THE COST OF EIGENSPACE RESTRICTION IN SHARED ATTENTION CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Soft prompts and prefixes control a frozen Transformer through learned records: a record's key decides when attention retrieves it, and its value decides what it returns. Confining records to a low-dimensional subspace cuts their parameters, and the value map's highest-gain eigenspace is the natural spectral choice. We show that this choice can require exponentially more records. In a two-head scalar attention interface with fixed projections and readout, we count the equal-weight records needed to approximate a growing-amplitude target $m \tanh q$, equivalently a fixed target under growing attention gain. At every fixed relative tolerance we determine the optimal exponential count rate exactly, both for unrestricted records and for the eigenspace restriction, and the restricted rate is strictly larger below an explicit tolerance. The converse is an entropy-area bound: equal-weight records pay for how far attention moves from uniform. Because the restriction ties every value to its key, attention must keep moving across the query interval, whereas unrestricted records can saturate retrieval early and let a palette of values supply the rest; counted constructions attain both rates. The loss persists for thin full-dimensional strips around the eigenspace, nearby fixed-key value maps and, under a disk-margin condition, an RMSNorm realization, and computer-certified explicit libraries exhibit it at moderate amplitude in four parameter settings. In frozen RoBERTa-base, one unrestricted shared key/value record gives lower SST-2 held-out loss than up to 256 subspace-confined records, and the strongest confined arms stay above it with a fivefold training budget; on AG News, at the same block, the differences are at most 0.021 nats. In the scalar interface, finite optimization favors unrestricted records at every tested count under uniform queries, with reversals under some query distributions.

1 INTRODUCTION

Parameter-efficient adaptation often controls a frozen Transformer by adding learned records, such as soft-prompt vectors that pass through the model's fixed projections (Lester et al., 2021) or trained prefixes at every layer (Li & Liang, 2021; Liu et al., 2022). A record's key determines which queries retrieve it, and its value determines what it returns; when fixed projections derive both from the same vector, the two roles are coupled, a constraint that has motivated attention-level prompts (Wang et al., 2023a). Prompting theory establishes approximation capabilities and limitations under several architectural choices (Wang et al., 2023b; Oymak et al., 2023; Petrov et al., 2024b;a; Hu et al., 2025). We ask a complementary resource question: how does restricting the allowed records change the number of records needed to control an attention response?

Confining records to a low-dimensional subspace is a natural way to cut their parameters: fine-tuning and prompt tuning retain much of their effect inside low-dimensional subspaces (Aghajanyan et al., 2021; Qin et al., 2024), and low-rank projections of cached keys and values compress attention memory (Saxena et al., 2024; Chang et al., 2025). For attention records the natural spectral choice is the value map's top eigenspace, where a record returns the largest value per unit norm before attention weighting. We study it as the choice most favorable to returned value, not as an existing method. It keeps value gain but removes the freedom to change one head's value while holding that head's key fixed. We show that this is expensive: in a specified shared-record interface, at every tolerance below $\tanh(Q)/2$, the restriction strictly increases the exponential rate at which the required number of records grows.

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.

The interface is deliberately small (Section 2). A record is a point $(x, y) \in [-m, m]^2$; two heads with a shared scalar query q use x and y as keys, a fixed coupled value map with top eigenvector $(1, -1)$ and eigenvalue $A = 1 + 2\epsilon$ produces their values, and the output averages the heads. The restriction confines every record except two fixed anchors to $y = -x$. The target is $m \tanh q$ on $[-Q, Q]$ at relative tolerance δ, equivalently a fixed target under attention gain m. Because records have equal weight, attention is uniform at $q = 0$ and can move at most $\log N$ in relative entropy from uniform; at small tolerance, tracking the target requires movement linear in m, so the minimum count grows exponentially and the question is its rate. Counting every copy with a fixed readout defines the resource (Table 1).

Our main result determines both optimal exponential rates, the limits of $m^{-1} \log N$, exactly at every fixed tolerance $\delta > 0$ (Theorems 1–3). A subset can never need fewer records; the content is the exact price of the restriction, from a matching converse and counted construction for each family. As $\delta \downarrow 0$ the unrestricted and restricted rates tend to $J(t/A)$ and $J(t)/A$, respectively, where $t = \tanh Q$ and $J(s)$ is the relative entropy of $\mathrm{Bern}((1 + s)/2)$ from $\mathrm{Bern}(1/2)$. The mechanism is an entropy-area budget (Figure 1c): the relative entropy attention has spent by query q equals the area between the level of the current mean retrieved key and its path up to q. On the eigenspace every record returns A times its key, so the normalized retrieved key must follow $\tanh(q)/A$ across the whole interval. Unrestricted records reach the same endpoint t/A along the steeper path $\tanh q$ and stop, while a palette of records with nearly equal keys but different values supplies the rest of the response. The steeper path encloses less area, and strict convexity of J makes the gap strict.

The loss is not an artifact of exact confinement, rank or the exact value matrix: it survives small attention-weighted deviations, thin strips with two-dimensional interior and full-rank Fisher information, nearby fixed-key value maps whose top singular direction rotates, and anchor changes, and both rates transfer to an RMSNorm realization under a disk-margin condition (Sections 3.3 and 4). Nor is it only asymptotic: at $Q = 1$, $\epsilon = 1/4$ and $\delta = 0.05$, explicit unrestricted libraries with computer-certified error have fewer records than the finite converse permits for any restricted library at every tested amplitude from $m = 70$ to 200, by factors from 1.4 to 450, and three further settings of (Q, ϵ) give certified separations as well.

The same resource comparison is measurable in trained models (Section 5). In frozen RoBERTa-base with shared key/value records at one block, a single unrestricted record lowers SST-2 held-out cross-entropy from 0.438 to 0.260, a level that no arm confined to the top 16- or 64-dimensional value/output eigenspace, or to a random subspace of the same dimension, reaches with up to 256 records, despite up to 21 times as many trainable coordinates; with five times the training budget the strongest confined arms converge but stay above it. On AG News one unrestricted record is again best at that block, but only by 0.006 nats over the best restricted arm, and Value-16 records are better at a later block. In the scalar interface, finite optimization lowers held-out RMSE by a median 49% at the primary count (12–40% at the other counts) over ten fresh pairs, with distribution-dependent reversals.

Our contributions are:

1. **Exact rates at positive tolerance**, with a strict restriction loss for $0 < \delta < \tanh(Q)/2$ and a certified separation at finite amplitude (Section 3).
2. **A mechanism with matching constructions**: an entropy-area converse for arbitrary libraries, matched by counted binomial bases, shifts and a finite palette (Section 3.2).
3. **Quantified robustness** to transverse deviations, strips, value-map perturbations, anchors and normalization (Sections 3.3–4).
4. **Evidence in pretrained encoders and finite optimization**, reported with every adverse cell (Section 5).

The scalar model isolates a retrieval–value trade-off; it does not establish how prevalent that trade-off is in trained Transformers (Section 7).

2 THE COUNTED SHARED-RECORD INTERFACE

Fix $Q > 0$, $\epsilon > 0$, and an integer $m \geq 1$. A library is a multiset $\mu = \{(x_i, y_i)\}_{i=1}^{N} \subset [-m, m]^2$ of records. All records and maps are fixed before the query $q \in [-Q, Q]$ is supplied. Head 1 uses x_i as

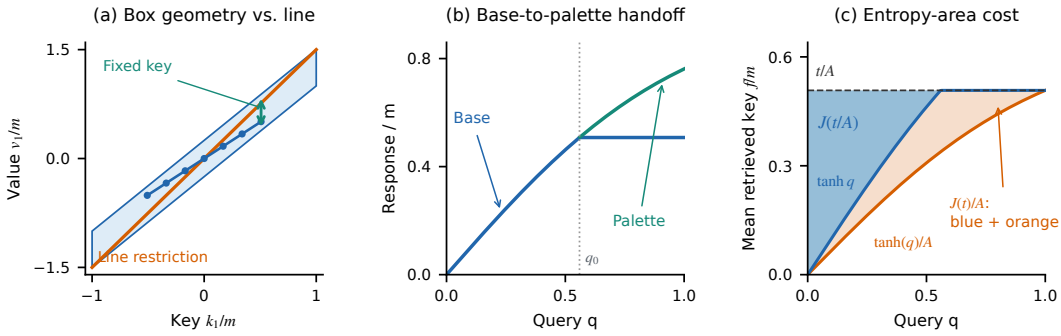


Figure 1: Retrieval–value geometry and its count cost ($Q = 1$, $\epsilon = 1/4$). (a) Head 1's key–value plane. Unrestricted records fill the parallelogram, and a vertical slice changes the value at a fixed head-1 key (while changing head 2's key); the eigenspace restriction $y = -x$ leaves the orange line $v_1 = Ak_1$. Blue dots are diagonal base records. (b) The unrestricted construction in the small-error limit: the base response saturates at t/A at $q_0 = \operatorname{artanh}(t/A)$; above this retrieval transition a palette dominates the common partition-weighted mixture and returns larger values (a limit sketch, not a fit). (c) The normalized mean retrieved key f/m along the query path. The line must follow $\tanh(q)/A$, enclosing area $J(t)/A \simeq 0.21854$ (blue plus orange); unrestricted records can follow $\tanh q$ and stop at t/A, enclosing $J(t/A) \simeq 0.13510$ (blue). The enclosed area lower-bounds $\log N/m$ by (9). Limits take $m \to \infty$ before $\delta \downarrow 0$; anchors are omitted from the drawing but retained in both theorems.

the key of record i, and head 2 uses y_i:

$$p_i(q) = \frac{e^{qx_i}}{Z_x(q)}, \qquad r_i(q) = \frac{e^{qy_i}}{Z_y(q)}, \qquad Z_x(q) = \sum_i e^{qx_i}, \quad Z_y(q) = \sum_i e^{qy_i}. \tag{1}$$

Both heads read the same record through a fixed value map with rows $(1+\epsilon, -\epsilon)$ and $(-\epsilon, 1+\epsilon)$, and the output is their scalar average:

$$F_\mu(q) = \frac{1}{2}\left(\mathbb{E}_p[(1+\epsilon)x - \epsilon y] + \mathbb{E}_r[(1+\epsilon)y - \epsilon x]\right) = (1+\epsilon)f(q) - \epsilon c(q), \tag{2}$$

where $f = (\mathbb{E}_p x + \mathbb{E}_r y)/2$ is the mean retrieved key and $c = (\mathbb{E}_p y + \mathbb{E}_r x)/2$ is the mean cross-coordinate. Two heads give a simple setting with this coupling: with head 1 alone, y would be a free value coordinate at a fixed key, whereas here it is also head 2's key, so every change of a record changes some key (Appendix H). A library is admissible at tolerance δ if it contains two designated anchor records $\pm(1,1)$ and

$$\sup_{|q| \le Q} |F_\mu(q) - m \tanh q| \le \delta m. \tag{3}$$

Additional copies of anchor coordinates are allowed and counted. There is no residual, trainable readout, query-dependent record generation, or free per-record bias; these choices specify a control problem, not an approximation of a pretrained network. Dividing coordinates by m gives the equivalent view of bounded records under attention gain m (Section 4).

The value map has orthogonal eigenvectors $(1,1)/\sqrt{2}$ and $(1,-1)/\sqrt{2}$ with eigenvalues 1 and $A = 1 + 2\epsilon$, so its highest-gain eigenspace is the line $y = -x$. Let $N_*(m,\delta)$ be the smallest admissible count, with $+\infty$ for an empty class. Let $N_{\text{anti}}(m,\delta)$ additionally restrict every nonanchor record to $y = -x$; such a multiset need not be sign symmetric. Only the two designated anchors are exempt, further copies of their coordinates are not permitted, and both anchors remain charged. The anchors lie on the other eigenvector, so the support of a restricted library with any nonzero line record affinely spans the plane and its Fisher information has full rank (Appendix H); the restricted rate is the same without them (Proposition 8). Write

$$R_-(\delta) = \liminf_{m\to\infty} \frac{\log N_*(m,\delta)}{m}, \qquad R_+(\delta) = \limsup_{m\to\infty} \frac{\log N_*(m,\delta)}{m}. \tag{4}$$

Table 1: Main rates and why the charged resources matter. The first two rows use the same fixed maps, count every equal-weight copy, and give the exact limit of $m^{-1}\log N$ at each fixed $\delta > 0$. The last two rows are elementary controls that change the interface; they are not novelty claims.

Control family	Record requirement	Regime / evidence
Unrestricted shared records	Rate $J\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right)$	Exact fixed-tolerance characterization; Theorems 1–2
Antisymmetric nonanchors	Rate $\frac{1}{A}J([t-2\delta]_+)$	Exact rate and strict loss; Theorem 3
Add a free readout gain m	Two anchors suffice	Exact box response; Appendix E.1
Store integer atom masses	$m+1$ distinct atoms	Exact for odd m; $O(m^2)$ weight bits instead of 2^m copies

Define R_-^{anti}, R_+^{anti} by replacing N_* with N_{anti} in (4). All logarithms are natural. Define $t = \tanh Q$, $[z]_+ = \max\{z, 0\}$, and

$$J(s) = s\,\mathrm{artanh}\, s + \tfrac{1}{2}\log(1-s^2), \qquad 0 \le s < 1. \tag{5}$$

This is the binary relative entropy $D_{\mathrm{KL}}(\mathrm{Bern}((1+s)/2)\|\,\mathrm{Bern}(1/2))$. It is increasing and strictly convex, with $J(0) = 0$.

What the count comparison means. Record count is the analog of prompt length. Equal N does not equalize adjustable parameters: the restriction removes degrees of freedom, and this loss is the object of study. Equal weights make count the resource: with free integer masses shared by both heads (equivalently, free logit biases), $m+1$ weighted diagonal atoms represent the unrestricted target exactly for odd m (Table 1), so both exponents concern the bias-free regime, where a record's weight is set by its key alone. In a d-dimensional model a key component along a direction shared by all queries would act as such a bias; the scalar query has none. $O(\log m)$-bit coordinates do not imply logarithmic total storage when the number of records is exponential, and independent key–value controls or direct updates are different interfaces (Appendix E).

3 Sharp count rates and a robust geometric loss

Theorem 1 (Unrestricted finite converse)**.** *Any equal-weight library satisfying* (3)*, whether or not it contains anchors, obeys*

$$\log N \ge mJ\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right) - \tfrac{1}{2}\log 2. \tag{6}$$

No coordinate correlation or symmetry of the original library is assumed.

Theorem 2 (Exact unrestricted rate at positive tolerance)**.** *For every fixed $\delta > 0$,*

$$R_-(\delta) = R_+(\delta) = J\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right). \tag{7}$$

For each fixed rational $0 < d < \min\{\delta, t/(A+1)\}$, counted constructions with rational $O(\log m)$-bit coordinates achieve count coefficient $J((t-(A+1)d)/A)$ and eventually satisfy tolerance δ.

Theorem 3 (Exact restricted rate and strict cost)**.** *For every fixed $\delta > 0$,*

$$R_-^{\mathrm{anti}}(\delta) = R_+^{\mathrm{anti}}(\delta) = \frac{1}{A}J([t-2\delta]_+). \tag{8}$$

This coefficient is strictly larger than (7) *for every $0 < \delta < t/2$. For $\delta \ge t/2$, both coefficients are zero. As $\delta \downarrow 0$, the unrestricted and restricted coefficients tend to $J(t/A)$ and $J(t)/A$, respectively.*

Reading the theorems. The theorems give exact limits of $m^{-1}\log N$ at each positive tolerance, not finite counts; for $0 < \delta < t/2$ the ratio N_{anti}/N_* grows like $\exp\{m(R^{\mathrm{anti}} - R + o(1))\}$, where R^{anti} and R are the limits in (8) and (7). At $Q = 1$, $\epsilon = 1/4$ and $\delta = 0.05$ the unrestricted and restricted coefficients are 0.0930 and 0.1590. A zero coefficient does not establish bounded or polynomial count. The constructions fix a strict error slack before $m \to \infty$, so their constants can diverge as the slack vanishes (Appendices A, B and I).

The separation at finite amplitude. The literal proof recipe has impractical constants (Appendix D), but the construction itself works at moderate amplitude. For $Q = 1$, $\epsilon = 1/4$ and $\delta = 0.05$ we build unrestricted libraries as in Section 3.2, with integer records, palettes chosen by integer programming and a computer-certified uniform error below δm. At every tested amplitude from $m = 70$ to 200 they have fewer records than the finite restricted converse behind Theorem 3 (Appendix B, Step 9) permits for any restricted library, by factors from 1.4 to 450, using at most twelve palette orbits; at $\delta = 0.1$ this holds from $m = 60$. Three further settings, $(Q, \epsilon) = (2, 1/4)$, $(1, 1/2)$ and $(1, 1/8)$, give certified separations at both tolerances from $m = 40$, $m = 50$–70 and $m = 80$–120, respectively (Appendix P). The counts are certified upper bounds on N_*, not optimal values.

3.1 Why the restriction costs records: an entropy-area budget

For $\Psi(s_1, s_2) = \log \sum_i e^{s_1 x_i + s_2 y_i}$, the Hessian is the covariance of (x, y), the exponential family's Fisher matrix (Nielsen, 2020). With u_N uniform on the counted records and $\Phi = (\log Z_x + \log Z_y)/2$, the exact identity is

$$\frac{D_{\mathrm{KL}}(p_q \| u_N) + D_{\mathrm{KL}}(r_q \| u_N)}{2} = \int_0^q [f(q) - f(s)]\, ds = \int_0^q s\Phi''(s)\, ds \le \log N, \quad q \ge 0. \tag{9}$$

Here $\Phi' = f$ and $\Phi'' = (\mathrm{Var}_p\, x + \mathrm{Var}_r\, y)/2$. The left side is the relative entropy that attention has spent by query q; the middle is the area between the level of the current mean retrieved key $f(q)$ and its path over $[0, q]$. Because attention starts at uniform, this area never exceeds $\log N$.

The two families differ in how fast this path may rise. For a centrally symmetric library, relative entropy and convexity give $f \le F \le Af$ (Appendix A). Accuracy therefore caps the path at about $m \tanh s$ while forcing its endpoint to about mt/A, so the area is at least about $mJ(t/A)$; the construction below attains this by following $m \tanh s$ and then stopping. On the eigenspace, both heads return A times the key of every nonanchor record, so $F = Af - D$ with an anchor correction $0 \le D \le A - 1$ for $q \ge 0$. The path must then stay near $m \tanh(s)/A$ over the whole interval, and the area is at least about $(m/A)J(t)$. Figure 1c draws both paths. The model-specific work is forcing these areas with exact tolerance bookkeeping and realizing matching libraries.

3.2 Converse and construction mechanisms

Converse for arbitrary libraries. An arbitrary library lacks the central symmetry used above, and adding its sign reflection directly would give the two heads different mixture weights, so their errors need not combine convexly. Instead, form all ordered-pair records $(x_i + y_j, y_i + x_j)$. Both normalizers factor as $Z_x Z_y$ and the new output is $2F_\mu$. Adding sign reflection then produces $2N^2$ central records approximating $2m \tanh q$, with one common mixture weight for both heads. For this auxiliary library, target accuracy forces an area of at least $2mJ(b_-)$, with $b_- = [(t - (A+1)\delta)/A]_+$, and equation (9) gives Theorem 1. This changes amplitude and count; it is not a same-budget symmetry reduction. The restricted converse needs no symmetrization, because the identity $F = Af - D$, with $0 \le D \le A - 1$ for $q \ge 0$, holds for every restricted library.

Matching constructions. A truncated binomial base at cutoff β has count exponent $J(\beta)$ and normalized response approaching $\min\{\tanh q, \beta\}$ for positive queries. Replacing each scalar key k by $k \pm a$, with both multiplicities charged, gives

$$\widetilde{Z} = 2\cosh(aq) Z_B, \qquad (\log \widetilde{Z})' = \mu_B + a \tanh(aq).$$

The count only doubles, while a shift of order m uses the allowed error near zero and spends at most $\log 2$ additional entropy area; we bound the non-uniform transition near zero directly.

On the restricted line, use $n = \lfloor m/A \rfloor$, cutoff $\beta = t - 2d$, and shift $a = dn$; its output divided by m stays within $d + o(1)$ of the target. Unrestricted records use cutoff $\beta = (t - (A+1)d)/A$ and a diagonal shift dm, so the leading key becomes $\lambda m + O(1)$, where $\lambda = \beta + d = (t - d)/A$. A counted palette of sign/exchange orbits, with keys at fixed offsets from this leading key and normalized values spread between about λ and $A\lambda$, dominates above the retrieval transition and approximates the target within the remaining slack. Below the transition the base dominates, on the overlap both are accurate, and their outputs form a common partition-weighted mixture, not an additive correction. Taking

the infimum over fixed $d < \delta$ proves the sharp rates. The successful library is invariant under sign reversal and coordinate exchange, so what the restriction imposes is pointwise confinement, not a different symmetry.

