and observed generation throughput. Compressed models use their final checkpoint at step 61,201. The full model has no additional domain training. Throughput is measured from synchronized generation calls in sequential runs on the same machine.

Configuration	Layers	Parameters (B)	Composite score	Generated tokens/s	Throughput ratio
Full model	48	14.770	0.5180	11.92	1.00×
Contiguous + CPT	33	10.641	0.5503	18.40	1.54×
Grassmannian + CPT	33	10.641	0.5668	16.03	1.35×
BI + CPT	33	10.641	0.5846	17.27	1.45×

I Historical Workload Scores and Controls

These archived prefix-rubric results are supporting diagnostics, not the central selection endpoint or semantic accuracy. The complete profiles and counterchecks are retained to prevent favorable subsets from standing in for answer quality.

I.1 Generated-prefix scores after compression and adaptation

All three compressed, adapted configurations exceed the unadapted full model on the overall domain composite score under the common evaluation protocol. The full model scores 0.5180, compared with 0.5503 for contiguous selection, 0.5668 for Grassmannian selection, and 0.5846 for BI selection: absolute increases of 0.0323, 0.0488, and 0.0666, respectively. Supplementary table S3 reports these scores over the same 1,118 questions using the fixed scoring function and generation budget. Together with the size and throughput measurements, this comparison identifies three configurations that combine fewer parameters and faster observed generation with a higher aggregate domain score. The result evaluates the complete compression-and-adaptation procedure, including continued pretraining on the target knowledge corpus.

At the same depth and final training step, the evaluated configurations exhibit different combinations of domain score and generation throughput. BI yields the highest aggregate score, while contiguous selection yields the highest observed throughput. The Grassmannian variant scores above the contiguous variant and below BI, with a throughput of 16.03 tokens per second versus 11.92 for the full model. Thus, geometry-guided selection produces a smaller, adapted model with higher fixed-budget heuristic rubric scores and observed token throughput than the original configuration. These prefix scores do not establish improved completed-answer quality. The comparison also retains the alternative selection strategies as explicit reference points for assessing the contribution of the geometric criterion.

I.2 PREFIX-RUBRIC PROFILES ACROSS SELECTION STRATEGIES

The category breakdown shows how the alternatives distribute performance across the evaluated workload (Supplementary table S4). Contiguous selection has the highest score on domain-knowledge questions, reaching 0.7361 compared with 0.5087 for the full model. Grassmannian selection has the highest score on the multistep category: 0.5820, compared with 0.3872 for the full model, 0.2925 for contiguous selection, and 0.4417 for BI. All three compressed variants also score above the full model on the Russian-language category. The full model retains higher scores on several other categories, including property reasoning and heat treatment. Category names identify groups of benchmark questions; their scores use the same composite rubric as the aggregate evaluation.

Question-level comparisons provide a complementary view of these profiles. Relative to the full model, the Grassmannian variant scores higher on 541 questions, equally on 226, and lower on 351. The corresponding counts are 456/300/362 for contiguous selection and 522/279/317 for BI. These comparisons use unrounded per-question scores. Grassmannian selection therefore improves the score on more individual questions than either alternative, while BI achieves the largest average improvement. The difference follows from the magnitudes of the score changes as well as their frequency.

The numerical component of the scoring function further distinguishes the configurations. Among the 850 questions with numerical reference values, its mean is 0.3800 for the full model, 0.3353 for contiguous selection, 0.3965 for Grassmannian selection, and 0.3529 for BI. The Grassmannian variant is the only compressed configuration above the full model on this component. This measure records whether generated numbers match the supplied point or range criteria; it does not independently assess whether each number is attached to the correct factual claim. Together, the category, paired-question, and component results describe the measured profile of geometric selection alongside the alternatives.

Supplementary table S4. Category-level composite scores. Category means use the stored per-question scores, rounded to three decimal places before aggregation, consistent with the main evaluation.