3.3 THE LOSS IS NOT A DIMENSION COUNT

Proposition 4 (A full-dimensional strip can remain costly)**.** *Require every nonanchor to satisfy $|x + y| \le \tau m$, keeping the same two anchors. Every accurate library obeys*

$$\frac{\log N}{m} \ge \frac{1}{A} J\left(\left[t - 2\delta - 2\epsilon\tau - \frac{A-1}{m}\right]_+\right). \tag{10}$$

For each fixed $0 < \delta < t/2$, sufficiently small fixed $\tau > 0$ leaves a strict gap above the unrestricted optimal rate.

The strip has two-dimensional interior and can support full-rank Fisher information, yet retains a rate penalty, so the loss is not explained by a dimension count or by the mere existence of fixed-key value freedom. The proof replaces the line identity by $F = Af - \epsilon E - D$, where E, the head-averaged attention-weighted sum of $x + y$ over nonanchor records, obeys $|E| \le \tau m$. Appendix J gives an explicit sufficient width, not the exact strip optimum.

Beyond exact support and exact value maps. The support strip is sufficient, but not necessary. If the all-record transverse moment obeys

$$\sup_{0 \le q \le Q} \left| \frac{\mathbb{E}_{p_q}(x + y) + \mathbb{E}_{r_q}(x + y)}{2m} \right| \le \tau,$$

then $\log N/m \ge A^{-1} J([t - 2\delta - 2\epsilon\tau]_+)$, without a symmetry or support assumption (Proposition 7). Exceptional records are charged by the attention mass they receive, not by their count fraction. The strict gap also persists for a nonzero neighborhood of the value matrix with fixed keys, even when the restricted line follows its rotated top right singular direction (Proposition 9); at $Q = 1$, $\epsilon = 1/4$ and $\delta = 0.05$, every such matrix within spectral distance 0.01 keeps a coefficient separation of at least 0.0285. The line's exact rate does not require the two anchors (Proposition 8; assumption ledger in Appendix L).

4 A NORMALIZED REALIZATION WITH EXPLICIT ATTENTION GAIN

Transformers normalize records before projecting them, so we separate the two resources: records are normalized, and growth is carried by an explicit attention gain m. Nonzero raw records $r_i \in \mathbb{R}^3$ pass through the declared RMSNorm map

$$n(r) = \frac{\sqrt{3}\, r}{\sqrt{\|r\|_2^2 + 3\eta}}, \qquad (u_i, v_i) = \frac{(n(r_i)_1, n(r_i)_2)}{\sqrt{3}}, \tag{11}$$

with fixed rational stabilizer $\eta \ge 0$, unit gains and no bias, following Zhang & Sennrich (2019). Fixed head projections give logits mqu_i, mqv_i and values $(1 + \epsilon)u_i - \epsilon v_i$ and $(1 + \epsilon)v_i - \epsilon u_i$, whose mean G_m targets $\tanh q$. The raw anchors $(1/m, 1/m, 1)$ and $(-1/m, -1/m, 1)$ project to $\pm(a_m, a_m)$ with $a_m = ((1 + 3\eta)m^2 + 2)^{-1/2}$; the gain and anchors vary with m, the maps do not. Let $N_{\mathrm{RMS}}(m, \delta)$ be the smallest count achieving $\sup_{|q| \le Q} |G_m(q) - \tanh q| \le \delta$, or $+\infty$ if infeasible, and define $R_{\pm}^{\mathrm{RMS}}$ by (4) over integers $m \ge 2$. Requiring $v_i = -u_i$ for nonanchors defines $N_{\mathrm{RMS,anti}}$ and $R_{\pm}^{\mathrm{RMS,anti}}$.

Corollary 5 (RMSNorm transfer with varying gain and anchors)**.** *For the family above, if $t/A < 1/\sqrt{2}$, then for every fixed $\delta > 0$,*

$$R_-^{\mathrm{RMS}}(\delta) = R_+^{\mathrm{RMS}}(\delta) = J\left(\left[\frac{t - (A+1)\delta}{A}\right]_+\right), \tag{12}$$

$$R_-^{\mathrm{RMS,anti}}(\delta) = R_+^{\mathrm{RMS,anti}}(\delta) = \frac{1}{A} J([t - 2\delta]_+).$$

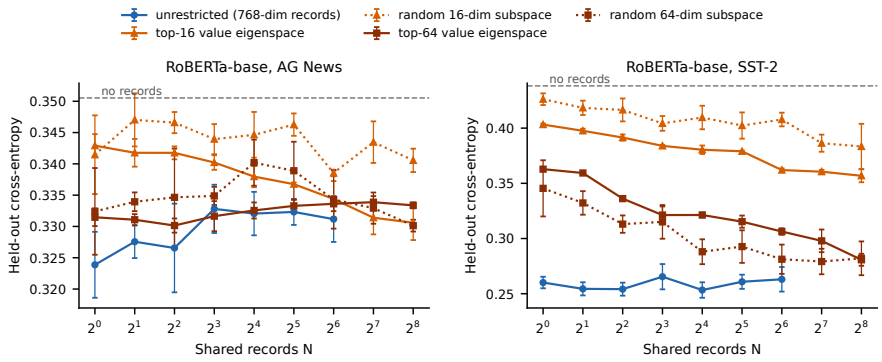


Figure 2: Held-out cross-entropy against the number of shared records N in frozen RoBERTa-base, 2000 updates (mean and seed standard deviation over three paired seeds; $N = 128, 256$ only for restricted arms). Dashed: no records. On SST-2 no restricted arm reaches the loss of one unrestricted record even at $N = 256$; on AG News the arms differ by at most 0.021 nats at each N.

The upper constructions use bounded nonzero raw records with rational $O(\log m)$-bit coordinates. The thin-strip lower bound also holds when normalized nonanchors satisfy $|u + v| \le \tau$.

Every projected record lies in the unit disk, so $(x, y) = m(u, v)$ is a box record with $F_m = mG_m$, which transfers the converse. The shifted constructions have limiting coordinate magnitudes at most $(t - d)/A < t/A$, hence a strict disk margin. A third raw coordinate realizes each projected pair, and rational rounding changes the output by $O(m^{-2})$ independently of count (Appendix C). For $\epsilon \ge (\sqrt{2} - 1)/2$, including $\epsilon = 1/4$, the condition $t/A < 1/\sqrt{2}$ holds for every finite Q; for smaller ϵ it restricts Q. This is not a fixed-gain or arbitrary-normalization theorem: at fixed gain and fixed value maps, normalization bounds the attainable output, so adding records cannot implement an arbitrarily growing-amplitude target.

5 Experiments

We test the paper's resource comparison in frozen pretrained encoders (Section 5.1), then the finite-budget geometry in the scalar interface (Section 5.2). Neither tests the asymptotic exponent or the construction itself.

5.1 Record count in frozen pretrained encoders

The theorem prices a restriction in records. We ask the same question in frozen RoBERTa-base (Liu et al., 2019). At block 6, N shared records $Z = Z_0 + CB^\top$ supply extra keys and values through the block's frozen projections, as prefix vectors do in the formalization of Petrov et al. (2024b); a frozen common head reads the first-token state, and only C is trained: 2000 AdamW updates, checkpoint selection on an internal development split of the training data, and three paired seeds (Appendix Q). Records are unrestricted ($B = I$, 768 coordinates each); confined to the top-k eigenvectors of the Gram $\sum_h (V_h O_h)(V_h O_h)^\top$ of the heads' value/output circuits (Elhage et al., 2021), the pretrained analog of the value map's highest-gain eigenspace; or confined to a random k-dimensional subspace, as in intrinsic-dimension studies (Li et al., 2018; Aghajanyan et al., 2021), with $k \in \{16, 64\}$. All arms start from the same function. As in the theorem, equal N means equal prompt length, not equal trainable coordinates.

Results. On SST-2 (Figure 2) one unrestricted record lowers held-out CE from 0.438 without records to 0.260 and raises accuracy from 80.4% to 88.8%. Within the 2000-update budget no restricted arm reaches that loss with up to 256 records; the best is 0.279 (Random-64 at $N = 128$), and Value-64 at $N = 256$ trains 21 times as many coordinates as the single unrestricted record. Paired by seed at equal N, unrestricted records have lower held-out CE in 83 of 84 pairs ($N \le 64$) and higher accuracy in all 84; at $N = 256$ the 64-dimensional arms reach similar accuracy (88.7–88.8%).

Because most restricted units selected their final checkpoint, whereas no unrestricted unit did, we refit one unrestricted record and the 64-dimensional arms at $N = 64$ and 256, the restricted cells closest to it (chosen after these results), for 10,000 updates. The restricted arms then converge before the end of training, but their held-out CE levels off at 0.271–0.290, above one unrestricted record trained as long (0.263) in all 12 seed pairs; their accuracy (89.0–89.5%) matches its 89.3%. The top value eigenspace beats a random subspace of the same dimension at $k = 16$ but not at $k = 64$, so value gain alone does not make a subspace a good record space.

On AG News the effect is small. One unrestricted record is best (0.324, against 0.351 without records), and no restricted arm matches it with up to 256 records, but at each N the arms differ by at most 0.021 nats and every accuracy lies between 88.7% and 89.5%. The SST-2 ordering holds against the 16-dimensional arms at blocks 3 and 9 and at learning rates 0.003 and 0.03 (one seed); on AG News, Value-16 is lower at learning rate 0.03 and $N = 64$ (0.330 versus 0.333). On AG News it is block-dependent: at block 9, Value-16 records have lower CE than unrestricted ones for $N \geq 4$ (0.327 versus 0.332 at $N = 16$), although unrestricted accuracy is higher in 11 of 12 pairs.

5.2 Finite optimization in the scalar interface

Protocol. Both arms use the RMSNorm model of Section 4 with $Q = 1$, gain $m = 16$, $\epsilon = 1/4$, $\eta = 10^{-5}$ and two counted raw anchors; raw coordinates lie in $[-2, 2]$. Unrestricted records have three free raw coordinates and restricted records $(a, -a, z)$ two; each pair starts at the same function and minimizes sampled training error with identical L-BFGS-B budgets, so the comparison does not isolate orientation from parameter dimension.

Fresh confirmation. After an exploratory study selected uniform queries on $[-1, 1]$, we froze ten fresh seeds with 512/128/4,096 training/validation/test queries. At the primary count $N = 32$ every pair favors unrestricted records, with median paired held-out RMSE reduction 49.4%; the prespecified bootstrap interval for the mean log ratio $\log(\mathrm{RMSE}_{\mathrm{restricted}}/\mathrm{RMSE}_{\mathrm{unrestricted}})$ is $[0.550, 0.684]$. Unrestricted records also win all ten pairs at $N = 16, 64, 128$, with median reductions 12.3%, 22.6% and 39.8%; this nonmonotone pattern does not estimate an exponent (Table 11 and Appendix N).

Boundaries. In the exploratory study (60 fits, five query laws) the restriction wins for center-concentrated queries (median reduction -49.8% at $N = 128$) and for Wine covariates at $N = 128$ (-22.3%). For edge-concentrated queries at $N = 32$ the grid maximum favors the restriction (0.184 versus 0.093) despite an 82.0% sampled-RMSE reduction. A genuine vector-query pilot (65 validation-only fits, test locked) gives median reductions of only 0.65–0.74% (Table 5). These results limit empirical generality; none contradicts the scoped theorems (Appendix D). An earlier RoBERTa-base study of three equal-rank record bases found no uniform winner, and its prespecified composite pattern failed (Appendices M and O).

6 Relationship to prior work

Approximation and structural resource bounds. Prompt-approximation work gives universality, length, memorization, or efficiency bounds under other architectural quantifiers (Wang et al., 2023b; Petrov et al., 2024a; Hu et al., 2025; Hsu & Lai, 2026). Structural separations already show severe costs: Amsel et al. (2025) prove that constant error on their target forces $dH \max_h \|V_h\|_2$ to be exponential in $d - r$ for rank-r attention (Theorem 3), and Yu et al. (2026) prove that generalized D-retrieval tasks with $s < D$ heads can require a parameter-count exponent growing with sequence length (Theorem 2); rank bottlenecks of attention heads were identified earlier (Bhojanapalli et al., 2020). Bohbot et al. (2025) instead bound convergence from i.i.d. empirical tokens to infinite-token attention. We fix two heads and optimize deterministic equal-weight record libraries, deriving the minimum record-count exponent as target amplitude grows at every fixed positive relative tolerance, with matched unrestricted and antisymmetric rates and a full-dimensional fixed-key strip result. We claim this exact characterization, not the first attention-restriction separation.

How prompt records act. Petrov et al. (2024b, Sec. 4) show that a prefix passed through frozen projections cannot change the relative attention among content tokens at its layer and only adds a bias in the span of its values. Our records act in this way; we ask how many are needed when

their geometry is restricted. Oymak et al. (2023) instead analyze a prompt that acts as the query. Low-dimensional parameterizations of adaptation are motivated by intrinsic-dimension findings (Li et al., 2018; Aghajanyan et al., 2021; Qin et al., 2024); our random-subspace arm follows Li et al. (2018), while the value-eigenspace arm uses the value/output circuits of Elhage et al. (2021).

Coupled versus independent control. APrompt motivates attention prompting through the constraints that input prompts impose on keys and values (Wang et al., 2023a), and prefix-tuning trains per-layer key/value activations through an MLP reparameterization (Li & Liang, 2021). We claim neither the first identification of that coupling nor the first evidence that relaxing it can help; both of our arms keep the same coupling and differ only in the allowed record geometry. Le et al. (2025) show statistical benefits of sharing latent prompt parameters between prefix keys and values in a mixture-of-experts regression model that permits learned atom biases (Theorems 4.1–4.3); their rates concern parameter estimation as the number of observations grows, a different question from our deterministic restriction loss. The linear-system prompt construction of Hsu & Lai (2026) relies on a rank condition allowing joint logit and value matching (Theorems 1 and 4), and Petrov et al. (2024a) transfer split-head approximation (Theorem 2) to a classical head (Theorem 3) through a larger representation (Lemma 3). Neither construction preserves our two-coordinate coupling and record accounting or supplies an arbitrary-library converse, and our scalar result gives no lower bound for those more flexible architectures.

Classical approximation ingredients. Softmax-mixture density (Nguyen et al., 2016, Theorem 7), the tilting coefficient J of importance-sampling sample sizes (Chatterjee & Diaconis, 2018) and equal-weight quadrature (Peled, 2010) are classical; none counts integer repetitions against every optimized deterministic library. We assign no novelty to them, the exponential-family identities or RMSNorm (Zhang & Sennrich, 2019) (Table 6).

7 Discussion, limitations, and conclusion

What the result says. The strongest value direction need not minimize the records needed to control attention. Equal-weight records pay for how far retrieval moves, and a family that ties returned values to keys must move retrieval farther, so retrieval selectivity and returned value jointly set the cost; dimension or value gain alone does not. The construction, an existence principle rather than an algorithm, also locates the savings: records whose keys nearly coincide while their values differ, which a restriction that ties values to keys cannot express. In a frozen encoder, records confined to a low-dimensional subspace, spectral or random, fail to match one unrestricted record with 256 records and 21 times its coordinates on one task, and cost little on another. The theory suggests judging a proposed record space by how much it lets values vary among records with nearly equal keys, not by its dimension or value gain alone; our pretrained study does not isolate this property.

Limitations. The theorems concern a specified interface: scalar queries, two prescribed heads, a fixed readout, equal-mass records without logit biases, growing gain and one target family, $m \tanh q$. They give exact asymptotic rates, and the certified finite separation covers four parameter settings. The pretrained-encoder study uses affine records around a small Z_0 at one block per run, a frozen head, two tasks, one encoder and a fixed update budget, extended fivefold only for the strongest restricted arms; its records can carry query-independent logit terms through the query bias, outside the theorems' bias-free regime; its restricted arms also cut dimension, and it does not test the scalar exponent. Exact strip rates, robustness beyond the fixed-key value-map neighborhood and multi-head theorems remain open.

Conclusion. Within this interface both optimal count rates are exact at every fixed positive tolerance, and the eigenspace restriction strictly raises the rate below $\tanh(Q)/2$; the loss survives the relaxations above, holds at certified finite amplitude, and has a measurable analog in frozen encoders.

AI USE STATEMENT

Generative AI tools were used in the research workflow to develop and critique mathematical formulations and proofs, design and implement experiment protocols and analytic-target data generation, preprocess public covariates, interpret results, search and organize related work, prepare figures, and draft and revise the manuscript. Scalar experiment targets were generated by explicit mathematical functions; the real-text diagnostics use original dataset classification labels, not language-model labels. Automated derivations and reports were checked against saved proof and numerical artifacts, including separate agent reviews and metric reconstruction; these checks do not replace author verification. The authors retain responsibility for the mathematical claims, empirical reporting, citations, and all AI-assisted content.

ETHICS STATEMENT

This work studies a mathematical attention interface, synthetic-target optimization, and bounded frozen-model classification diagnostics on public research datasets. It does not recruit participants or claim deployed-system performance. The main reporting concern is scope: projected public covariates are not downstream task validation, finite-grid maxima are not uniform certificates, real-label diagnostics do not validate the scalar theorem, and negative conditions are retained. Raw dataset text and model weights are not redistributed; original sources and attribution are recorded.

REPRODUCIBILITY STATEMENT

Appendices A–C and G, I–J, and K give full proofs, including count and encoding details. Appendix H derives the information-geometric identities and explains the conceptual figures. Appendix D specifies the frozen optimization protocols, seeds, metrics, and full grid summaries. The accompanying anonymous supplement contains all 140 scalar fitted-record arrays, 15 frozen scalar query arrays, complete scalar result rows, and all 65 vector validation rows. Its portable entry point recomputes the scalar train, validation, test and grid metrics from saved records, reconstructs aggregate statistics, and regenerates figures and tables. This is saved-fit forward replay, not independent refitting or a continuous uniform-error certificate. The vector test remains excluded and unevaluated. Frozen scalar optimization sources and optional refit instructions are also included, with source hashes and dataset attribution.

Appendices M and O record the real-label protocols, all 36 primary fits and 12 matched-control fits. The anonymous supplement provides complete compact per-fit and paired rows, an aggregation verifier, and the real-task research code. It does not contain dataset text, model weights, checkpoints, per-example logits, or full learning curves; it therefore cannot independently recompute the model evaluation metrics or replay training from saved states. The original logits, curves and audit receipts remain in the study records and were checked for this manuscript. Appendix P's explicit integer libraries for all four settings, their certificates and the verifiers are included. For Appendix Q, the supplement provides the runners, evaluation gate, compact per-unit rows, the aggregate analysis with a script that recomputes every cell mean, and the separate audit program and its reports; checkpoints and per-example logits are not included.

REFERENCES

Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992. URL https://doi.org/10.24432/C5PC7J. Dataset, CC BY 4.0.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.),

International Conference on Learning Representations, volume 2025, pp. 62431–62483, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9c537882044c8b5352c363e840872ddb-Paper-Conference.pdf.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 864–873. PMLR, 2020. URL https://proceedings.mlr.press/v119/bhojanapalli20a.html.

Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and François-Xavier Vialard. Token sample complexity of attention, 2025. URL https://arxiv.org/abs/2512.10656. arXiv:2512.10656v3, revised 8 June 2026.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, and Kai-Chiang Wu. Palu: KV-cache compression with low-rank projection. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 50222–50249, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7da6e0e00702c60607a6ae05c802ef85-Paper-Conference.pdf.

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2), April 2018. ISSN 1050-5164. doi: 10.1214/17-aap1326. URL http://dx.doi.org/10.1214/17-AAP1326.

Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine quality. UCI Machine Learning Repository, 2009. URL https://doi.org/10.24432/C56S3T. Dataset, CC BY 4.0.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Alexander Hsu and Rongjie Lai. Training-free universal approximation by prompting random transformers, 2026. URL https://arxiv.org/abs/2608.09558. arXiv preprint.

Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. In *International Conference on Learning Representations*, 2025. doi: 10.48550/ARXIV.2411.16525. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/48dfc849640344e2d58df0b5bb78c33b-Paper-Conference.pdf.

Minh Le, Chau Nguyen, Huy Nguyen, Quyen Tran, Trung Le, and Nhat Ho. Revisiting prefix-tuning: Statistical benefits of reparameterization among prompts. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QjTSaFXg25.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.08838.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Hien D. Nguyen, Luke R. Lloyd-Jones, and Geoffrey J. McLachlan. A universal approximation theorem for mixture-of-experts models. *Neural Computation*, 28(12):2585–2593, December 2016. ISSN 1530-888X. doi: 10.1162/neco_a_00892. URL http://dx.doi.org/10.1162/NECO_a_00892.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020. doi: 10.3390/e22101100. URL https://arxiv.org/abs/1808.08271.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26724–26768. PMLR, 2023. URL https://proceedings.mlr.press/v202/oymak23a.html.

Ron Peled. Simple universal bounds for chebyshev-type quadratures. *Journal of Approximation Theory*, 162(12):2317–2348, 2010. doi: 10.1016/j.jat.2010.08.002. URL https://arxiv.org/abs/0903.4625.

Aleksandar Petrov, Philip Torr, and Adel Bibi. Prompting a pretrained transformer can be a universal approximator. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40523–40550. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/petrov24a.html.

Aleksandar Petrov, Philip Torr, and Adel Bibi. When do prompting and prefix-tuning work? A theory of capabilities and limitations. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 6031–6054, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/18a367688479c1b08b001584218a4443-Paper-Conference.pdf.

Yujia Qin, Xiaozhi Wang, Yusheng Su, Yankai Lin, Ning Ding, Jing Yi, Weize Chen, Zhiyuan Liu, Juanzi Li, Lei Hou, Peng Li, Maosong Sun, and Jie Zhou. Exploring universal intrinsic task subspace for few-shot learning via prompt tuning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3631–3643, 2024. doi: 10.1109/TASLP.2024.3430545. URL https://doi.org/10.1109/TASLP.2024.3430545. Preprint: arXiv:2110.07867, "Exploring Universal Intrinsic Task Subspace via Prompt Tuning".

Utkarsh Saxena, Gobinda Saha, Sakshi Choudhary, and Kaushik Roy. Eigen Attention: Attention in low-rank space for KV cache compression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15332–15344. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.899. URL https://aclanthology.org/2024.findings-emnlp.899/.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013. URL https://aclanthology.org/D13-1170/.

Athanasios Tsanas and Angeliki Xifara. Energy efficiency. UCI Machine Learning Repository, 2012. URL https://doi.org/10.24432/C51307. Dataset, CC BY 4.0.

Qifan Wang, Yuning Mao, Jingang Wang, Hanchao Yu, Shaoliang Nie, Sinong Wang, Fuli Feng, Lifu Huang, Xiaojun Quan, Zenglin Xu, and Dongfang Liu. APrompt: Attention prompt tuning for efficient adaptation of pre-trained language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9147–9160, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.567. URL https://aclanthology.org/2023.emnlp-main.567/.

Yihan Wang, Jatin Chauhan, Wei Wang, and Cho-Jui Hsieh. Universality and limitations of prompt tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 75623–75643. Curran Associates, Inc., 2023b. doi: 10.52202/075280-3305. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/eef6aecfe050b556c6a48d9c16b15558-Paper-Conference.pdf.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The effect of attention head count on transformer approximation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 77573–77607, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7d72a514fc6948515af3ee69c1059776-Paper-Conference.pdf.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019. doi: 10.48550/ARXIV.1910.07467. URL https://arxiv.org/abs/1910.07467.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015. URL https://arxiv.org/abs/1509.01626.

GUIDE TO THE APPENDICES

Table 2 lists where each main-text result is proved or reported.

Table 2: Where each result is proved or reported.

Main-text result	Appendix
Theorem 1 (unrestricted converse)	A: balanced symmetrization and entropy-area bound
Theorem 2 (exact unrestricted rate)	B: base-and-palette construction; I: fixed-tolerance shifts
Theorem 3 (exact restricted rate, strict gap)	B, Step 9: restricted converse; G: matching construction; I
Entropy-area identity (9), Figure 1	H
Proposition 4 (full-dimensional strip)	J
Moment, anchor and value-map robustness	K; assumption ledger in L
Corollary 5 (RMSNorm)	C
Certified finite-amplitude separation	P; construction constants in D
Resource controls (Table 1)	E; related-work comparison in F
Record count in frozen encoders (Section 5.1)	Q; earlier diagnostics in M and O
Scalar optimization (Section 5.2)	D (protocol) and N (displays)

A CONVERSE AND BALANCED SYMMETRIZATION

We prove Theorem 1 and record the exact auxiliary symmetrization and anchor-restoration statements used to interpret the reduction. The notation is that of Section 2. In this appendix only, $b = [(\tanh Q - (A+1)\delta)/A]_+$ is the lower-bound parameter. Write $H_x = \mathbb{E}_p[(1+\epsilon)x - \epsilon y]$ and $H_y = \mathbb{E}_r[(1+\epsilon)y - \epsilon x]$, so $F_\mu = (H_x + H_y)/2$. The multiset union symbol $\uplus$ retains every duplicate.

A.1 STEP 1. BALANCE THE HEAD NORMALIZERS WITHOUT AVERAGING THEIR ERRORS

For each ordered pair $(i, j) \in \{1, \dots, N\}^2$, introduce

$$u_{ij} = (x_i + y_j,\ y_i + x_j).$$

Call the resulting N^2-record multiset ν. Each coordinate lies in $[-2m, 2m]$. Coordinate exchange maps u_{ij} to u_{ji}, so the multiset is coordinate-exchange invariant even if μ is not.

The first partition sum factors as

$$Z_{\nu,x}(q) = \sum_{i,j} e^{q(x_i + y_j)} = Z_x(q) Z_y(q).$$

The second factors as $Z_{\nu,y}(q) = Z_y(q) Z_x(q)$. Under the first head, indices i, j have product distribution $p(q) \otimes r(q)$. Its value is the sum of the old first-head value at i and old second-head value at j. Hence

$$H_{\nu,x}(q) = H_x(q) + H_y(q) = 2F_\mu(q).$$

Under the second head the product distribution is $r(q) \otimes p(q)$, and its value is the sum of the old second-head value at i and old first-head value at j. Therefore

$$H_{\nu,y}(q) = H_y(q) + H_x(q) = 2F_\mu(q).$$

Both normalizers and both head outputs are now equal. In particular $F_\nu(q) = 2F_\mu(q)$; no query rescaling occurred.

A.2 Step 2. Add sign reflection with one common mixture weight

Let $\sigma = \nu \uplus (-\nu)$, where $\uplus$ denotes multiset union. It has exactly $2N^2$ records, remains coordinate-exchange invariant, and is centrally symmetric. Reflection negates values and replaces q by $-q$. Because both head normalizers of ν agree, both heads use the same mixing coefficient

$$w(q) = \frac{Z_x(q)Z_y(q)}{Z_x(q)Z_y(q) + Z_x(-q)Z_y(-q)} \in [0, 1].$$

Consequently

$$F_\sigma(q) = 2w(q)F_\mu(q) - 2(1 - w(q))F_\mu(-q).$$

Both $2F_\mu(q)$ and $-2F_\mu(-q)$ lie within $2\delta m$ of $2m \tanh q$, since $\tanh$ is odd and the original hypothesis covers both query signs. Their displayed convex combination has the same error bound. This proves the balanced symmetrization assertion. It does not assert that direct sign union of μ would work.

A.3 Step 3. Establish the central-symmetry comparison

First, for every library and $q > 0$, direct substitution of the head distributions gives

$$D(p\|r) + D(r\|p) = \sum_i (p_i - r_i) \log(p_i/r_i) = 2q(f_\mu(q) - c_\mu(q)) \geq 0.$$

Indeed, $\log(p_i/r_i) = q(x_i - y_i) + \log Z_y - \log Z_x$, and the constant terms cancel because $\sum_i (p_i - r_i) = 0$. Relative entropy is nonnegative for these strictly positive finite probability distributions. It follows that $F_\mu = f_\mu + \epsilon(f_\mu - c_\mu) \geq f_\mu$.

Now assume the multiset is centrally symmetric. Define

$$\Psi(s, t) = \log \sum_i e^{sx_i + ty_i}.$$

Its Hessian is the covariance matrix of (x_i, y_i) under the corresponding positive tilt, hence is positive semidefinite. Integrating this Hessian along a line segment yields gradient monotonicity:

$$(a - b) \cdot (\nabla\Psi(a) - \nabla\Psi(b)) \geq 0.$$

Central symmetry makes $\Psi(-s, -t) = \Psi(s, t)$ and $\nabla\Psi(-s, -t) = -\nabla\Psi(s, t)$. Choose $a = (q, 0)$ and $b = (0, -q)$. The preceding inequality becomes

$$q\big(\mathbb{E}_p x + \mathbb{E}_p y + \mathbb{E}_r x + \mathbb{E}_r y\big) = 2q(f_\mu(q) + c_\mu(q)) \geq 0.$$

Thus $c_\mu(q) \geq -f_\mu(q)$ for $q > 0$. Combining the two inequalities gives

$$f_\mu(q) \leq F_\mu(q) \leq (1 + 2\epsilon)f_\mu(q) = Af_\mu(q).$$

At $q = 0$ all three outputs are zero under central symmetry, so the comparison holds there without division by q. The proof uses central symmetry, not antisymmetric support.

A.4 Step 4. Derive the entropy-area inequality for any finite multiset

For a library of L records define

$$\Phi(q) = \tfrac{1}{2}(\log Z_x(q) + \log Z_y(q)).$$

Then $\Phi(0) = \log L$, $\Phi' = f$, and

$$f'(q) = \tfrac{1}{2}\big(\operatorname{Var}_p(x) + \operatorname{Var}_r(y)\big) \geq 0.$$

Each head's Shannon entropy is $\log Z - q\mathbb{E}[\text{key}] \geq 0$, because it equals $-\sum_i p_i \log p_i$ and every probability is at most one. Averaging gives $\Phi(q) - qf(q) \geq 0$. Therefore

$$\log L \geq qf(q) - \int_0^q f(t)\,dt = \int_0^q \big(f(q) - f(t)\big)\,dt.$$

No anchor or key-bound assumption is needed for this inequality. All functions are differentiable finite sums, so the differentiations and integrals require no limiting interchange.

A.5 Step 5. Obtain the finite obstruction for a central library

Suppose a central library satisfies $|F(q) - M\tanh q| \leq \delta M$ on $[0, Q]$, where $M > 0$ is only the target amplitude in this display. This step uses no coordinate-box assumption. Step 3 yields

$$f(Q) \geq M\frac{\tanh Q - \delta}{A}, \qquad f(t) \leq M(\tanh t + \delta) \quad (0 \leq t \leq Q).$$

If $b = 0$, the desired $\log L \geq MJ(b) = 0$ follows from $L \geq 1$. If $b > 0$, then $0 < b \leq \tanh Q < 1$. Define $q_0 = \operatorname{artanh} b \in (0, Q]$. The nonnegative integrand in Step 4 may be restricted to $[0, q_0]$, by monotonicity of f. On that interval its value is at least $M(b - \tanh t)$. Hence

$$\begin{aligned}\log L &\geq M\int_0^{q_0} (b - \tanh t)\,dt \\ &= M\big(b\operatorname{artanh} b - \log\cosh(\operatorname{artanh} b)\big) \\ &= MJ(b).\end{aligned}$$

The last equality uses $\cosh(\operatorname{artanh} b) = (1 - b^2)^{-1/2}$. Thus the central-library bound holds in every case.

A.6 Step 6. Transfer it to the original unrestricted library

Apply Step 5 to σ from Steps 1–2 with $L = 2N^2$ and $M = 2m$. It gives

$$\log(2N^2) \geq 2mJ(b),$$

which is exactly Theorem 1 after dividing by two. Although the auxiliary multiset need not contain the prescribed anchors, Step 5 is valid without anchors. This is a proof device, not a claim that the auxiliary multiset is an admissible fixed-m deployment.

A.7 Step 7. Restore anchors and track the exact loss

For the sequence statement let

$$t_m = \lceil m^2 \cosh Q \rceil$$

and repeat every record of σ exactly t_m times. Its count is $L = 2t_m N^2$, its amplitude parameter is $M = 2m$, and its output is unchanged. Central symmetry implies each unnormalized head partition sum is at least L, since each sign pair contributes $2\cosh(qx) \geq 2$. Coordinate exchange implies the two head partition sums are equal; call their common value $Z(q)$.

Append one copy of each mandatory diagonal anchor $\pm(1,1)$. Their common head partition contribution is $a(q) = 2\cosh q \le 2\cosh Q$. Their isolated output under this value matrix is $\tanh q$. Both heads therefore have the same new mixture, giving

$$F_{\mathrm{new}}(q) = \frac{Z(q)}{Z(q)+a(q)} F_\sigma(q) + \frac{a(q)}{Z(q)+a(q)} \tanh q.$$

Because $M \ge 2$, the normalized approximation error is bounded by

$$\frac{|F_{\mathrm{new}}(q) - M\tanh q|}{M} \le \delta + \frac{a(q)}{Z(q)+a(q)} \frac{M-1}{M} \le \delta + \frac{\cosh Q}{t_m N^2} \le \delta + m^{-2}.$$

The count is $N_{\mathrm{new}} = 2t_m N^2 + 2$. Consequently

$$\frac{\log N_{\mathrm{new}}}{2m} - \frac{\log N}{m} = \frac{\log(2t_m + 2/N^2)}{2m} = O_Q\left(\frac{\log(m+1)}{m}\right).$$

The anchors lie in the new coordinate box. Every repetition and appended anchor has been charged. This proves the anchor-restoration assertion and completes the stated claims. □

B Counted base-and-palette construction

We first record the direct construction whose ingredients are reused in Appendix I.

Proposition 6 (Direct counted construction)**.** *Suppose*

$$0 < \delta < \min\left\{4(A-t), \frac{4t(A-1)}{A+1}\right\}, \qquad b = \frac{t+\delta/4}{A}. \tag{13}$$

For every sufficiently large integer m, an admissible anchored library with rational coordinates exists with count N_m such that

$$\lim_{m\to\infty} \frac{\log N_m}{m} = J(b). \tag{14}$$

Every repetition is counted, and each coordinate requires $O_{Q,\epsilon,\delta}(\log(m+1))$ bits. Consequently,

$$J\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right) \le R_-(\delta) \le R_+(\delta) \le J\left(\frac{t+\delta/4}{A}\right), \tag{15}$$

$$\lim_{\delta\downarrow 0} R_-(\delta) = \lim_{\delta\downarrow 0} R_+(\delta) = J(t/A). \tag{16}$$

The amplitude limit is taken before the error limit.

Throughout this appendix, $b = (t+\delta/4)/A$ under (13). Put $a = \delta/(4\epsilon)$, $q_0 = \operatorname{artanh} b$, $\eta_0 = \min\{\delta/8, q_0/4\}$ and $q_* = q_0 - \eta_0$. The overlap width η_0 is distinct from the RMSNorm stabilizer η used in Appendix C. The hypotheses imply $0 < b < 1$, $b + \delta/4 < t$, and $0 < a < b$. Constants depend on fixed Q, ϵ, δ; every $o(1)$ on a displayed compact interval is uniform in the query.

B.1 Step 1. Define the diagonal base and count every record

For all sufficiently large integers m, let k_m be the largest nonnegative integer of parity m satisfying $k_m \le bm$. Set $b_m = k_m/m$, so $b - 2/m < b_m \le b$. For $k = -k_m, -k_m+2, \dots, k_m$, define

$$C_m(k) = \binom{m}{(m+k)/2}, \quad C_b = C_m(k_m), \quad r_m(k) = \lceil C_m(k)/C_b \rceil.$$

The base B_m contains $r_m(k)$ copies of (k,k). It is invariant under both sign reversal and coordinate exchange. Each head's value at (k,k) is k, so $F_B = f_B$ and is the logarithmic derivative of its common partition sum Z_B. In particular f_B' is a variance and is nonnegative.

Write $H(p) = -p \log p - (1-p)\log(1-p)$, using $0 \log 0 = 0$. For $j \in \{0, \dots, m\}$ the elementary type estimate is

$$\frac{e^{mH(j/m)}}{m+1} \le \binom{m}{j} \le e^{mH(j/m)}.$$

Here is a proof. For $0 < j < m$, in a binomial distribution with success probability $p = j/m$, the mass at j is a mode: the ratio of mass at $j+1$ to that at j is $j/(j+1) < 1$, and the ratio at j to that at $j-1$ is $(m-j+1)/(m-j) > 1$; successive mass ratios decrease with the index. Its mass therefore lies in $[1/(m+1), 1]$. Dividing that mass by $p^j(1-p)^{m-j} = e^{-mH(p)}$ proves both inequalities. For $j = 0, m$ the inequalities hold directly. Algebra gives $J(s) = \log 2 - H((1+s)/2)$.

The retained range includes a maximum central binomial coefficient for all sufficiently large m. That coefficient is at least $2^m/(m+1)$, because all $m+1$ coefficients sum to 2^m. Using the type bounds at k_m and $\lceil z \rceil \le z + 1$ gives

$$\frac{e^{mJ(b_m)}}{m+1} \le N_B := \sum_k r_m(k) \le (m+1)e^{mJ(b_m)} + (m+1).$$

Thus $m^{-1} \log N_B \to J(b) > 0$. No random sampling, uncharged scalar weights, or uncounted duplicates enter this count.

B.2 Step 2. Prove uniform base accuracy below the transition

The full tilted binomial partition is

$$W_m(q) = \sum_{k=-m,-m+2,\dots,m} C_m(k) e^{qk} = (2\cosh q)^m.$$

Under its normalized law, the index S is the sum of m independent signs with mean $\mathbb{E}_q S = m \tanh q$. Put $d_* = b - \tanh q_* > 0$ and $h = \min\{d_*/2, b/2\} > 0$. For sufficiently large m, and uniformly on $[0, q_*]$, both endpoints of the retained band lie at distance at least mh from the mean, since $b_m \to b$ and $0 \le \tanh q \le \tanh q_*$.

For a single tilted sign X, its centered log moment-generating function is

$$\psi_q(s) = \log\cosh(q+s) - \log\cosh q - s \tanh q.$$

It has $\psi_q(0) = \psi_q'(0) = 0$ and $\psi_q''(s) = \operatorname{sech}^2(q+s) \le 1$. Taylor's integral formula on the segment from zero to s, for either sign of s, gives $\psi_q(s) \le s^2/2$. Applying Markov's inequality to $e^{h(S - \mathbb{E}_q S)}$ and to $e^{-h(S-\mathbb{E}_q S)}$ therefore gives the uniform tail estimate

$$\tau_m(q) := \mathbb{P}_q\{|S| > k_m\} \le 2e^{-mh^2/2}.$$

Let $W_{B,m}$ be the sum restricted to the retained indices. It satisfies $W_{B,m} = (1-\tau_m)W_m$. Conditional and unconditional expectations of $S/m \in [-1, 1]$ differ by at most $2\tau_m/(1-\tau_m)$, which tends uniformly to zero.

For the integer rounding, write $r_m(k) = C_m(k)/C_b + \rho_m(k)$ with $0 \le \rho_m(k) < 1$. Then

$$Z_B(q) = W_{B,m}(q)/C_b + R_m(q), \qquad 0 \le R_m(q) \le (m+1)e^{qk_m} \quad (q \ge 0).$$

The upper type bound $C_b \le \exp\{m(\log 2 - J(b_m))\}$ implies

$$\frac{R_m(q)}{W_{B,m}(q)/C_b} \le \frac{m+1}{1-\tau_m(q)} \exp\{-m[J(b_m) + \log\cosh q - q b_m]\}.$$

For $g_b(q) = J(b) + \log\cosh q - qb$, we have $g_b(q_0) = 0$ and $g_b'(q) = \tanh q - b < 0$ when $q < q_0$. Hence g_b has a strictly positive minimum on $[0, q_*]$. Continuity and $b_m \to b$ imply that the bracket in the exponent is uniformly at least a fixed positive constant for all sufficiently large m. Rounding therefore contributes a uniformly vanishing mass fraction. The corresponding normalized mean changes by at most twice that fraction because every k/m lies in $[-1, 1]$.