Category	Questions	Full model	Contiguous + CPT	Grassmannian + CPT	BI + CPT
Applied scenarios	20	0.8539	0.8304	0.8360	0.8131
Domain knowledge (GOST)	439	0.5087	0.7361	0.6382	0.6856
Property reasoning	170	0.7294	0.6588	0.4721	0.6853
Cross-standard questions	20	0.6584	0.6001	0.6585	0.6127
Heat treatment	25	0.7107	0.5366	0.4907	0.5620
Hallucination-trap questions	59	0.6161	0.4266	0.4125	0.5087
Russian-language questions	30	0.0700	0.2034	0.2233	0.2227
Multistep questions	320	0.3872	0.2925	0.5820	0.4417
Specification compliance	20	0.7003	0.5335	0.5001	0.5668
Exploration	15	0.4933	0.1693	0.2553	0.2747

The aggregate score weights each question equally. Domain-knowledge and multistep questions account for 759 of the 1,118 items (67.9%), so improvements on these groups have a substantial effect on the overall result. As a descriptive sensitivity comparison, weighting the ten categories equally instead gives means of 0.5728 for the full model, 0.4987 for contiguous selection, 0.5069 for Grassmannian selection, and 0.5373 for BI. The higher question-weighted scores of the compressed configurations thus coexist with category-specific trade-offs. Reporting both aggregations makes the evaluated workload explicit while preserving the common protocol across models.

All 1,118 generations from each compressed configuration reach the 300-token limit; 1,111 of the full model's generations reach that limit. The reported scores therefore characterize the generated text within the fixed budget. Within this protocol, the completed experiment demonstrates three smaller,

domain-adapted configurations with higher overall domain composite scores and higher observed generation throughput than the original full model. Their different performance profiles provide the empirical basis for discussing how structural layer selection and target-corpus adaptation should be used together.

I.3 Discriminative controls and explicit numerical assertions

The original rubric gives an unanswered-question echo a composite of 0.4830 overall and 0.5768 in the multistep category, compared with the geometric configuration's 0.5820 in that category. No model is involved in this control: each question is supplied to the scorer as its own response. Topic words already occur in the prompt, and a temperature can numerically match a strength reference because the scorer does not bind values to properties or units. This control does not establish that actual model outputs are echoes. It shows why a category advantage under that rubric alone is insufficient evidence of answering the requested task.

We therefore add a separate, retrospective measurement on all 849 numerical questions from the generated workload, covering 1,127 reference slots. A conservative parser extracts explicit property–number–unit constructions for elemental composition, tensile strength, and yield strength. It converts supported strength units to MPa, removes verbatim questions and interrogative segments, excludes recognized hypothetical/negative constructions, and rejects numeric operands immediately followed by arithmetic. A slot matches only when at least one recognized assertion exists and every recognized assertion for that property agrees with the reference. Point references use 15% relative error with a near-zero denominator floor; interval references retain the original 2% bounds expansion. This changes both extraction and, for small composition values, tolerance; it is not an isolated ablation of unit binding.

The parser and constructed controls were frozen before computing the first supplementary four-model comparison. Inspecting its outputs exposed two extraction defects: accepting an arithmetic operand and missing an intervening entity clause such as "strength of steel grade … is …". Version 2 addresses these defects and was frozen before recomputing the four-model results. Version 1 outcomes were already known; both versions, specifications, hashes, and every extraction are retained. Neither version was selected by a requirement that geometry rank first.

The controls include correct and incorrect values, wrong properties and units, conflicting recognized claims, intervals, unit conversion, prompt echo, and naked reference-number lists. Version 2 passes 21 constructed cases; on all 849 items, both prompt echo and bare reference-number lists produce zero supported slots, while synthetically formatted reference assertions reach all 1,127 slots. These checks establish grammar behavior and reachability, not complete semantic validity on natural responses.

Supplementary table S5. Supplementary explicit property–unit numerical matches. Every row uses the same 1,127 reference slots. Coverage is the number of slots with a recognized assertion. V1 and V2 use the same population and numerical criteria; their extraction grammar differs. These are parser-mediated counts, not factual-accuracy estimates.

Configuration	V1 matching slots	V2 extracted slots	V2 matching slots	V2 matches / all slots
Full model	17	63	17	1.51%
Contiguous + CPT	14	611	81	7.19%
Grassmannian + CPT	69	451	77	6.83%
BI + CPT	6	112	26	2.31%

All three compact configurations have more recognized matching slots than the original model under V2. The scenario breakdown specifies where these observations occur (Supplementary table S6). Contiguous selection has the largest overall count and strongest named-composition result, while geometric selection has the largest count on composition-to-property questions. This complements the original profile analysis with explicit property and unit associations under the same generation budget.

Supplementary table S6. V2 matching / extracted slots by task. Fixed denominators are 409 slots for composition recall, 140 for strength recall, and 578 across the 300 composition-to-property questions. Lack of extraction is not proof of an incorrect answer.