Combining the tail and rounding bounds proves

$$\sup_{0 \le q \le q_*} |f_B(q)/m - \tanh q| \longrightarrow 0.$$

B.3 Step 3. Control the base above and around the transition

The exact ratio of adjacent retained binomial coefficients is

$$\frac{C_m(k-2)}{C_m(k)} = \frac{m+k}{m-k+2}.$$

This ratio increases with k. Thus for $j \ge 0$ with $k_m - 2j$ retained,

$$\frac{C_m(k_m - 2j)}{C_b} \le \left(\frac{m + k_m}{m - k_m + 2}\right)^j \le \left(\frac{1+b}{1-b}\right)^j = e^{2q_0 j}.$$

For $q \ge q_0 + \eta_0$, divide each partition summand by e^{qk_m} and use $r_m \le C_m/C_b + 1$. Summing the two geometric series gives

$$Z_B(q) \le e^{qk_m}\left\{\frac{1}{1 - e^{-2\eta_0}} + \frac{1}{1 - e^{-2(q_0+\eta_0)}}\right\} = C_B e^{qk_m}.$$

The constant C_B is finite and independent of m, q. This proves a uniform bound on $[q_0 + \eta_0, Q]$ whenever this interval is nonempty; if it is empty no high-range bound is needed.

On $[q_*, q_0 + \eta_0] \cap [0, Q]$, monotonicity of f_B and its support give

$$b - \eta_0 - o(1) \le \tanh q_* - o(1) \le f_B(q)/m \le b_m \le b.$$

The first inequality uses $|\tanh u - \tanh v| \le |u - v|$, obtained by integrating its derivative $\mathrm{sech}^2 \le 1$. On this overlap the target lies between $b - \eta_0$ and $b + \eta_0$. Therefore the base's normalized error there is at most $2\eta_0 + o(1) \le \delta/4 + o(1)$.

B.4 Step 4. Build a finite positive-exponential palette

Let $I = [q_*, Q]$, $L = b + \delta/4$, $U = t$, and

$$g(q) = \max\{L, \tanh q\} \quad (q \in I), \qquad \rho = \delta/8.$$

We prove the existence of finitely many positive integers R_j, rational constants κ_j, c_j, and a function

$$G(q) = (1+\epsilon)b - \epsilon\frac{\sum_{j=1}^M R_j e^{\kappa_j q} c_j}{\sum_{j=1}^M R_j e^{\kappa_j q}},$$

such that $|c_j| \le b - a/2$ and $\sup_{q \in I} |G(q) - g(q)| \le \rho$. All these constants are fixed independently of m.

First choose $d > 0$ such that $|g(s) - g(q)| \le \rho/4$ whenever $s, q \in I$ and $|s - q| \le d$. Such a d exists because g is continuous on the compact interval. Choose a finite grid $s_1, \ldots, s_M$ in I with every

$q \in I$ within $d/4$ of some grid point. Set $h_j = g(s_j) \in [L, U]$. For $T > 0$ use Gaussian weights $w_j(q) = e^{-T(q-s_j)^2}$. Their sum is at least $e^{-Td^2/16}$. Hence the normalized total weight of points with $|s_j - q| > d$ is at most $Me^{-15Td^2/16}$, uniformly in q. Choose finite T large enough that

$$(U - L)Me^{-15Td^2/16} \le \rho/4.$$

The weighted average of h_j then differs from $g(q)$ by at most $\rho/4$ from the near points plus $\rho/4$ from the far points. It is uniformly accurate to $\rho/2$.

Canceling the common factor e^{-Tq^2} converts these weights into $\alpha_j e^{\kappa_j^0 q}$, where $\alpha_j = e^{-Ts_j^2} > 0$ and $\kappa_j^0 = 2Ts_j$. Define

$$c_j^0 = \frac{(1+\epsilon)b - h_j}{\epsilon}.$$

Since $U = Ab - \delta/4$ and $L = b + \delta/4$, we have $-b + a \le c_j^0 \le b - a$. Choose rational c_j with

$$|c_j - c_j^0| \le \min\{a/4, \rho/(4\epsilon)\}.$$

Then $|c_j| \le b - 3a/4 < b - a/2$. Replacing h_j by $h_j' = (1+\epsilon)b - \epsilon c_j$ changes any normalized mixture by at most $\rho/4$. Also $|h_j'| \le Ab$.

It remains to realize rational key offsets and counted integer weights. Choose $0 < \gamma < 1/2$ so that

$$\frac{2Ab\gamma}{1-\gamma} \le \rho/4.$$

Choose rational κ_j with $|\kappa_j - \kappa_j^0| \le \tau$, where $\tau = \log(1+\gamma/4)/Q > 0$. Let $\alpha_{\min} = \min_j \alpha_j > 0$. Choose an integer $D \ge 4/(\gamma\alpha_{\min})$ and let $R_j = \lceil D\alpha_j \rceil$. Uniformly for $q \in I \subset [0, Q]$, the weight multiplier

$$u_j(q) = \frac{R_j e^{\kappa_j q}}{D\alpha_j e^{\kappa_j^0 q}}$$

lies between $1-\gamma$ and $1+\gamma$. For the lower bound, $R_j/(D\alpha_j) \ge 1$ and $e^{-Q\tau} = 1/(1+\gamma/4) \ge 1-\gamma$. For the upper bound, $R_j/(D\alpha_j) \le 1 + \gamma/4$ and $e^{Q\tau} = 1 + \gamma/4$, whose product is at most $1+\gamma$ for the chosen range of γ.

To quantify the normalized change, if p_j are the old normalized weights, the new ones are $p_j' = p_j u_j/\bar{u}$, where $\bar{u} = \sum_j p_j u_j \in [1-\gamma, 1+\gamma]$. Thus

$$\sum_j |p_j' - p_j| \le \sum_j p_j \frac{|u_j - \bar{u}|}{\bar{u}} \le \frac{2\gamma}{1-\gamma}.$$

The mixture of h_j' changes by at most $\rho/4$. Total error is at most $\rho/2 + \rho/4 + \rho/4 = \rho$. This proves the stated palette approximation with rational κ_j, c_j and positive integer R_j. The possibly large D, R_j are constants with respect to m; their multiplicities are counted in the next step.

B.5 Step 5. Realize the palette using shared records

For each j take $m^3 R_j$ copies of each member of the four-point multiset

$$\mathcal{O}_j = \{(x_{mj}, y_{mj}), (y_{mj}, x_{mj}), (-x_{mj}, -y_{mj}), (-y_{mj}, -x_{mj})\},$$

where $x_{mj} = k_m + \kappa_j$ and $y_{mj} = mc_j$. The resulting P_m is invariant under sign reversal and coordinate exchange. Its record count is exactly

$$N_P = 4m^3 \sum_{j=1}^{M} R_j.$$

Coincident orbit members, if any, would still count with their full multiplicities. For sufficiently large m, the bounds $b < 1$ and $|c_j| \le b - a/2$ put all coordinates in $[-m, m]$.

For the first head, the leading member in each orbit has key x_{mj}. Indeed

$$x_{mj} - |y_{mj}| \ge ma/2 - 2 - \max_j |\kappa_j| \ge ma/4$$

for sufficiently large m, and $2x_{mj} \ge ma/4$ as well, because $a < b$ and $x_{mj}/m \to b$. Each of the other three keys is therefore at least $ma/4$ below x_{mj}. On I, where $q \ge q_* > 0$, their total partition weight is at most $3e^{-q_* ma/4}$ times the total leading weight. Coordinate exchange maps this calculation exactly to the second head.

Every normalized head value is bounded in absolute value by A, since both coordinates divided by m lie in $[-1, 1]$. Thus omitting the suppressed members changes the normalized scalar output by at most $6Ae^{-q_* ma/4}$. The retained first-head value at orbit j, divided by m, is

$$\frac{(1+\epsilon)x_{mj} - \epsilon y_{mj}}{m} = (1+\epsilon)b - \epsilon c_j + (1+\epsilon)(b_m - b + \kappa_j/m).$$

The second head's exchanged member has exactly this same value. In their normalized leading weights, the common factor $m^3 e^{qk_m}$ cancels, leaving $R_j e^{\kappa_j q}$. The last displayed remainder has absolute value at most $(1+\epsilon)(2 + \max_j |\kappa_j|)/m$. We have therefore proved

$$\sup_{q \in I} |F_P(q)/m - G(q)| \le 6Ae^{-q_* ma/4} + \frac{(1+\epsilon)(2 + \max_j |\kappa_j|)}{m} = o(1).$$

On I, $\tanh q \ge \tanh q_* \ge b - \eta_0$, so

$$0 \le g(q) - \tanh q \le \delta/4 + \eta_0.$$

Steps 4 and 5 give the uniform palette error

$$|F_P(q)/m - \tanh q| \le \delta/4 + \eta_0 + \delta/8 + o(1) \le \delta/2 + o(1) \quad (q \in I).$$

B.6 Step 6. Compare partitions and join the accurate pieces

Both B_m and P_m have equal first- and second-head partition sums by coordinate exchange. Consequently, for $T_m = B_m \uplus P_m$, the output is the common partition-weighted mixture

$$F_T(q) = \frac{Z_B(q)}{Z_B(q) + Z_P(q)} F_B(q) + \frac{Z_P(q)}{Z_B(q) + Z_P(q)} F_P(q).$$

This identity uses one common weight for both heads; no unequal-normalizer averaging is assumed. Each normalized output is bounded by A on the whole real line because each record has both coordinates in $[-m, m]$.

Put $C_\kappa = \max_j |\kappa_j|$. For all sufficiently large m, every palette coordinate is at most $k_m + C_\kappa$. Thus for $q \in [0, q_*]$,

$$Z_P(q) \le 4 \left(\sum_j R_j \right) e^{QC_\kappa} m^3 e^{qk_m}.$$

Step 2 gives $Z_B \geq W_{B,m}/C_b \geq (1-\tau_m)\exp\{m[J(b_m)+\log\cosh q]\}$. The positive gap proved there implies $Z_P/Z_B \to 0$ uniformly on $[0, q_*]$. Since the two outputs are bounded, $F_T/m - F_B/m \to 0$ uniformly on this range. Its target error therefore tends to zero.

On the high range $[q_0+\eta_0, Q]$, if nonempty, the palette leading terms give

$$Z_P(q) \geq m^3 e^{qk_m} \sum_j R_j e^{\kappa_j q} \geq m^3 e^{qk_m} e^{-QC_\kappa} \sum_j R_j.$$

Step 3 now gives $Z_B/Z_P = O(m^{-3})$ uniformly. The combined output is therefore within $o(1)$, after division by m, of the palette, whose error is at most $\delta/2 + o(1)$.

On the overlap $[q_*, q_0+\eta_0] \cap [0, Q]$, Steps 3 and 5 bound both normalized errors by $\delta/2 + o(1)$. Their convex mixture satisfies the same bound regardless of the mixture weight. These three ranges cover $[0, Q]$, including their endpoints. If the high range is empty, the first range and overlap alone cover it. Hence

$$\sup_{0\leq q\leq Q} |F_T(q)/m - \tanh q| \leq \delta/2 + o(1).$$

Central symmetry gives $Z_x(-q) = Z_x(q)$ and reverses the sign of the head value numerators. Thus $F_T(-q) = -F_T(q)$. Since the target is odd as well, the same bound holds on $[-Q, Q]$, including $q = 0$.

B.7 Step 7. Add the mandatory anchors and finish Proposition 6 (construction)

Append one copy of each of $(1,1)$ and $(-1,-1)$, counting both even if the base already has those coordinates. These two records have common partition contribution $Z_a(q) = 2\cosh q$ and isolated output $\tanh q$. They preserve coordinate-exchange and central symmetry. For a centrally symmetric multiset of N_T records, each head partition is at least N_T: pairing opposite coordinates uses $e^{qx} + e^{-qx} = 2\cosh(qx) \geq 2$; zero coordinates contribute one per record.

The new output is the common convex mixture of F_T and $\tanh q$. Its anchor mixture weight is bounded by $2\cosh Q/N_T \leq 2\cosh Q/N_B$. For $m \geq 1$,

$$\left|\frac{\tanh q}{m} - \tanh q\right| \leq 1.$$

It follows that

$$\sup_{|q|\leq Q} |F_m(q)/m - \tanh q| \leq \delta/2 + o(1) + \frac{2\cosh Q}{N_B} \leq \delta$$

for all sufficiently large m, since N_B grows exponentially with positive rate $J(b)$. The strict unused $\delta/2$ margin absorbs every term tending to zero.

The exact final count is $N_m = N_B + 4m^3 \sum_j R_j + 2$. As $J(b) > 0$, Step 1 implies $m^{-1}\log N_m \to J(b)$. All coordinates lie in $[-m, m]$ for the same sufficiently large range. The base coordinates are integers of size $O(m)$; the other coordinates are $k_m + \kappa_j$ and mc_j with fixed rational κ_j, c_j. A rational's fixed denominator and an $O(m)$ numerator use $O_{Q,\epsilon,\delta}(\log(m+1))$ bits. No compression of repeated records is used to define N_m. This proves Proposition 6 (construction). □

B.8 Step 8. Derive the ordered small-error rate

For every positive tolerance in the claim's range, Step 7 constructs an admissible library for all sufficiently large integers m, so $N_*(m, \delta)$ is finite eventually. A nonempty set of attainable positive integer counts has a least element; no compactness of the continuous coordinate set is needed to define this minimum.

Theorem 1 applies to every admissible library, with the original m and without a symmetry assumption:

$$\log N \geq mJ\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right) - \tfrac{1}{2}\log 2.$$

Taking the lower limit after division by m supplies the lower inequality in Proposition 6 (rate). For every sufficiently large m, Proposition 6 (construction) gives $N_*(m,\delta) \leq N_m$; taking the upper limit after division of their logarithms by m gives $R_+(\delta) \leq J(b)$. Continuity of the explicit function J on $[0,1)$ implies that both bounding coefficients tend to $J(t/A)$ as $\delta \downarrow 0$. The squeeze therefore proves both limits in Proposition 6 (rate). This step assumes no continuity of R_- or R_+ and no equality between them at a positive error tolerance. It proves neither an exact-error-0 finite representation nor an interchange of the two limits. □

B.9 Step 9. Compare with antisymmetric appendages while retaining anchors

For completeness, the entropy-area identity for any finite library of N equal-weight records follows by setting $\Phi(q) = (\log Z_x(q) + \log Z_y(q))/2$. We have $\Phi(0) = \log N$, $\Phi' = f$, and $f' = (\mathrm{Var}_p\, x + \mathrm{Var}_r\, y)/2 \geq 0$. Each head entropy equals $\log Z - q\mathbb{E}[\text{key}] \geq 0$, so

$$\log N \geq \int_0^Q [f(Q) - f(s)]\, ds.$$

The integrand is nonnegative. These are differentiations of finite sums and require no symmetry assumption.

If every record lies on $y = -x$ and there are no anchors, then $c = -f$ and $F = Af$ exactly. The anchored comparator in Theorem 3 is different: it has exactly two designated diagonal anchors and all remaining records on $y = -x$. For $q \geq 0$, write Z_x, Z_y for the full partition sums including the anchors. Directly splitting their two value numerators gives

$$F(q) = Af(q) - D(q), \qquad D(q) = \frac{A-1}{2}\, 2\sinh q \left(\frac{1}{Z_x(q)} + \frac{1}{Z_y(q)}\right).$$

Each $Z_h \geq 2\cosh q$, even when the appended multiset is not sign symmetric. Hence

$$0 \leq D(q) \leq (A-1)\tanh q \leq A-1.$$

If the comparator is accurate to δm, the endpoint and path inequalities are

$$f(Q) \geq \frac{m}{A}(t-\delta), \qquad f(s) \leq \frac{m}{A}(\tanh s + \delta) + \frac{A-1}{A} \quad (0 \leq s \leq Q).$$

Put $z_m = [t - 2\delta - (A-1)/m]_+$. If $z_m = 0$, the next lower bound is just $\log N \geq 0$. If $z_m > 0$, then $z_m \leq t$ and the entropy-area integral may be restricted to $[0, \operatorname{artanh} z_m]$. On that interval its integrand is at least $(m/A)(z_m - \tanh s)$. Integrating gives the finite comparator bound

$$\frac{\log N}{m} \geq \frac{1}{A} J\left(\left[t - 2\delta - \frac{A-1}{m}\right]_+\right).$$

This inequality counts the anchors and does not rely on any cancellation between unequal head normalizers. Taking the lower limit proves the lower-rate inequality in Theorem 3, including the case when no comparator exists for some m.

Finally, $J(0) = 0$, $J'(s) = \operatorname{artanh} s$, and $J''(s) = 1/(1-s^2) > 0$. Since $0 < 1/A < 1$ and $t > 0$, strict convexity gives

$$J(t/A) < \frac{1}{A} J(t).$$

The two explicit functions $J((t+\delta/4)/A)$ and $A^{-1}J((t-2\delta)_+)$ are continuous at zero. Their strict gap at zero persists for every sufficiently small positive δ, also chosen within the claim's feasibility range. For each such fixed δ, the constructed sequence's exponent is strictly below the lower limit for every sequence of anchored antisymmetric-appendage libraries. This proves the restricted converse and the small-tolerance comparison; Appendix I supplies the sharp fixed-tolerance upper rate in Theorem 3. It refutes the specified antisymmetric restriction's rate optimality, not a finite-m comparison at an arbitrary tolerance. □

B.10 An explicit positive-tolerance range

The strict coefficient comparison can use a specified sufficient tolerance rather than only a continuity argument. Put

$$\Delta = \frac{J(t)}{A} - J(t/A) > 0$$

and choose

$$\delta_0 = \min\left\{\frac{1}{4}, \frac{t}{4}, 2(A-t), \frac{2t(A-1)}{A+1}, \frac{2A\Delta}{9Q}\right\} > 0. \tag{17}$$

For $0 < \delta < \delta_0$, the upper construction's feasibility conditions hold and both $t - 2\delta$ and $(t+\delta/4)/A$ lie in $(0,t)$. Since $J'(s) = \operatorname{artanh} s \le Q$ on $[0,t]$, the mean-value theorem gives

$$\frac{J(t) - J(t-2\delta)}{A} \le \frac{2Q\delta}{A}, \qquad J((t+\delta/4)/A) - J(t/A) \le \frac{Q\delta}{4A}.$$

Consequently,

$$\frac{J(t-2\delta)}{A} - J((t+\delta/4)/A) \ge \Delta - \frac{9Q\delta}{4A} > \frac{\Delta}{2}.$$

If $t/A < 1/\sqrt{2}$, a sufficient normalized tolerance range is

$$\delta_{0,\mathrm{RMS}} = \min\{\delta_0,\, 4(A/\sqrt{2} - t)\} > 0. \tag{18}$$

Indeed, $(t+\delta/8)/A < 1/\sqrt{2}$ in this range, the box construction at tolerance $\delta/2$ is feasible, and its normalized upper coefficient is no larger than $J((t+\delta/4)/A)$. The same strict coefficient gap follows. This does not bound a useful finite starting amplitude: the anchor correction in the finite converse persists, and the sufficient construction constants can be enormous. The shifted construction in Appendix I identifies the fixed-positive-tolerance optimum and sharpens this sufficient separation range.

C RMSNorm realization

This appendix supplies the raw realization and direct ordered-rate construction used in Corollary 5; Appendices G and I supply the restricted and sharp fixed-tolerance constructions for the full raw-record domain declared in Section 4. Here η always denotes the fixed rational RMSNorm stabilizer. In the proof, u_i, v_i denote an ideal projected pair and u_i', v_i' its realized approximation.

C.1 Every normalized library lies in the original box family

Set $(x_i, y_i) = m(u_i, v_i)$. Then $(x_i, y_i) \in [-m, m]^2$. The logits of the box interface in Section 2 equal the present logits, and its scalar output is exactly $F_m(q) = mG_m(q)$. The unrestricted lower proof applies to arbitrary records and does not require the original fixed diagonal anchors. Hence its finite lower bound immediately gives the first inequality for every normalized library. This implication uses the entire admitted raw-record domain, not just our construction. It uses the original arbitrary-record lower bound on this disk subset; it does not claim that every point of the original square is normalization-realizable.

C.2 THE UPPER CONSTRUCTION FITS STRICTLY INSIDE THE UNIT DISK

Apply Proposition 6 with tolerance $\delta/2$, and put

$$b = (t + \delta/8)/A, \qquad a = \delta/(8\epsilon).$$

Choose δ small enough that $\delta/2$ satisfies all that theorem's hypotheses and $b < 1/\sqrt{2}$. Its nonanchor base records are (k, k) with $|k| \le bm$. Its palette records are coordinate/sign orbits of

$$(k_m + \kappa_j, mc_j), \qquad k_m/m \to b, \quad |c_j| \le b - a/2,$$

where the rational κ_j, c_j and integer multiplicities are fixed with respect to m. After dividing coordinates by m, every base point has squared norm at most $2b^2 < 1$, and every palette point has limiting squared norm at most $b^2 + (b - a/2)^2 < 1$. Finitely many offsets imply a common strict margin. For example choose s strictly between $\sqrt{2}b$ and one. There is a threshold $m_0 = m_0(Q, \epsilon, \delta)$ beyond the inherited upper-proof threshold and large enough that $(b + C_\kappa/m)^2 + (b - a/2)^2 \le s^2$, where $C_\kappa = \max_j |\kappa_j|$. Then every projected nonanchor point satisfies $\|w_i\|_2 \le s$ for $m \ge m_0$. All later thresholds may additionally depend on the fixed η; none is asserted uniform as Q varies or δ tends to zero.

Steps 6–7 of that proof actually show that the nonanchor union alone has uniform normalized error at most $\delta/4 + o(1)$ at tolerance $\delta/2$; its count has rate $J(b)$. We use that nonanchor union, and supply the present anchors below. All nonanchor u_i, v_i are rational with $O(\log m)$ bits. No record multiplicity changes.

C.3 REALIZE THE DISK WITH RATIONAL RAW VECTORS, CONTROLLING NORMALIZATION ERROR

Choose an integer $C \ge 1$ with $C^2(1 - s^2) \ge 3\eta + 1$. It is a fixed constant with respect to m. For each rational projected pair define

$$z_i = \sqrt{C^2(1 - u_i^2 - v_i^2) - 3\eta}, \qquad \widetilde{z}_i = \frac{\lceil m^3 z_i \rceil}{m^3}, \qquad \widetilde{r}_i = (Cu_i, Cv_i, \widetilde{z}_i).$$

The radicand is at least one, so each raw vector is nonzero. Before rounding, its squared raw norm is $C^2 - 3\eta$, and the normalization denominator $\sqrt{\|r_i\|_2^2 + 3\eta}$ equals C. Hence the desired projected pair is realized exactly by the unrounded vector.

Let $h_m = m^{-3}$. Upward rounding gives $0 \le \widetilde{z}_i - z_i < h_m$. For the rounded normalization denominator D_i, augment the raw vector by the fixed fourth coordinate $\sqrt{3\eta}$ only to compute its Euclidean norm. The triangle inequality and upward rounding imply

$$C \le D_i = \sqrt{\|\widetilde{r}_i\|_2^2 + 3\eta} \le C + h_m.$$

This fourth coordinate is a proof device, not an added neural state or stored record. The actual normalized projected pair is $(u_i', v_i') = C(u_i, v_i)/D_i$, remains in the unit disk, and satisfies

$$\max\{|u_i' - u_i|, |v_i' - v_i|\} \le h_m/C \le h_m.$$

The inherited normalized base coordinates are k/m, and palette coordinates are $k_m/m + \kappa_j/m$ and c_j, where κ_j, c_j have fixed rational denominators. Their denominators therefore divide a fixed integer times m; their numerators have $O(\log m)$ bits. Multiplication by the fixed integer C preserves that bound. The filler grid adds denominator m^3 and numerator at most $Cm^3 + 1$. Because η is a fixed rational, the radicand is rational of $O(\log m)$ bit length; the upward integer can be found by comparing rational squares to it, including exact-square cases. No irrational raw coordinate is stored. The constructed raw norms are bounded above by $C + h_m$, independently of m.

For completeness, if softmax logit vectors $\ell, \ell+d$ have $\|d\|_\infty \le \kappa$, the path $p_i(\theta) = \mathrm{softmax}(\ell+\theta d)_i$ satisfies

$$\sum_i |p_i'(\theta)| = \sum_i p_i(\theta)|d_i - \mathbb{E}_{p(\theta)}d| \le 2\kappa.$$

Integration gives $\|p(1) - p(0)\|_1 \le 2\kappa$, independent of record count. Here $\kappa \le mQh_m$ in either head, old values have absolute value at most A, and corresponding values change by at most Ah_m. Thus rational raw realization changes the scalar output by at most

$$Ah_m + 2AmQh_m = O(m^{-2})$$

uniformly over all $q \in [-Q, Q]$. This estimate is independent of the exponentially many counted copies. Upward rounding is unchanged by coordinate exchange or simultaneous sign reversal of the first two coordinates, so the projected multiset retains the two symmetries and its head partition sums remain equal.

C.4 Add the declared normalized anchors

The realized nonanchor union is centrally symmetric in projected coordinates and has equal head partition sums, each at least its count N_T. The two raw anchors have common partition contribution

$$Z_a(q) = 2\cosh(mqa_m) \le 2\cosh Q$$

because $ma_m < 1$. Their isolated scalar output is $a_m \tanh(mqa_m)$, of absolute value at most $1/m$. The whole library is therefore a common mixture of the accurate nonanchor output and this anchor output, with anchor weight at most $2\cosh Q/N_T$. The anchor target error is at most $1 + 1/m \le 2$. Combining all bounds yields

$$\sup_{|q|\le Q} |G_m(q) - \tanh q| \le \delta/4 + o(1) + O(m^{-2}) + 4\cosh Q/N_T \le \delta$$

for sufficiently large m. All $o(1)$ and $O(m^{-2})$ output terms here are uniform over $[-Q, Q]$, with fixed-parameter constants. Since $m^{-1}\log N_T \to J(b) > 0$, adding two anchors preserves the count rate. For every sufficiently large m, $N_{\mathrm{RMS}}(m, \delta) \le N_T(m, \delta) + 2$; take the limsup to obtain the stated upper inequality. This proves the upper bound with rational raw records; the raw anchor norms are also uniformly bounded. Continuity of J now proves the two ordered small-error limits.

C.5 The antisymmetric restriction still loses rate

For a competitor with $v_i = -u_i$ outside the anchors, use the scaled coordinates from Appendix C.1. Its anchors are $\pm(c_m, c_m)$, where $c_m = ma_m \in (0, 1)$. Splitting those two diagonal contributions exactly as in Step 9 of Appendix B gives, for $q \ge 0$,

$$F(q) = Af(q) - D_m(q), \quad D_m(q) = \frac{A-1}{2}\, 2c_m \sinh(qc_m)\left(\frac{1}{Z_x(q)} + \frac{1}{Z_y(q)}\right).$$

As both partition sums include $2\cosh(qc_m)$,

$$0 \le D_m(q) \le (A-1)c_m \tanh(qc_m) \le A - 1.$$

No symmetry of the appended competitor is assumed. The same entropy-area argument therefore gives

$$\frac{\log N}{m} \ge \frac{1}{A} J\left(\left[t - 2\delta - \frac{A-1}{m}\right]_+\right).$$

Its lower-limit coefficient is $J((t-2\delta)_+)/A$. Since $A > 1$, $t > 0$, $J(0) = 0$ and J is strictly convex, $J(t/A) < J(t)/A$. Continuity of the explicit upper and restricted lower coefficients gives the strict comparison for all sufficiently small fixed positive δ. This proves the stated restriction loss under the same RMSNorm interface.

D Experimental protocol and complete descriptive summaries

D.1 Parameterization and paired initialization

For each N, raw anchors are $(1/16, 1/16, 1)$ and $(-1/16, -1/16, 1)$. The unrestricted arm learns $N-2$ three-coordinate records. The restricted arm learns (a_i, z_i) and expands each to $(a_i, -a_i, z_i)$. Both use the same RMSNorm denominator $\sqrt{\|r_i\|^2 + 3 \times 10^{-5}}$, gain 16, coupling $1/4$, and scalar averaging. Every optimized raw coordinate is bounded by $[-2, 2]$. This parameterization makes the restricted arm a subset of the unrestricted arm; it does not equalize dimension.

For seed s and count N, initialization uses NumPy's legacy `RandomState` at $s+N$. A two-column base is first drawn uniformly from $[-0.6, 0.6]$, and its second column is then overwritten with uniform draws from $[0.7, 1.3]$. The unrestricted initialization expands this same base to $(a, -a, z)$. Thus the paired initial predictions agree. L-BFGS-B uses analytical gradients, maximum iterations 6,000, maximum evaluations 16,000, function tolerance 10^{-12}, and projected-gradient tolerance 10^{-7}. The final optimizer iterate is evaluated; there is no selection of a favorable test checkpoint. A successful stopping flag is not proof of global optimality.

D.2 Confirmation protocol and uncertainty

The confirmation uses data and optimizer seeds $s = 2026091701, \ldots, 2026091710$, with 512/128/4,096 train/validation/test queries from the uniform law on $[-1, 1]$ for each seed. Each target is evaluated analytically as $\tanh(q)$. These seeds are fresh relative to exploration. Counts are $16, 32, 64, 128$, with $N = 32$ prospectively designated primary for this confirmation. Uniform queries were selected after the exploratory outcomes, so the confirmation supports that selected regime rather than an unselected population of tasks.

For pair j, write a_j for unrestricted test RMSE and b_j for restricted test RMSE. The reported reduction is $\mathrm{median}_j(1 - a_j/b_j)$; the two arm medians are reported separately. The uncertainty target is $\bar{\ell} = 10^{-1} \sum_j \log(b_j/a_j)$. At each count, the saved analysis resampled ten paired log ratios with replacement 10,000 times using `RandomState`$(517 + N)$, then took the 0.025 and 0.975 quantiles of their means. At the primary count this interval is $[0.5503338, 0.6837892]$. No correction or simultaneous claim is made for secondary counts. Counts reuse the same seed set and are not 40 independent data tasks.

The prespecified primary success rule required median reduction at least 20%, at least eight of ten positive pairs, at least 18 of 20 converged fits, and a positive lower endpoint for the mean-log-ratio interval. The primary passes these conditions. The secondary $N = 16$ cell has a positive effect in every pair but falls below the effect-size rule. Such a rule is a prespecified decision threshold, not a theorem about when a benefit exists.

Table 3: Confirmation grid diagnostics. Each value is the median, across ten fits, of the maximum absolute error over 4,097 equally spaced queries on $[-1, 1]$. These finite grids do not certify the continuous-domain supremum.

N	Unrestricted grid max.	Restricted grid max.	Fits converged
16	0.079201	0.086427	20/20
32	0.023914	0.042722	20/20
64	0.008049	0.010931	20/20
128	0.001678	0.002558	20/20

D.3 Exploration and its selection boundary

The exploratory seeds are $2026091501, 2026091502, 2026091503$ and counts are $32, 128$. Five frozen query arrays are used: Energy covariates, Wine covariates, uniform, center-concentrated, and edge-concentrated. The synthetic laws are respectively $\mathrm{Uniform}[-1, 1]$, $\tanh Z$ for $Z \sim \mathcal{N}(0, 0.35^2)$, and a balanced sign times $\mathrm{Uniform}[0.7, 1]$, with 512/128/256 train/validation/test queries. In each cell the two arms share training queries, target values, and initial predictions. The optimizer expansion followed an earlier incomplete optimization diagnostic; the 60-fit exploration and 80-fit fresh confirmation use the same expanded budget.

Energy uses 768 building-simulation rows from UCI Energy Efficiency (Tsanas & Xifara, 2012) with eight input features. Twelve geometry groups are defined by the first five features; a seed-20260915 permutation assigns eight groups to training and two each to validation and test. Features are standardized using training statistics. With $v = (1, -1, 1, -1, 1, 1, -1, 1)/\sqrt{8}$, the query is $q = \tanh(v^\top \widetilde{x}/2)$ and the label is $\tanh(q)$. Original energy targets are unused. The scalar Wine dataset is UCI Wine (Aeberhard & Forina, 1992): 178 rows, 13 chemical-analysis features, and three unused class labels. Its frozen split has 106/36/36 train/validation/test rows. It uses the same query formula with the alternating 13-coordinate sign vector divided by $\sqrt{13}$ and saved training-standardization statistics; this formula reproduces the saved queries exactly. This dataset differs from the Wine Quality covariates used in the vector pilot (Cortez et al., 2009). All 15 scalar query arrays, their splits, and content hashes accompany the anonymous supplement, together with source attribution and a portable replay entry point. The current manuscript does not reinterpret those arrays as independently validated downstream datasets.

Table 4: All exploratory strata. U/A denote unrestricted/antisymmetric medians. Grid maxima use 4,097 points.

Distribution	N	Reduction (%)	U RMSE	A RMSE	U/A grid max.
Energy	32	54.79	0.008840	0.019611	0.03636/0.04918
Energy	128	30.25	0.000738	0.001059	0.00760/0.00549
Wine	32	30.64	0.001470	0.002129	0.03542/0.05175
Wine	128	-22.25	0.000265	0.000248	0.00785/0.00742
Uniform	32	49.29	0.006226	0.012277	0.02328/0.04172
Uniform	128	46.49	0.000238	0.000467	0.00131/0.00228
Center	32	-5.88	0.000551	0.000500	0.05757/0.05647
Center	128	-49.81	0.000049	0.000033	0.01579/0.01323
Edge	32	82.01	0.002579	0.014341	0.18445/0.09332
Edge	128	0.12	0.000300	0.000311	0.07016/0.06712

All plotted medians and paired reductions were reconstructed directly from the saved 140 fit-result rows when preparing this draft. Earlier independent metric replays checked stored fitted records against the attention computation to floating-point precision. Regenerating a plot is not a new optimization run or a renewed proof certificate.

D.4 Vector pilot boundary

A separate Wine-covariate pilot uses genuine vector queries and a synthetic four-direction odd teacher. It freezes one teacher, feature map, readout, and split. Four attention arms across three counts and five seeds, plus five multilayer-perceptron fits, give 65 fits. Only validation outcomes were inspected; the held-out test remains locked. The 30 primary fits passed the frozen capability and plateau checks, yet none of the three median effects reached the prespecified 10% success criterion. A same-dimensional random-subspace control performed worse, suggesting that restriction direction matters; it does not establish a causal account of the antisymmetric result. Independent-key–value controls changed parameter count, normalization, anchors, and feasible geometry and were still improving under the plateau diagnostic. Their lower errors therefore cannot isolate a benefit from removing coupling. These limitations motivate keeping the vector result as a boundary of the scalar empirical claim.

Table 5: Vector-query pilot: median validation RMSE and paired reduction, five pairs per count sharing one teacher and split. All pairs favor unrestricted records, but reductions remain below 1%. The held-out test remains unopened, so no held-out performance is claimed. These values are not directly comparable to scalar test RMSE.

N	Unrestricted RMSE	Restricted RMSE	Paired reduction (%)	Positive
16	0.128956	0.129874	0.6538	5/5
32	0.123969	0.124868	0.7076	5/5
64	0.122059	0.123053	0.7419	5/5

D.5 Finite feasibility of the literal construction

An algebraic audit applied the conservative Gaussian-palette choices to $Q = 1$, $\epsilon = 1/4$, normalized tolerance $1/20$, and box-construction tolerance $1/40$. The direct proof allocation uses 1,122 grid points and a Gaussian concentration parameter at least 22,307,810.83. Its nominal key offsets range from about 25.08 million to 44.62 million. Consequently, the palette coordinates lie outside $[-m, m]$ for all audited gains $m = 16, 32, 64, 128$, before evaluating any output.

The sufficient common integerizer needs at least 32,183,381 bits, and one palette multiplicity needs at least 22,012,279 bits. These figures describe integers that count explicitly repeated records; they are not the memory cost of the materialized library, which would be far larger. The controlling conservative sufficient gain is about 7.14×10^9. These are sufficient choices for this particular proof recipe, not lower bounds on every possible palette or on the gain needed by all constructions. No library was materialized. In particular, evaluating a compressed weighted formula would not constitute a realization of the equal-weight record model.

E A finite comparison with different control interfaces

This appendix retains a finite analytic control example. Its comparisons change the control interface and are distinct from the shared-family restriction studied in the main theorems. They establish feasibility and infeasibility statements under declared caps, not a general prompt-versus-weight efficiency ranking.

E.1 Elementary controls for the charged resources

The following identities explain why readout and multiplicity accounting matter. They are elementary controls for the original box interface, not additional novelty claims or statements about a fixed-gain normalized network.

A free readout gain. With only the two designated anchors $\pm(1, 1)$, both value rows equal the corresponding key, since the rows of the coupling matrix sum to one. Each head and their prescribed average therefore output $\tanh q$. Allowing an additional readout gain m realizes $m \tanh q$ exactly with two records. The gain is an extra control excluded from the main problem.

Stored weights instead of counted copies. Suppose a positive integer mass may be stored once with each distinct atom and used in both softmax heads, and only distinct atoms are counted. For odd m, take

$$(x_k, y_k) = (k, k), \quad k = -m, -m+2, \dots, m, \qquad w_k = \binom{m}{(m+k)/2}.$$

The support contains both designated anchors. Diagonal records have value k in each head, so the binomial identity gives

$$Z_x(q) = Z_y(q) = \sum_k w_k e^{qk} = (2\cosh q)^m, \qquad F(q) = \frac{d}{dq} \log Z_x(q) = m \tanh q.$$

There are $m + 1$ distinct atoms, each mass needs at most $m + 1$ bits, and straightforward storage of all masses uses $O(m^2)$ bits in addition to the coordinate descriptions. Replacing each mass by

equal-weight copies counts $\sum_k w_k = 2^m$ records. Thus this weighted construction changes the charged resource; it does not contradict a lower bound for equal-weight records. We state the identity on the odd-m subsequence to include the exact prescribed anchors without a further approximation.

E.2 A FINITE INDEPENDENT-CONTROL COMPARISON

For any library in $[-m, m]^2$, symmetric relative entropy gives $F \geq f$ on positive queries. Since $c \geq -m$, (2) also gives $F \leq (1+\epsilon)f + \epsilon m$. Let

$$m = 64, \quad \epsilon = 1/64, \quad \delta = 1/20, \quad q_1 = \log(4/3), \quad q_2 = \log 2.$$

If both positive-query errors are at most δm, then $\tanh q_1 = 7/25$ and $\tanh q_2 = 3/5$ imply

$$f(q_2) - f(q_1) \geq m \left[\frac{3/5 - 1/20 - 1/64}{1 + 1/64} - \frac{7}{25} - \frac{1}{20} \right].$$

Monotonicity of f and the entropy-area inequality at q_2 give

$$\log N \geq q_1[f(q_2) - f(q_1)] \geq \frac{816}{65} \log(4/3) \geq \frac{204}{65}.$$

Here $\log(1+x) \geq x/(1+x)$ at $x = 1/3$ proves the last step. But $N \leq 8$ would imply $\log N \leq 3 \log 2 < 9/4$, a contradiction. This proof uses two positive queries and does not invoke the two-sided symmetrization theorem.

Now allow independent keys and values in each head. Retain the two public anchors with keys and values ± 1, and append keys ± 1 with values ± 127 in each head. At each sign, the two equal-key records have mean value ± 64, so each head produces $64 \tanh q$ exactly. The eight appended integer fields fit in signed eight-bit storage, totaling 64 bits. Six additional shared two-coordinate records use 96 such bits, while the impossibility above holds even for continuous shared coordinates in the larger declared box. In this finite comparison, shared coordinates have magnitude at most 64, derived shared values have magnitude at most 66, and the common permitted value cap for the independent arm is 127. The public architecture and anchors are fixed across the count. A direct public-direction update can instead add $63 \tanh q$ to the anchor response $\tanh q$, storing the integer gain 63 in one eight-bit field.

The independent and direct controls are more permissive interfaces. The direct update assumes the nonlinear direction $\tanh q$ is public; it is not a learned solution to the original constrained representation problem. Finite-width bookkeeping for these specific integer constructions does not assign a finite storage cost to arbitrary real-valued shared coordinates. The count converse applies to that continuous relaxation and therefore to its finite-precision subset.

F ADDITIONAL COMPARISON OF THEOREM RESOURCES

Table 6 compares the resource questions of the closest theorems, and Table 7 lists what the two arms of the main comparison share.

G MATCHING THE ANTISYMMETRIC ORDERED COEFFICIENT

We give a direct upper construction for the ordered small-error case. Appendix I adds a counted shift to obtain the exact fixed-tolerance statement of Theorem 3. Fix $\beta \in (t, 1)$ and set $n = n_m = \lfloor m/A \rfloor$. For sufficiently large m, let k_n be the largest nonnegative integer of parity n with $k_n \leq \beta n$. Place

$$w_n(k) = \left\lceil \frac{\binom{n}{(n+k)/2}}{\binom{n}{(n+k_n)/2}} \right\rceil \quad \text{copies of } (k, -k), \qquad k = -k_n, -k_n + 2, \ldots, k_n. \tag{19}$$

All copies are charged. The coordinates are integers with $O(\log m)$ bits and satisfy $|k| \leq \beta n < m$. This nonanchor base is sign symmetric; $w_n(k) = w_n(-k)$ gives equal head partition sums. If $\mu_n(q)$ is the tilted mean of k, its two head outputs are both $A\mu_n(q)$, hence $F_B = A\mu_n$ exactly.

Table 6: How the compared results differ in their resource questions. Different interfaces prevent direct ordering of these theorems.