Configuration	Composition recall (409)	Strength recall (140)	Composition-to-property (578)
Full model	6 / 24	11 / 39	0 / 0
Contiguous + CPT	46 / 401	24 / 138	11 / 72
Grassmannian + CPT	6 / 32	6 / 24	65 / 395
BI + CPT	8 / 41	12 / 17	6 / 54

Coverage is highly unequal, and the contiguous count changes from 14 to 81 between parser versions, changing which model has the largest overall matching count. This demonstrates sensitivity to output phrasing; these results cannot establish a robust quality ranking. The parser does not resolve every material identity, anaphoric reference, table layout, qualification, or contradiction outside its recognized constructions. It also operates on the generated prefix, including reasoning text, without certifying a completed final answer. Consequently, matching counts are neither independently validated factual accuracy nor a rigorous lower bound on it. They supply an inspectable diagnostic beyond unbound number matching; deployment-quality claims still require a validated task-level endpoint. The positive compression result and these explicit-assertion observations should be interpreted together, without replacing the historical rubric or extrapolating either metric beyond its coverage.

I.4 Structural evidence for the selection criterion

The historical single-block intervention experiment tests the diagnostic before the combined deletion-and-adaptation result. Each of the 48 blocks is removed separately, and damage is defined as the increase in gold-answer NLL over the full model. The recovered helper defines 14 fixed, short metallurgy question–answer pairs. It appends each gold answer to the templated prompt, computes its mean token NLL under teacher forcing, and averages the 14 question-level losses equally. Thus, the 48 observations are block interventions on the same small fact set. The helper sources and questions are preserved in a separate supplement; their exact historical execution revision remains unresolved. Blocks are grouped by geometric score, independently of the measured damage. Supplementary table S7 reports the recovered measurements and descriptive sensitivity calculations from the saved per-block records.

Supplementary table S7. Historical single-block deletion damage. Negative values denote lower gold-answer NLL after the individual deletion. Groups contain separate one-block interventions, not simultaneous deletions; the recovered evaluation protocol uses the same 14 QA pairs for each intervention. The low/high groups use this probe's own calibration and need not match the final 15-block removal set.

Group ordered by geometric score	Blocks	Mean NLL change	Median NLL change
Lowest 15	15	−0.0124	−0.0132
Highest 15	15	0.7547	0.0730
Highest 15, excluding block 0	14	0.0900	0.0698
All blocks	48	0.2518	0.0364

The geometric score has a Spearman correlation of 0.3247 with intervention damage. Removing block 0 produces an unusually large NLL increase of 10.0611 and strongly affects the high-group mean. Excluding this block leaves a positive correlation of 0.2806; the fixed high group's remaining mean is 0.0900, compared with −0.0124 for the unchanged low group. The median comparison has the same ordering. Thus, the association is not solely due to the block-0 outlier, although its magnitude is modest and individual exceptions remain. These findings support the criterion as a screening signal for deletion candidates. They do not establish additive effects when several blocks are removed.

This diagnostic result and the final comparison answer consecutive questions: whether the signal is associated with individual deletion sensitivity, and whether a model constructed from geometry-selected deletions produce the recorded resource and prefix-score profile after adaptation. Their distinct calibration and evaluation protocols prevent interpreting them as a single controlled causal experiment.

J Retrospective Semantic Reading

One AI evaluator read all 120 saved responses for 30 questions shared by the original model and three historical compact+CPT configurations. Before reading, three question IDs per each of ten categories were selected by a fixed hash rule. Model labels and previous rubric scores were hidden in the review packet, but style and prior familiarity could reveal identities. This is limited masking, not independent human validation. All texts have the original 300-token budget. A whole-response ordinal scale records 0 (central error or does not solve the task), 1 (partly useful with substantive errors or omissions), 2 (substantively correct with acceptable gaps), and 3 (fully adequate); unresolved cases are kept separate. EOS is not required for credit.

Supplementary table S8. Semantic reading of saved prefixes. Each row covers the same 30 questions. No response receives 3; the 2–3 column therefore contains only score-2 responses.

Configuration	Scores 2–3	Score 1	Score 0	Unresolved
Original	4	23	3	0
Geometry + CPT	0	9	20	1
BI + CPT	0	9	17	4
Contiguous + CPT	1	8	19	2

A separate retrospective annotation, added after reading but before revealing model names, records whether the core response to the original question is correct, partially correct, incorrect, absent, or unresolved. Correct core content occurs in 7 original, 1 geometry+CPT, 0 BI+CPT, and 3 contiguous+CPT responses. This preserves correct answers followed by faulty continuations; it is not a second independent evaluation. In the paired whole-response scores, geometry is higher than the original on one question, lower on 19, tied on nine, with one unresolved pair. This slice does not support preserved answer quality for the historical geometry checkpoint under this protocol.