Work	Principal question/result	Distinction from the present result
Petrov et al. (2024a)	Prefix universality and approximation-length upper bounds	No matching converse for the present fixed coupled family
Hu et al. (2025)	Universality, memorization token lower bounds, inference efficiency	Different targets and architectural quantifiers
Hsu & Lai (2026)	Kernel approximation through a rank-admissible prompt construction	Joint logit/value solvability and different norm/width resources
Amsel et al. (2025); Yu et al. (2026)	Rank/head restrictions with exponential resource lower bounds	Head-count/output-weight or parameter costs, not deterministic record-count rates at fixed two heads
Bohbot et al. (2025)	Convergence of i.i.d. finite-token attention to a mean-field limit	Stochastic sample size, not optimized record count under growing amplitude
Nguyen et al. (2016)	Density of softmax mixtures	Does not charge this shared-record realization or give its count exponent
Chatterjee & Diaconis (2018)	Sample size of importance sampling	Random samples from a proposal, rather than optimized deterministic records
This work	Exact fixed-tolerance rates and thin-strip restriction loss	Two specified heads, scalar query, equal weights, growing gain

Table 7: Resources and restrictions in the main comparison. The normalized realization is defined in Section 4.

Quantity	Unrestricted shared family	Antisymmetric family
Records	N equal-weight records, including anchors	Same counting convention
Adjustable coordinates	$(x, y) \in [-m, m]^2$	$(x, -x)$; same box
Keys, values, readout	Fixed maps in (2)	Identical maps
Scaling	Growing box and target; or normalized gain m	Same scaling and anchors
Precision of upper construction	$O(\log m)$ bits per rational coordinate	Matching upper uses $O(\log m)$-bit rational coordinates

Uniform accuracy and count. The truncated-binomial count and low-query estimates in Steps 1–2 of Appendix B apply with amplitude n, cutoff β, and the entire interval $[0, Q]$: the strict inequality $\tanh Q < \beta$ puts that interval below $\operatorname{artanh}\beta$. More explicitly, the tilted-binomial mean lies a distance at least nh from either truncation boundary for a fixed $h > 0$ and large n, so the discarded mass is at most $2e^{-nh^2/2}$. The ceiling remainder is negligible because

$$\min_{0\le q\le Q}\{J(\beta) + \log\cosh q - q\beta\} > 0.$$

Consequently $\sup_{[0,Q]} |\mu_n/n - \tanh q| \to 0$ and $\log N_B/n \to J(\beta)$. Since $An/m \to 1$,

$$\sup_{[0,Q]}\left|\frac{F_B(q)}{m} - \tanh q\right| \le \frac{An}{m}\sup_{[0,Q]}\left|\frac{\mu_n(q)}{n} - \tanh q\right| + \left|\frac{An}{m} - 1\right| \longrightarrow 0.$$

Oddness extends the estimate to $[-Q, Q]$. Parity is imposed only at n, not between m and n.

The two designated anchors. Append one copy of each $\pm(1, 1)$. Neither occurs in the antisymmetric base, so exactly two diagonal exceptions are present. Both components have equal head normalizers and the base satisfies $Z_B(q) \ge N_B$. The anchor mixture weight is at most $2\cosh Q/N_B$. The anchor-only normalized output is $\tanh q/m$, whose distance from the target is at most one for $m \ge 1$. The final error is thus bounded by the base error plus $2\cosh Q/N_B$, and tends uniformly to zero. The total count $N_B + 2$ has rate $J(\beta)/A$.

For every fixed $\delta > 0$, this is admissible for all sufficiently large m. Bounding the limsup by each fixed-β construction and then taking the infimum over $t < \beta < 1$ gives

$$\frac{J([t-2\delta]_+)}{A} \le R_-^{\mathrm{anti}}(\delta) \le R_+^{\mathrm{anti}}(\delta) \le \inf_{t<\beta<1} \frac{J(\beta)}{A} = \frac{J(t)}{A}. \tag{20}$$

The first inequality is the existing converse. Continuity as $\delta \downarrow 0$ proves the matching ordered coefficient. This direct construction alone asserts no uniform-in-β threshold or fixed-positive-error optimality.

Normalized realization. Under $t/A < 1/\sqrt{2}$, choose $t < \beta < \min\{1, A/\sqrt{2}\}$. The ideal projected records $(u_k, v_k) = (k/m, -k/m)$ have norm at most $\sqrt{2}\beta/A < 1$. Apply the rational raw realization from Appendix C: the first two raw coordinates are $(Cu_k, -Cu_k)$, and the third is rounded upward. The common normalization denominator preserves $v'_k = -u'_k$ *exactly*. The norm-based rounding also preserves sign and exchange symmetries. Raw records are nonzero, have bounded norms and $O(\log m)$-bit rational coordinates. The inherited uniform output perturbation is $O(m^{-2})$, independently of count. Adding the declared varying anchors contributes at most $4\cosh Q/N_B$ to target error and does not change the exponent. Thus $R_\pm^{\mathrm{RMS,anti}}$ obey the analogue of (20) and both ordered small-error limits equal $J(t)/A$.

H Retrieval geometry and the entropy-area identity

This appendix gives the exact identities behind the geometric interpretation. The exponential-family facts are standard (Nielsen, 2020); their role here is to explain the counted converse, not to supply a separate novelty claim.

H.1 Feasible records and headwise fibers

For head 1, $(x, y) \mapsto (k_1, v_1) = (x, (1+\epsilon)x - \epsilon y)$ has determinant $-\epsilon$. The box maps to

$$|k_1| \le m, \qquad (1+\epsilon)k_1 - \epsilon m \le v_1 \le (1+\epsilon)k_1 + \epsilon m.$$

Its fixed-key slices contain a range of values. Restricting $y = -x$ leaves $v_1 = Ak_1$, so each such slice contains just one value. Changing y also changes head 2's key: there is no joint key-null direction because the two-head key map is the identity. The four-point palette orbits coordinate the two heads rather than make their keys and values globally independent.

The full value matrix has orthogonal eigenvectors $(1,1)/\sqrt{2}$ and $(1,-1)/\sqrt{2}$, with positive eigenvalues 1 and A. The latter is its maximal Euclidean norm-gain direction. This statement concerns the matrix, not a single row's gradient or the final attention-weighted scalar output.

H.2 Two paths in one exponential family

Let $z_i = (x_i, y_i)$ and define

$$\Psi(\theta) = \log \sum_i e^{\theta^\top z_i}, \qquad \pi_i(\theta) = e^{\theta^\top z_i - \Psi(\theta)}.$$

Differentiating finite sums gives $\nabla\Psi = \mathbb{E}_\pi z$ and

$$\nabla^2\Psi = \mathrm{Cov}_\pi(z) = \mathbb{E}_\pi[(\nabla \log \pi_i)(\nabla \log \pi_i)^\top].$$

This is the Fisher information matrix in natural coordinates. It is positive semidefinite, and positive definite precisely when the record support affinely spans the plane. The heads trace $p_q = \pi(q, 0)$ and $r_q = \pi(0, q)$; Figure 3 draws both paths for a three-record library. Direct cancellation of their log normalizers yields

$$D_{\mathrm{KL}}(p_q \| r_q) + D_{\mathrm{KL}}(r_q \| p_q) = 2q(f(q) - c(q)),$$

and for $q \neq 0$,

$$F(q) = f(q) + \frac{\epsilon}{2q}[D_{\mathrm{KL}}(p_q \| r_q) + D_{\mathrm{KL}}(r_q \| p_q)].$$

This is an identity, not an objective to maximize divergence: changing the library also changes f, the coefficient is signed, and accuracy is relative to the target. At zero use the original output formula.

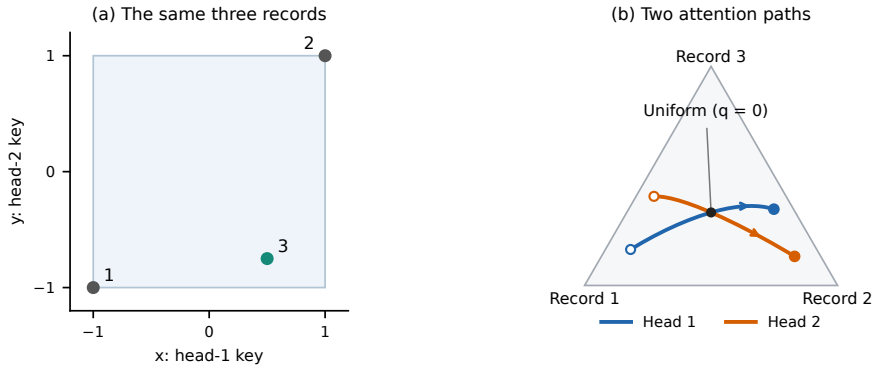


Figure 3: Two geometric views of a fixed three-record library: the designated anchors $\pm(1,1)$ and one record $(1/2,-3/4)$. Left: record coordinates. Right: the head distributions p_q, r_q in the probability simplex for $q \in [-1,1]$, meeting at uniform weights when $q=0$. Arrows indicate increasing query; filled endpoints indicate $q=1$. This is an exact illustrative calculation, not an optimized library, an empirical result or a depiction of metric distances.

H.3 An integrated Fisher-information budget

Let u_N be uniform on the counted records and $\Phi = (\log Z_x + \log Z_y)/2$. Since $\Phi(0) = \log N$ and $\Phi' = f$, for $q \geq 0$,

$$\begin{aligned} \frac{D_{\mathrm{KL}}(p_q\|u_N) + D_{\mathrm{KL}}(r_q\|u_N)}{2} &= q\Phi'(q) - \Phi(q) + \log N \\ &= \int_0^q [f(q) - f(s)]\, ds = \int_0^q s\Phi''(s)\, ds \leq \log N, \\ \Phi''(s) &= \tfrac{1}{2}[\mathrm{Var}_{p_s}(x) + \mathrm{Var}_{r_s}(y)]. \end{aligned} \tag{21}$$

The last integrand is the mean scalar Fisher information of the two query paths, weighted by s. The bound follows from nonnegative Shannon entropies. For example, if $\Phi''(s) \geq v$ throughout $[a,b] \subset [0,Q]$, then $\log N \geq v(b^2-a^2)/2$. This elementary consequence interprets sustained retrieval sensitivity as consuming the same information budget. It is not a new sharp rate theorem by itself.

Uniformly repeating every record multiplies the charged count while leaving the output, Fisher matrix and spent entropy in (21) unchanged. The result constrains the minimum feasible count, not a metric invariant.

The target and the coupled-output inequalities supply the additional constraints needed in the converse. Balanced-pair reduction handles arbitrary libraries before the area is bounded below by the appropriate $mJ(\cdot)$. Also

$$J(b) = \int_0^{\operatorname{artanh} b} (b - \tanh s)\, ds = D_{\mathrm{KL}}(\mathrm{Bern}((1+b)/2)\|\,\mathrm{Bern}(1/2)).$$

This is the shaded area in Figure 1. It is not a Fisher–Rao distance.

Rank and topology boundaries. For antisymmetric nonanchors alone, Ψ_{bulk} depends only on $\theta_1 - \theta_2$ and its Hessian has rank at most one. The two diagonal anchors can restore full rank: with any nonzero antisymmetric point, the full support affinely spans the plane. Thus a global Fisher degeneracy does not explain the anchored theorem. The feasible region and line segment are both contractible; no homological obstruction is shown. Neither these identities nor the occurrence of $\tanh$ establishes constant negative curvature. The paper's geometric claim is the scoped count separation and its entropy interpretation.

I SHARP RATES AT A FIXED POSITIVE TOLERANCE

This appendix strengthens the ordered-limit constructions by using part of the allowed error in a binary shift. Every shift is implemented by counted records. We keep the finite converses unchanged.

I.1 A BINOMIAL BASE AND AN EXACT BINARY SHIFT

For any fixed $0 < \beta < 1$, let k_n be the largest nonnegative integer of parity n not exceeding βn, and retain scalar keys $k = -k_n, -k_n + 2, \dots, k_n$ with multiplicities

$$w_n(k) = \left\lceil \binom{n}{(n+k)/2} \Big/ \binom{n}{(n+k_n)/2} \right\rceil .$$

Write $Z_{n,\beta}$ for their partition sum and $\mu_{n,\beta} = (\log Z_{n,\beta})'$. The count estimates in Appendix B, Steps 1–3, imply

$$\frac{\log N_{n,\beta}}{n} \to J(\beta), \qquad \sup_{0 \le q \le Q} \left| \frac{\mu_{n,\beta}(q)}{n} - \min\{\tanh q, \beta\} \right| \to 0. \tag{22}$$

Here the second assertion also covers intervals above the cutoff. To see the full uniform statement, fix a small $h > 0$ and put $q_\beta = \operatorname{artanh} \beta$. Below $q_\beta - h$, Step 2 applies. Above this point, monotonicity of $\mu_{n,\beta}$ and its support give $\tanh(q_\beta - h) - o(1) \le \mu_{n,\beta}/n \le k_n/n \le \beta$. The discrepancy from the clipped target is at most $h + o(1)$ there. First take $n \to \infty$ and then $h \downarrow 0$. If the cutoff is beyond Q, the low-query estimate alone suffices.

Replace every scalar key k by two keys $k + a$ and $k - a$, each with its original multiplicity. Exactly $2N_{n,\beta}$ records result, with

$$\widetilde{Z}(q) = 2\cosh(aq) Z_{n,\beta}(q), \qquad (\log \widetilde{Z})' = \mu_{n,\beta}(q) + a \tanh(aq). \tag{23}$$

The added entropy area is $aq \tanh(aq) - \log\cosh(aq) \in [0, \log 2]$. Thus a shift of order n can consume a fixed error allowance while adding only $\log 2$ to the log count. Its limiting response has a boundary layer near zero; we bound that layer rather than assume uniform convergence to a sign function.

I.2 RESTRICTED UPPER RATE

Fix rational $0 < d < \min\{\delta, t/2\}$, set $\beta = t - 2d > 0$ and $n = \lfloor m/A \rfloor$. For each retained k, take both records

$$(k + dn, -k - dn), \qquad (k - dn, -k + dn)$$

with multiplicity $w_n(k)$. They satisfy $|k \pm dn| \le (\beta + d)n = (t - d)n < m$, have rational $O(\log m)$-bit coordinates, and are sign symmetric. Their scalar output is exactly

$$F_B(q) = A\{\mu_{n,\beta}(q) + dn \tanh(dnq)\}.$$

For $0 \le q \le q_\beta$, the difference from $\tanh q$ after division by m lies in $[-o(1), d + o(1)]$, by (22), $An/m \to 1$, and $0 \le \tanh(dnq) \le 1$. For $q \ge q_\beta > 0$, the binary factor converges uniformly to one; the limiting error is $\beta + d - \tanh q = t - d - \tanh q \in [-d, d]$. Consequently the uniform error is at most $d + o(1)$ on both signs. Adding the two designated anchors contributes $o(1)$, as in Appendix G. The total count rate is $J(t - 2d)/A$.

For $0 < \delta < t/2$, take the infimum over rational $d \uparrow \delta$ after the amplitude limit. For $\delta \ge t/2$, instead take $d \uparrow t/2$, obtaining a nonnegative rate at most zero. These are bounds on the optimal limsup from a family of fixed-d constructions, not a uniform threshold in d. Together with the unchanged finite converse they give, for every fixed $\delta > 0$,

$$R_-^{\mathrm{anti}}(\delta) = R_+^{\mathrm{anti}}(\delta) = \frac{1}{A} J([t - 2\delta]_+). \tag{24}$$

I.3 Unrestricted upper rate

Fix rational $0 < d < \min\{\delta, t/(A+1)\}$ and define the distinct parameters

$$\beta = \frac{t-(A+1)d}{A} > 0, \qquad \lambda = \beta + d = \frac{t-d}{A} < 1, \qquad q_\beta = \operatorname{artanh}\beta.$$

Use the $n = m$ binomial base above and replace each diagonal record (k, k) by $(k + dm, k + dm)$ and $(k - dm, k - dm)$, with the original multiplicity for each. Its count exponent is $J(\beta)$; its extreme positive key is $\ell_m = k_m + dm$, with $\ell_m/m \to \lambda$. Its normalized output is

$$F_B(q)/m = \mu_{m,\beta}(q)/m + d\tanh(dmq).$$

The entropy cutoff is β, whereas the palette's leading key is λ. Interchanging them would give the wrong count or partition comparison.

Choose fixed $h > 0$ and $a > 0$, with $h < q_\beta/2$, $q_\beta + h < Q$, and $0 < a < \lambda$, small enough that

$$d + 2h < \delta, \qquad d + h + \epsilon a < \delta.$$

Choose a further positive palette approximation tolerance r so that $d + h + \epsilon a + r < \delta$. On $I = [q_\beta - h, Q]$, clip $\tanh q$ to the interval $[\lambda + \epsilon a, A\lambda - \epsilon a]$ and call the resulting continuous function g. This interval is nonempty because $a < \lambda$ and $A - 1 = 2\epsilon$. Since $A\lambda = t - d$, and $\tanh(q_\beta - h) \geq \beta - h$, clipping changes the target by at most $d + h + \epsilon a$.

The Gaussian-weight construction and rational/integer rounding in Appendix B, Step 4, apply with leading coordinate λ, desired values in this interval, and accuracy r. Specifically, choose sample values $g(s_j)$, set $c_j^0 = ((1+\epsilon)\lambda - g(s_j))/\epsilon$, and approximate them by rational c_j with $|c_j| \leq \lambda - a/2$. The same positive exponential approximation supplies rational offsets κ_j and positive integers R_j, fixed independently of m. Use $m^3 R_j$ copies of every member of the sign/exchange orbit of

$$(\ell_m + \kappa_j, mc_j).$$

This palette has exactly $4m^3 \sum_j R_j$ records, eventually lies in the box, and its normalized output approximates g uniformly on I to $r + o(1)$. Indeed, its leading key exceeds every other orbit key by a positive multiple of m, uniformly on I, exactly as in Step 5 of that appendix.

Both the base and palette have equal head normalizers, so their union has the common convex mixture $F_T = (1 - w)F_B + wF_P$. Its accuracy follows in three regions:

- On $[0, q_\beta - h]$, the base error is at most $d + o(1)$. The palette is negligible: its partition sum is bounded by a fixed polynomial times $e^{q\ell_m}$, while
$$\widetilde{Z}_B(q) = 2\cosh(dmq) Z_{m,\beta}(q) \geq e^{dmq} Z_{m,\beta}(q).$$
Thus the shift cancels in the ratio, leaving the strictly positive exponent $J(\beta) + \log\cosh q - q\beta$ from the original low-query argument.
- On $[q_\beta - h, q_\beta + h]$, the base error is at most $d + 2h + o(1)$ by monotonicity and (22). The palette error is at most $d + h + \epsilon a + r + o(1)$. Every common convex mixture remains accurate to less than δ for large m.
- On $[q_\beta + h, Q]$, Step 3 gives $Z_{m,\beta}(q) \leq Ce^{qk_m}$, hence $\widetilde{Z}_B \leq 2Ce^{q\ell_m}$. The palette has partition at least $cm^3 e^{q\ell_m}$ for a fixed $c > 0$, so the base mixture weight tends uniformly to zero. The palette is already accurate.

Sign symmetry supplies the negative interval. All values divided by m are bounded, so the vanishing mixture weights give vanishing output errors. The base alone has exponentially many records since $\beta > 0$; appending the two anchors changes the normalized response by $o(1)$ and does not change the exponent. All coordinates are rational with $O(\log m)$ bits: k_m is integral, d and the palette offsets/values are fixed rational numbers. Neither β nor λ is stored as a record coordinate.

We have obtained $R_+(\delta) \leq J((t-(A+1)d)/A)$ for every such fixed rational d. Taking $d \uparrow \delta$ if $\delta < t/(A+1)$, or $d \uparrow t/(A+1)$ otherwise, and applying Theorem 1, proves

$$R_-(\delta) = R_+(\delta) = J\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right), \qquad \delta > 0. \tag{25}$$

A zero exponent is a subexponential optimal-rate statement; no bounded-count construction follows from this limiting argument.

I.4 Normalized transfer and strict comparison

Under $t/A < 1/\sqrt{2}$, all nonanchor coordinates in the unrestricted construction divided by m have limiting magnitudes at most $\lambda = (t-d)/A < t/A$, with finitely many offsets. The restricted construction has limiting coordinate magnitudes at most $(t-d)/A$ as well. Both therefore have a strict disk margin. Apply to each nonanchor union the rational realization and count-independent $O(m^{-2})$ output perturbation in Appendix C; it preserves antisymmetry when present. The varying normalized anchors add $o(1)$. Each construction had strict tolerance slack, so the same upper rates hold. The inherited lower bounds yield (24) and (25) for the corresponding RMS rates.

Finally, if $0 < \delta < t/2$, let $s = t - 2\delta > 0$. The unrestricted argument of J is $[(s-(A-1)\delta)/A]_+ < s/A$. Monotonicity and strict convexity give

$$J\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right) < J(s)/A.$$

For $\delta \geq t/2$, both rates are zero. Thus the strict loss holds throughout $0 < \delta < t/2$, and the original ordered coefficients follow by $\delta \downarrow 0$. No equality at finite m, useful uniform starting amplitude, or practical implementation of the palette is asserted.

J Robustness to a full-dimensional strip

For a fixed $\tau \geq 0$, require every nonanchor to satisfy $|x_i + y_i| \leq \tau m$, retaining exactly the two designated anchors $\pm(1,1)$ and charging all copies. Additional records at an anchor coordinate are allowed only if they themselves satisfy the strip condition. No symmetry of the nonanchor multiset is assumed.

Define

$$E(q) = \tfrac{1}{2}\{\mathbb{E}_{p_q}[(x+y)\mathbf{1}_{\mathrm{bulk}}] + \mathbb{E}_{r_q}[(x+y)\mathbf{1}_{\mathrm{bulk}}]\}.$$

Each expectation uses the full normalized head distribution; it is not conditioned on the bulk. Hence $|E(q)| \leq \tau m$. The identity $F = Af - \epsilon(f+c)$, with the anchors split off, gives

$$F(q) = Af(q) - \epsilon E(q) - D(q), \qquad D(q) = 2\epsilon \sinh q\{Z_x(q)^{-1} + Z_y(q)^{-1}\}.$$

For $q \geq 0$, each partition sum is at least $2\cosh q$, so $0 \leq D(q) \leq 2\epsilon = A - 1$. Accuracy therefore implies

$$f(Q) \geq \frac{m}{A}(t - \delta - \epsilon\tau), \qquad f(s) \leq \frac{m}{A}(\tanh s + \delta + \epsilon\tau) + \frac{A-1}{A} \quad (0 \leq s \leq Q).$$

Apply the nonnegative entropy-area integral from Appendix A over $[0, \operatorname{artanh} z]$, where $z = [t - 2\delta - 2\epsilon\tau - (A-1)/m]_+ \leq t$. This yields

$$\frac{\log N}{m} \geq \frac{1}{A} J\left(\left[t - 2\delta - 2\epsilon\tau - \frac{A-1}{m}\right]_+\right). \tag{26}$$

For $0 < \delta < t/2$, write

$$\Delta_\delta = \frac{1}{A}J(t-2\delta) - J\left(\left[\frac{t-(A+1)\delta}{A}\right]_+\right) > 0.$$

Since $J'(s) \leq Q$ on $[0,t]$, including truncation at zero gives

$$\frac{J(t-2\delta) - J([t-2\delta-2\epsilon\tau]_+)}{A} \leq \frac{2\epsilon Q\tau}{A}.$$

Thus every fixed $0 < \tau < A\Delta_\delta/(4\epsilon Q)$ leaves at least $\Delta_\delta/2$ between the strip lower rate and the unrestricted optimal rate. This is a sufficient width, not the maximal allowable width or an exact strip rate.

For the normalized family impose $|u_i + v_i| \leq \tau$ on nonanchors. After scaling $(x,y) = m(u,v)$, the designated anchors are $\pm(c_m, c_m)$ with $c_m = ma_m < 1$. Their correction becomes

$2\epsilon c_m \sinh(c_m q)(Z_x^{-1} + Z_y^{-1})$, which is in $[0, 2\epsilon c_m] \subset [0, A-1]$. The same lower bound follows. Under the existing disk margin, compare it with the exact unrestricted RMS rate in Appendix I.

For every $\tau > 0$ the allowed strip has nonempty two-dimensional interior and permits full-rank support. Under the head-1 map it obeys $|v_1 - Ak_1| \le \epsilon\tau m$, intersected with the original parallelogram. It therefore has value fibers of positive width at interior keys, yet can retain a strict rate penalty. This excludes an explanation based only on a zero-versus-nonzero fiber width or a support-dimension test. It is an existence/converse statement in the same prescribed interface, not a causal explanation of the inherited line-versus-plane optimization results.

K MOMENT, ANCHOR, AND FIXED-MAP ROBUSTNESS

This appendix records three robustness statements for the specified equal-mass two-head interface. They retain the fixed attention keys and the baseline value matrix unless a perturbation is explicitly stated. They do not concern a general pretrained Transformer. Write

$$A = 1 + 2\epsilon, \quad t = \tanh Q, \quad f = \frac{\mathbb{E}_p x + \mathbb{E}_r y}{2}, \quad c = \frac{\mathbb{E}_p y + \mathbb{E}_r x}{2}.$$

Thus the baseline output is

$$F_0 = Af - \epsilon(f + c). \tag{27}$$

Proposition 7 (All-record transverse moment)**.** *Let a finite library in $[-m, m]^2$ satisfy*

$$\sup_{|q| \le Q} |F_0(q) - m \tanh q| \le \delta m, \qquad \sup_{0 \le q \le Q} \left| \frac{\mathbb{E}_{p_q}(x+y) + \mathbb{E}_{r_q}(x+y)}{2m} \right| \le \tau.$$

Then

$$\frac{\log N}{m} \ge \frac{1}{A} J([t - 2\delta - 2\epsilon\tau]_+)\,. \tag{28}$$

No support, symmetry, anchor, or partition-normalizer condition is used.

Proof. Set $h = (f + c)/m$. The hypotheses and (27) imply

$$f(Q) \ge \frac{m}{A}(t - \delta - \epsilon\tau), \qquad f(s) \le \frac{m}{A}(\tanh s + \delta + \epsilon\tau), \quad 0 \le s \le Q.$$

If $z = [t - 2\delta - 2\epsilon\tau]_+$ is positive, then $q_0 = \operatorname{artanh} z \le Q$ and

$$f(Q) - f(s) \ge \frac{m}{A}(z - \tanh s), \qquad 0 \le s \le q_0.$$

The entropy-area inequality

$$\log N \ge \int_0^Q \{f(Q) - f(s)\}\, ds$$

has a nonnegative integrand. Restricting it to $[0, q_0]$ proves (28); $z = 0$ is immediate. □

The condition is strictly weaker than a support strip. For $m \ge 2$, the multiset $K \cdot (0,0) \uplus \{(m,m), (1,1), (-1,-1)\}$ violates every $|x + y| \le \rho m$ with $\rho < 2$, but its all-record moment is at most $(2me^{Qm} + 4\sinh Q)/(mK)$ uniformly on $[-Q, Q]$.

For the original unrestricted rate

$$U(\delta) = J\left(\left[\frac{t - (A+1)\delta}{A}\right]_+\right),$$

let $\Delta_\delta = A^{-1}J(t - 2\delta) - U(\delta)$. If $0 < \delta < t/2$, then $\Delta_\delta > 0$: set $s = t - 2\delta > 0$ and $a_U = [(s - (A-1)\delta)/A]_+$. Even after truncation, $0 \le a_U < s/A$, while $J(s)/s$ is strictly increasing on $(0, 1)$. Since $J' \le Q$ on $[0, t]$, the lower bound (28) remains strictly above $U(\delta)$ whenever

$$0 \le \tau < \frac{A\Delta_\delta}{2\epsilon Q}.$$

Exceptional mass. For $0 \le \rho \le 2$, let good records obey $|x+y| \le \rho m$, let $\bar{b}$ and $\bar{a}$ be the mean two-head attention masses on arbitrary nonanchor outliers and on the two designated anchors. These sets and the good set partition all counted copies. Each extra copy at an anchor coordinate belongs to the good set if it satisfies the strip, and to the outlier set otherwise. The preceding proposition applies with

$$\tau_{\text{eff}} = \sup_{0 \le q \le Q} \left\{ \rho(1 - \bar{b}(q) - \bar{a}(q)) + 2\bar{b}(q) + \frac{2}{m}\bar{a}(q) \right\}.$$

If the exceptional mass includes anchors and is at most β in each head, then $\tau_{\text{eff}} \le \rho + (2-\rho)\beta$. If β excludes anchors of per-head mass at most α, then

$$\tau_{\text{eff}} \le \rho + (2-\rho)\beta + \max\{0, 2/m - \rho\}\alpha.$$

Proposition 8 (Anchor-free and small arbitrary anchors)**.** *If every record is antisymmetric, $y = -x$, then for every fixed $\delta > 0$ its optimal count exponent equals*

$$\frac{1}{A} J([t - 2\delta]_+).$$

The same is true if a prescribed counted anchor multiset $\mathcal{C}_m$ is exempted, provided

$$\log(1 + |\mathcal{C}_m|) = o(m), \qquad \max_{(x,y) \in \mathcal{C}_m} \max\{|x|, |y|\} = o(m).$$

Proof. For the anchor-free line, $f + c = 0$, so $F_0 = Af$ and the area proof gives the finite lower bound $A^{-1}J([t-2\delta]_+)$. The shifted truncated binomial construction already consists entirely of antisymmetric records; its fixed-d error is $d + o(1)$ and its exponent is $J(t-2d)/A$. Taking the prescribed infimum over fixed d proves the rate, including a zero exponent when $\delta \ge t/2$.

For arbitrary anchors of coordinate bound b_m, their total contribution to $|f + c|$ is at most $2b_m$, yielding the lower bound

$$\frac{1}{A} J([t - 2\delta - 4\epsilon b_m/m]_+).$$

For each fixed $0 < d < \min\{\delta, t/2\}$, the symmetric binomial bulk has a positive count exponent and each partition sum at least its count. The anchor partition mass is at most $|\mathcal{C}_m| e^{Qb_m} = e^{o(m)}$, hence is a vanishing mixture fraction. Its head values are $O(b_m)$ while bulk head values are $O(m)$, so the normalized output change is $o(1)$. Strict fixed-d slack and the same infimum complete the proof. If prescribed anchors have arbitrary real coordinates, this is a count-rate statement; it does not assert an all-record finite-bit representation. □

Proposition 9 (Fixed-key value-map robustness)**.** *Let $W_0 = \left(\begin{smallmatrix} 1+\epsilon & -\epsilon \\ -\epsilon & 1+\epsilon \end{smallmatrix}\right)$, let $W = W_0 + \Delta W$ with $\|\Delta W\|_2 \le \eta$, and keep the attention keys x, y fixed. Put $g = A^2 - 1$ and $e = 2A\eta + \eta^2$. If $e < g/2$, the top right singular direction u of W is unique up to sign. If every nonanchor belongs to* $\operatorname{span}(u)$*, then*

$$|x + y| \le \tau m, \qquad \tau = \frac{2e}{g - e}.$$

Moreover $|F_W - F_0| \le \sqrt{2}\eta m$. Therefore a perturbed library accurate at tolerance δ obeys

$$\frac{\log N}{m} \ge \frac{1}{A} J\left(\left[t - 2(\delta + \sqrt{2}\eta) - 2\epsilon\tau - \frac{A-1}{m} \right]_+ \right).$$

For $\delta > \sqrt{2}\eta$, the unrestricted perturbed limsup rate is at most $J([t - (A+1)(\delta - \sqrt{2}\eta)]_+/A)$.

Proof. Let $M = W^\top W$, $M_0 = W_0^\top W_0$, and $E = M - M_0$. Then $\|E\|_2 \le e$. The top and bottom eigenvectors of M_0 are respectively $v_- = (1,-1)/\sqrt{2}$ and $v_+ = (1,1)/\sqrt{2}$, with eigenvalues A^2 and 1. Weyl's inequality gives a unique top eigenvalue λ of M. Projecting $Mu = \lambda u$ onto v_+ gives

$$(\lambda - 1)\langle v_+, u\rangle = \langle E v_+, u\rangle, \qquad |\langle v_+, u\rangle| \le \frac{e}{g - e}.$$

Writing a box-constrained line record as au gives $|a| \le \sqrt{2}m$ and the strip bound follows. The fixed-key output bound follows from $\|z\|_2 \le \sqrt{2}m$. The stated lower bound is the original strip converse after this tolerance sandwich; the upper bound applies the old unrestricted construction with strict tolerance $\delta - \sqrt{2}\eta$. □

For $0 < \delta < t/2$, the original strict gap survives under a concrete nonzero perturbation neighborhood. With $\Delta_\delta = A^{-1}J(t-2\delta) - U(\delta)$, it is sufficient that

$$\sqrt{2}\eta < \min\{\delta, t/2 - \delta\}, \quad e < g/2, \quad (A+3)\sqrt{2}\eta + \frac{4\epsilon e}{g-e} < \frac{A\Delta_\delta}{Q}.$$

The result is a perturbative separation, not an exact rate theorem for a general W. If keys move, a separate value-envelope bound is required; it is not covered by this proposition.

Numerical scale of the bound. For $Q = 1$, $\epsilon = 1/4$, $\delta = 0.05$ and $\eta = 0.01$, the proposition covers every fixed W with $\|W - W_0\|_2 \le 0.01$ under its fixed-key and top-right-singular-line conditions. Here $e = 0.0301 < g/2 = 0.625$ and $\tau \le 0.0493484$. Evaluation of the proved bounds gives a restricted asymptotic lower coefficient at least 0.1325 and an unrestricted limsup construction coefficient at most 0.1040, hence a separation at least 0.0285. The sufficient strict-gap inequality evaluates to $0.0883138 < 0.0990641$. The finite restricted inequality still contains $-(A-1)/m = -0.5/m$. These numbers illustrate the size of a sufficient neighborhood; they are neither exact perturbed rates nor an empirically measured robustness radius.

L WHICH ASSUMPTIONS DEFINE THE QUESTION?

The original interface is a tractable control problem rather than an approximation theorem for a pretrained Transformer. The following distinctions prevent a robustness result from being mistaken for evidence that all assumptions hold in practice.

Table 8: Assumption ledger. A formal relaxation changes a theorem hypothesis. A motivating example explains relevance but does not verify the hypothesis empirically.

Assumption	Status	Evidence and remaining boundary
Exact line support	Relaxed	Proposition 7 needs only uniformly small attention-weighted transverse moment. Outliers are charged by retrieved mass; a small count fraction alone is insufficient.
Exact coupled value matrix	Locally relaxed	Proposition 9 preserves a strict count separation for nearby value matrices, even when their top right singular direction rotates. Keys remain fixed; perturbed rates are not characterized exactly.
Two designated anchors	Relaxed for line rate	Proposition 8 covers no anchors and prescribed subexponential collections with sublinear coordinates.
Fixed backbone and shared key/value source	Motivated, not universal	Frozen-model soft prompting and input key/value coupling motivate this control choice (Lester et al., 2021; Wang et al., 2023a). They do not imply our specific maps.
Fixed readout and counted equal masses	Essential resource definition	Free readout gain gives a two-record solution; free integer atom weights change what is charged (Appendix E.1).
Scalar queries, two heads, fixed keys	Still specialized	They permit sharp converse/construction matching. No general trained multi-head model theorem is claimed.
Growing gain and target amplitude	Asymptotic regime	RMS normalization gives the stated equivalent varying-gain realization. This is neither a claim about natural training trajectories nor a runtime advantage.

The theorem's geometric quantity is not merely the number of available coordinates. The transverse moment measures the part of each retrieved record that changes the relation between retrieval and output. Since attention reweights a library with the query, a record far from the preferred line can be harmless when it receives little mass, while a small number of highly retrieved outliers can matter. The moment proposition quantifies this distinction; it does not establish that a finite calibration sample certifies its uniform-in-query premise.

The constructive base/palette division of roles explains how the larger feasible set can reduce an optimal count exponent. It remains an existence construction with large constants. The practical lesson is to evaluate retrieval and returned values jointly when choosing a constrained record space,

Table 9: Real-label diagnostic, means of three paired training seeds. Each cell is accuracy (%)/CE (nats); higher accuracy and lower CE are better. AG News uses 7,600 official test examples; SST uses 872 official development roots, separate from the train-derived selection split. No-prefix accuracy/CE: AG News 88.71/0.3505; SST 80.39/0.4383. Every task/count cell is shown.

Task	N	Value	Random	Joint
AG News	8	88.79/0.3480	88.78/0.3492	88.80/0.3433
AG News	32	88.89/0.3446	88.79/0.3474	88.96/0.3421
SST	8	82.38/0.4108	80.81/0.4302	81.84/0.4132
SST	32	82.38/0.4109	81.61/0.4252	82.61/0.4053

rather than to deploy this construction or expect a general accuracy improvement. General relevance is tested separately from the exact theorem, with its own comparison and optimization limitations.

M REAL-LABEL PRETRAINED-MODEL DIAGNOSTIC

This supplementary study was specified after the scalar theory and its finite studies. AG News was followed by the predeclared SST task regardless of the first outcome. It asks a narrower empirical question than the theorem: at matched rank and record count, do different shared K/V record orientations give different fixed-budget adaptation results in a frozen pretrained model? It is not a literal test of the two-head count exponent or a benchmark comparison against full prompt tuning.

Main-text summary of this study. The theory suggests that choosing record directions by value gain alone can neglect retrieval. We probe this with shared K/V memories in frozen RoBERTa-base (Liu et al., 2019) on AG News classification (Zhang et al., 2015) and SST binary sentence-root sentiment (Socher et al., 2013). At zero-based block 6, N record embeddings $Z = Z_0 + CB^\top$ supply extra keys and values through the frozen projections, and only $C \in \mathbb{R}^{N\times 16}$ is trained: 128 or 512 coefficients at $N = 8, 32$. The three rank-16 bases B are the top eigenvectors of a static value/output Gram (Value, closest in spirit to the eigenspace restriction), a Haar-random basis (Random), and the top eigenvectors of label-free, locally calibrated post-output-projection record Jacobians that include both the key and value routes (Joint). The backbone and a common training-derived classifier remain frozen. All 36 predetermined fits use 500 AdamW updates and internal-development CE selection. Appendix M provides the complete protocol, seed rows and diagnostics.

Results. Joint has lower CE than Value in all six AG News pairs, but its mean accuracy changes are only $+0.013$ and $+0.079$ percentage points at $N = 8, 32$ (Table 9, Figure 4). On SST, Value wins all three accuracy pairs at $N = 8$ (Joint minus Value mean -0.535 points), whereas Joint wins all three at $N = 32$ ($+0.229$ points); CE follows the same mean reversal. In all 12 task/count/seed cells, both structured arms have lower CE than Random. A calibration-matched value-route control (VJ), which removes only the key-route term from the Joint Jacobian, has higher CE than Joint in all 12 fresh-held-out pairs but higher mean accuracy in three of four task/count cells, so the prespecified composite support pattern fails (Appendix O). In this outcome-informed follow-up, Joint, which adds the key-route term, has lower CE than VJ in all 12 pairs but lower mean accuracy in three of four task/count cells, including every SST pair.

These are task- and count-dependent observations, not broad dominance. The 500-step budget selects the endpoint in 27/36 primary fits; learned-parameter counts match, but calibration information and basis-construction costs do not. The affine, multi-head pretrained interface differs from the origin-centered scalar theorem. Neither this comparison nor its descriptive Jacobian diagnostics identifies a causal retrieval mechanism or verifies an asymptotic rate.