Category balancing changes the mixture substantially: three categories account for 929 of 1,118 workload items but only nine of these 30 questions. Counts here are consequently not estimates of whole-workload accuracy. The token cap also truncates the original model, and unresolved material conditions or references are not treated as model errors. The full texts, rubric, question selection, source checks, quotations, and annotations are retained in `revisions/semantic-review-20260925/`. Replaying the summary verifies its arithmetic, not independent correctness of the AI judgments. These observations motivate separating reference-loss diagnostics, direct answer content, and turn termination.

K Analytical Illustration

$$x = Me_1 + ze_2, \quad y = Me_1 + ze_3, \quad z \in \{-1, +1\}$$

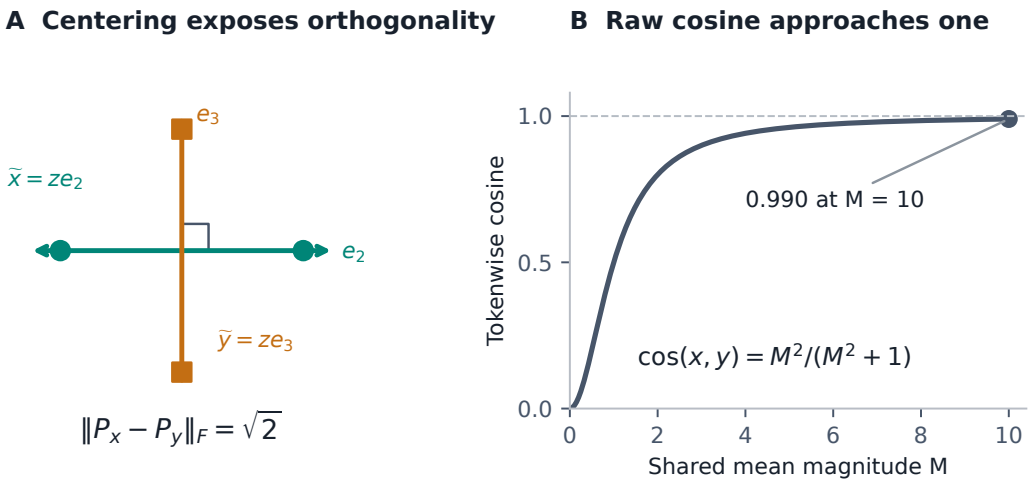


Figure 7: The exact construction from Section 3.3. For $x = Me_1 + ze_2$ and $y = Me_1 + ze_3$, centering exposes orthogonal rank-one subspaces while raw paired-token cosine tends to one. This illustrates a distinction between statistics, not a deletion-safety theorem.

L HISTORICAL PROCEDURE AND RESOURCES

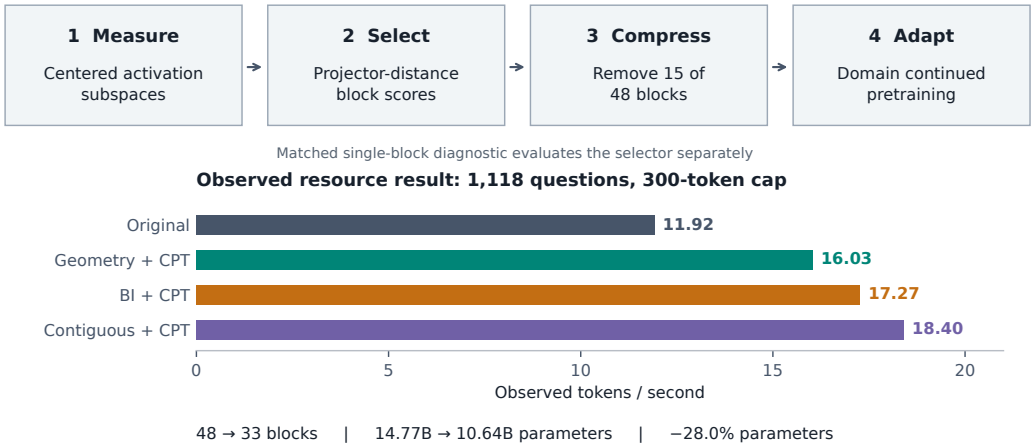


Figure 8: The historical pruning-and-CPT branch and its resource measurements. The matched individual-block experiment evaluates selection separately; its set differs by one block. Throughput uses 1,118 questions with a 300-token cap and is not an answer-quality endpoint.