Model and interface. We use `FacebookAI/roberta-base`, immutable revision `e2da8e2f811d1448a5b465c236feacd80ffbac7b`, in eager attention mode without its pooler. At zero-based encoder block 6, N shared embeddings provide additional keys and values through the original frozen projections; they do not add query/output positions. Natural context keys and values remain present. The update is $Z = Z_0 + CB^\top$, where the columns of $B \in \mathbb{R}^{768\times 16}$ are orthonormal and C is the only trainable state. $C = 0$ gives identical functions across all three arms

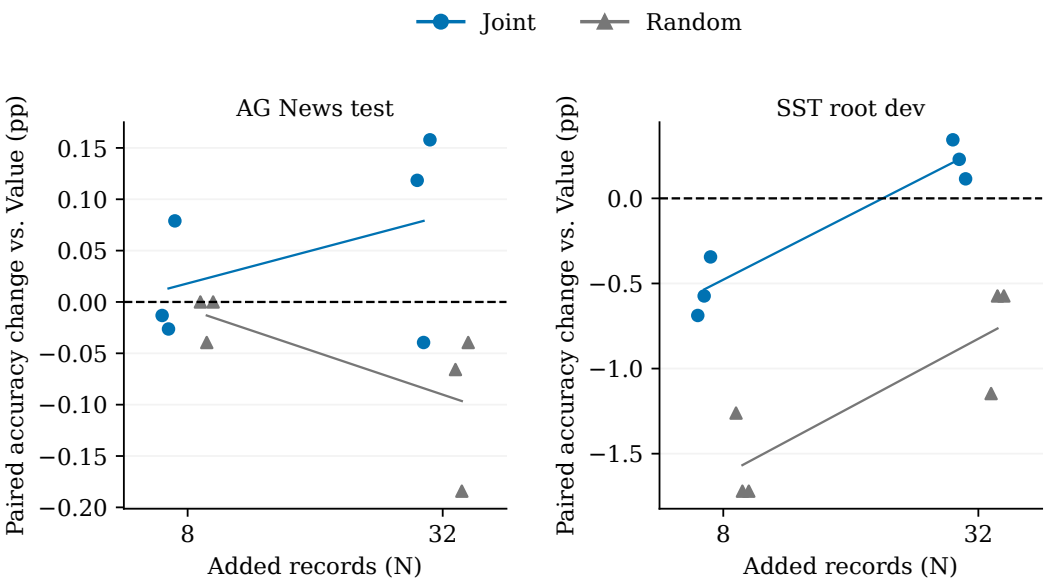


Figure 4: All paired real-task accuracy changes relative to Value. Each point is one training seed; lines join three-seed means at the two tested counts and imply no intermediate curve. The black dashed line is zero difference. Joint's direction reverses with count on SST; there is no uniform winner. Panel scales differ.

for each paired seed and count. The pair-specific initial records, common norm-radius constraint, data order and model are shared. The radius is the median calibration hidden-state norm; every update row is projected into that ball.

Basis definitions. Use a row-vector convention and head maps K_h, V_h, O_h. The static Value Gram is $\sum_h (V_h O_h)(V_h O_h)^\top$; Random is Haar. Joint uses the full derivative of a selected natural query's attention output with respect to one prefix record, after the output projection and before LayerNorm. With prefix probability a_{thi}, prefix value v_{hi}, query q_{th} and full natural-plus-prefix attention output c_{th} (all row vectors), its input-by-output derivative matrix is

$$D_{ti} = \sum_h \left[a_{thi} V_h O_h + (K_h q_{th}^\top / \sqrt{d_h}) \{ a_{thi} (v_{hi} - c_{th}) O_h \} \right].$$

The basis contains the top 16 eigenvectors of the mean $D_{ti} D_{ti}^\top$ over records and selected calibration queries. The cross-head sum is taken before forming the Gram. We use 32 label-free calibration texts and up to four eligible non-special query positions per text. Only queries are subsampled; their attention denominators retain the full nonpadding natural memory. The basis is computed separately for each (N, seed) at the shared Z_0. Thus the arms match learned parameter count, but not information used to construct the basis or preprocessing cost. Actual RoBERTa attention Jacobians were checked against autograd, including padded contexts; maximum absolute discrepancy was 2.1×10^{-17} in the checked double-precision case.

Data and information boundary. AG News uses the public classification corpus attributed to Zhang et al. (2015); its pinned mirror revision is `eb185aade064a813bc0b7f42de02595523103ca4`. SST uses the original Stanford tree archive and only complete sentence roots: labels 0/1 are negative, 3/4 positive, and neutral roots are omitted. Subtree phrases are never split across preparation/evaluation. For each task, normalized-exact-text groups from the original training source are deterministically partitioned into 2,000 balanced fit examples, 500 calibration examples, 500 internal-development examples, and 64 disjoint smoke examples (split seed 20260922). Ambiguous-label duplicate groups are excluded. AG News final evaluation uses all 7,600 official test examples; SST uses all 872 official binary

development roots, not the hidden GLUE test or the original SST test file. No final-evaluation row is loaded by fitting or basis construction. Final evaluation occurs only after the source/configuration freeze, all 18 task fits and explicit stage authorization. No exact-text overlap with preparation remained in either evaluation; overlap-excluded metrics therefore match the full metrics.

Optimization. A common linear classifier is fitted on no-prefix training features for 500 updates (seed 721), selected by internal-development CE, then frozen for every arm. Each record fit uses AdamW, learning rate 0.01, zero weight decay, batch 16, 500 updates, maximum token length 128 and disabled dropout. Development CE is evaluated every 25 steps; its minimum selects the saved checkpoint. Record counts are 8 and 32 and paired seeds 4101, 4102 and 4103, giving 18 fits per task and 36 total. The design controls learned coefficient counts (128/512); it does not establish optimized capacity or equal total compute. In 27/36 fits the best development checkpoint is at step 500. This is a material fixed-budget limitation, not grounds for extending selected arms after reading evaluation outcomes.

Verification and interpretation. A separate audit program recomputed accuracy and CE from all per-example saved logits, checked canonical evaluation IDs/labels and denominators, source/head/basis/checkpoint hashes, all 500-update receipts, coefficient counts, finite values, norm constraints and development checkpoint selection. Both 18-unit audits passed. This computation is separate from the training metric path; it is not a statistical guarantee or a mathematical proof. All raw checkpoints and logits are retained for audit; the displayed values come from the complete compact export. Figures 5 and 6 show every fit in accuracy and CE, Table 10 lists every seed row, and Figure 7 shows all development trajectories.

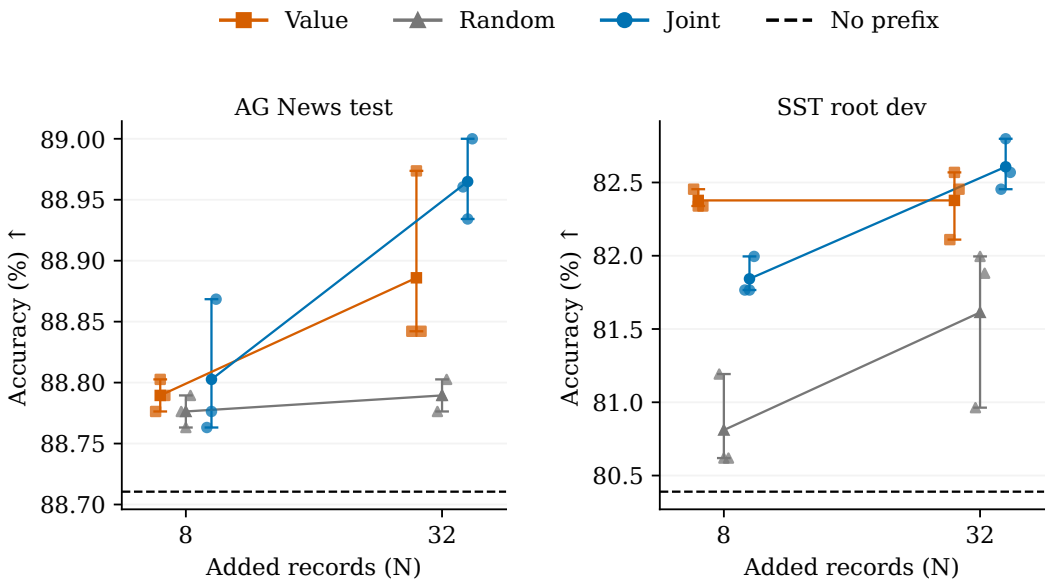


Figure 5: Every real-task training seed, mean and observed three-seed range (not a confidence interval). Black dashed lines show the common no-prefix reference. Joining the two tested counts does not establish an intermediate resource curve. Task panels use different vertical scales.

Descriptive geometry. Calibration diagnostics record principal-subspace alignment, captured Gram trace fractions, learned retrieval-logit changes, prefix attention mass and finite post-output-projection response changes. Joint selects its own Gram objective, so greater captured Joint energy is a construction check rather than independent validation. The mean squared principal cosine between Value and Joint ranges from 0.176 to 0.203, establishing that the tested bases differ without establishing which difference caused a task outcome. In particular, Joint also uses calibration information absent from static Value. The complete diagnostic values are retained alongside the paired CE and accuracy records.

Table 10: All 36 real-task fits. Accuracy is a percentage; CE uses natural logarithms. Step is the minimum-internal-development-CE checkpoint. No seed was removed.

Task	N	Seed	Arm	Accuracy	CE	Step
AG News	8	4101	Joint	88.763	0.34459	400
AG News	8	4102	Joint	88.776	0.34347	500
AG News	8	4103	Joint	88.868	0.34175	500
AG News	32	4101	Joint	88.961	0.34172	450
AG News	32	4102	Joint	88.934	0.34240	425
AG News	32	4103	Joint	89.000	0.34232	375
AG News	8	4101	Random	88.776	0.34947	500
AG News	8	4102	Random	88.763	0.34932	500
AG News	8	4103	Random	88.789	0.34878	500
AG News	32	4101	Random	88.776	0.34586	450
AG News	32	4102	Random	88.789	0.34811	400
AG News	32	4103	Random	88.803	0.34818	500
AG News	8	4101	Value	88.776	0.34787	500
AG News	8	4102	Value	88.803	0.34773	500
AG News	8	4103	Value	88.789	0.34845	500
AG News	32	4101	Value	88.842	0.34446	500
AG News	32	4102	Value	88.974	0.34406	500
AG News	32	4103	Value	88.842	0.34523	500
SST	8	4101	Joint	81.766	0.41403	225
SST	8	4102	Joint	81.766	0.41382	300
SST	8	4103	Joint	81.995	0.41177	350
SST	32	4101	Joint	82.454	0.40832	500
SST	32	4102	Joint	82.798	0.40149	500
SST	32	4103	Joint	82.569	0.40612	500
SST	8	4101	Random	81.193	0.42786	500
SST	8	4102	Random	80.619	0.43211	500
SST	8	4103	Random	80.619	0.43059	500
SST	32	4101	Random	80.963	0.43047	500
SST	32	4102	Random	81.995	0.42488	500
SST	32	4103	Random	81.881	0.42029	500
SST	8	4101	Value	82.454	0.40706	500
SST	8	4102	Value	82.339	0.41343	500
SST	8	4103	Value	82.339	0.41201	500
SST	32	4101	Value	82.110	0.40993	500
SST	32	4102	Value	82.569	0.41186	500
SST	32	4103	Value	82.454	0.41099	500

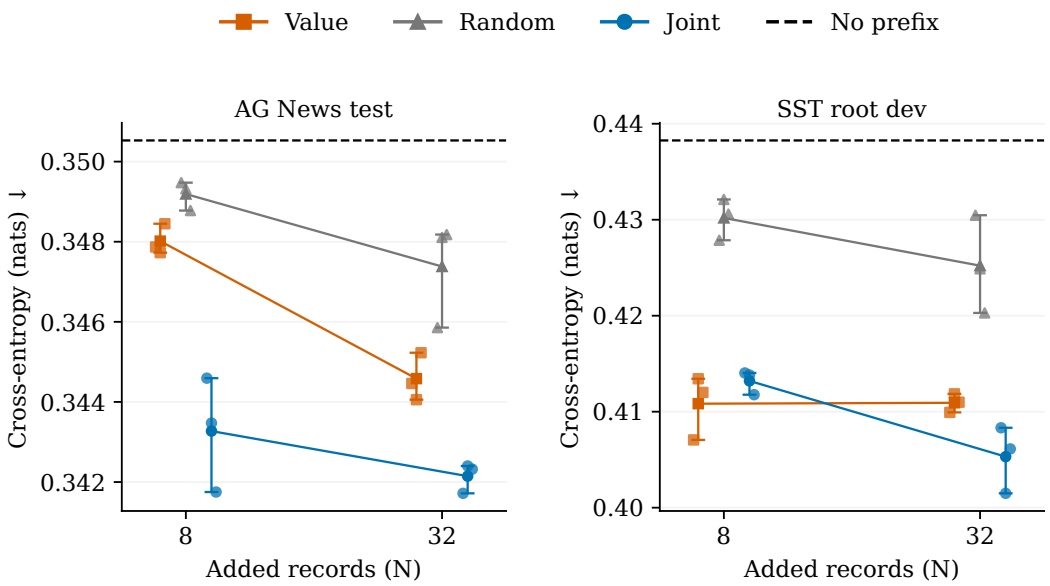


Figure 6: The same complete comparison in cross-entropy. Joint is lower in every AG News pair; SST reverses its mean ordering between the two counts. The no-prefix reference is a single common head/model evaluation, not three independently trained baselines.

Scope. The real labels make this evidence distinct from the earlier Energy/Wine projected-covariate studies with synthetic targets. Nevertheless, it is one frozen model, one intervention block, one common head and two record counts. The common head was trained without prefixes; no end-to-end adaptation baseline is established. Affine subspaces centered at Z_0 do not match the theorem's origin-centered library restriction. Mixed directions are retained, and no general retrieval-causality, Transformer-rate, state-of-the-art or computational-speed claim is drawn. Dataset raw text and pretrained weights are not redistributed with the manuscript; provenance and retrieval recipes are separate from compact result tables.

N Additional finite scalar displays

Table 11 gives the fresh uniform confirmation of Section 5.2 at every count.

Table 11: Fresh uniform confirmation, ten pairs per count. U denotes unrestricted and A antisymmetric. RMSE and grid-max columns are medians within each arm; $\widetilde{\Delta}_{\text{test}}$ is the median paired RMSE reduction. Wins count positive paired reductions. The primary row is $N = 32$; all other counts are secondary. Grid maxima use 4,097 points and are not uniform certificates.

N	$\widetilde{\text{RMSE}}^{U}_{\text{test}}$	$\widetilde{\text{RMSE}}^{A}_{\text{test}}$	$\widetilde{\Delta}_{\text{test}}$ (%)	$\widetilde{E}^{U}_{\text{grid}}$	$\widetilde{E}^{A}_{\text{grid}}$	U wins/10
16	2.693e-2	3.070e-2	+12.28	7.920e-2	8.643e-2	10/10
32	6.342e-3	1.256e-2	+49.41	2.391e-2	4.272e-2	10/10
64	1.906e-3	2.489e-3	+22.56	8.049e-3	1.093e-2	10/10
128	2.778e-4	5.046e-4	+39.76	1.678e-3	2.558e-3	10/10

Where the primary fits make errors. Figure 8 resolves the primary score into query-wise errors for every saved $N = 32$ fit. We replay the existing 4,097-point grid using the frozen fitted records; all 20 grid RMSE and maximum-error values reproduce the saved reports within numerical precision. No seed is selected and no model is refitted. Thin curves are individual fits; the thick pointwise median is a descriptive summary, not an additional fitted predictor. These curves expose error location

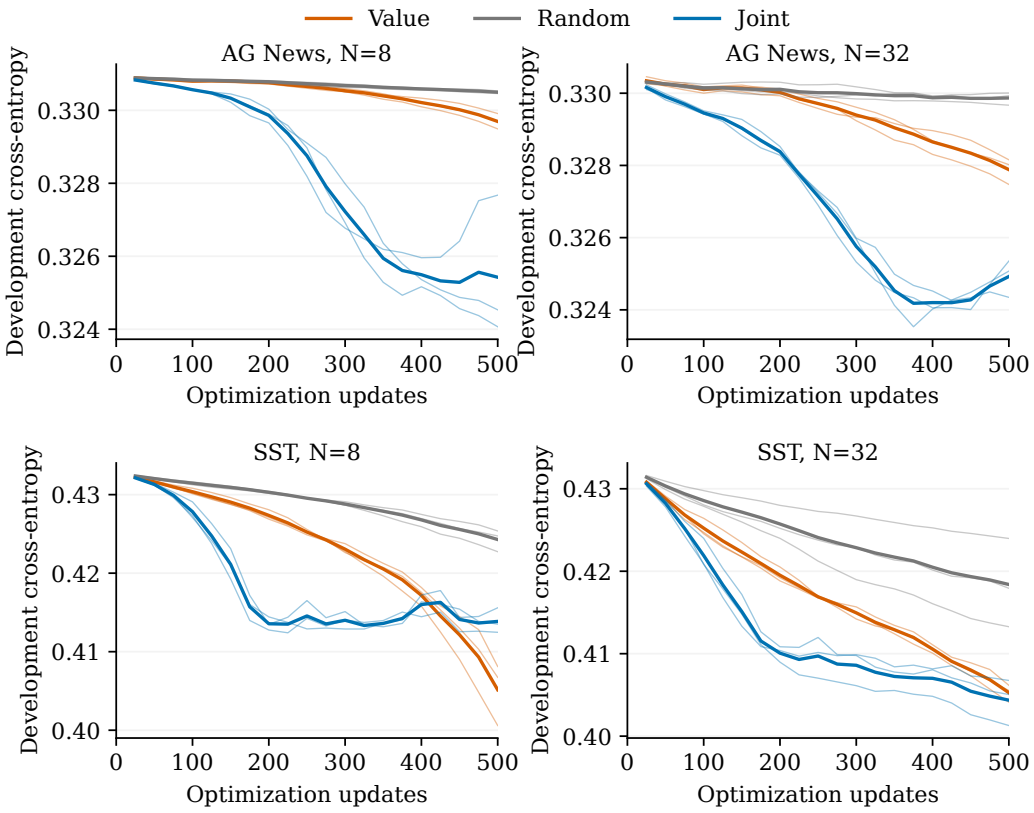


Figure 7: All development CE trajectories on the frozen 500-update budget. Curves describe optimization and checkpoint selection on internal development data; they are not additional held-out results. Endpoint selection in 27/36 fits prevents an optimizer-converged ranking claim.

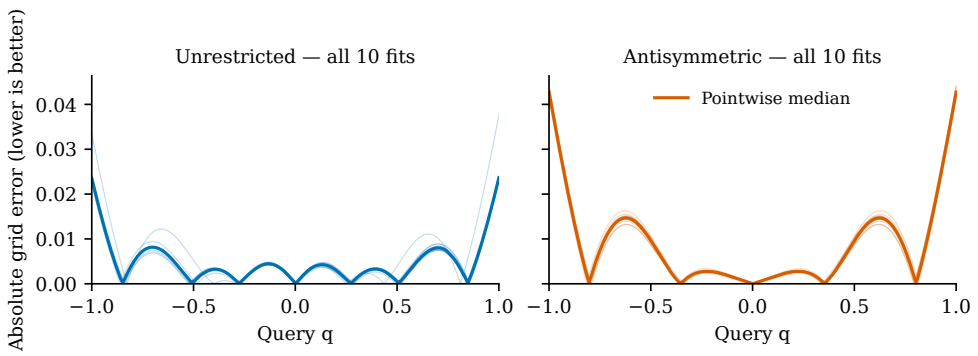


Figure 8: Primary-count errors across the query interval, with the same axes in both panels. All ten fits per arm are shown faintly, with a thick pointwise median. The target is $\tanh q$, the count is 32 and the gain is 16. These are replays of existing fits, not a new confirmation or a seed-selected example.

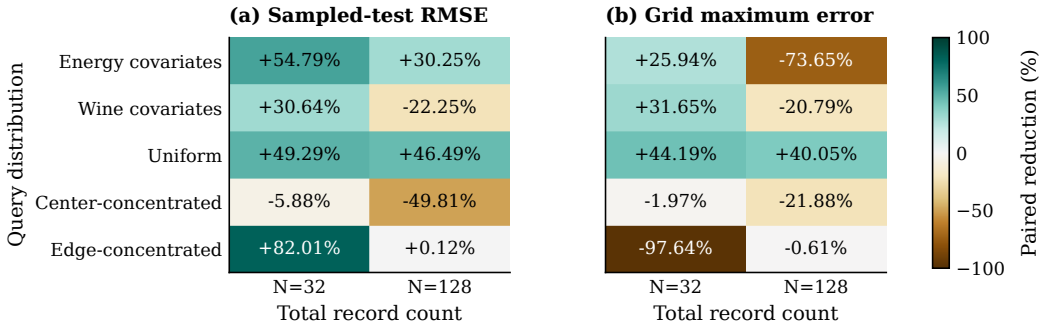


Figure 9: Distribution sensitivity across every exploratory condition, three paired seeds per cell. Left: median paired sampled-test RMSE reduction; right: median paired grid-maximum-error reduction. Both use $100(1 - U/A)$ for each seed before taking a median, with one common color scale. Positive favors unrestricted records. The two endpoints can disagree, notably for edge queries at $N = 32$. Energy/Wine use projected covariates and synthetic targets. Grid maxima are not uniform-error certificates; these are descriptive exploratory summaries.

and fit variability that a single RMSE hides; they do not identify a causal mechanism or certify error between grid points.

Additional metric distinctions. The primary median maximum error on a 4,097-point grid is 0.02391 for unrestricted records and 0.04272 for the restriction. A grid maximum is not a certified uniform error. Secondary median paired RMSE reductions are 12.28%, 22.56%, and 39.76% at $N = 16, 64, 128$. The smaller $N = 16$ effect is retained alongside all paired outcomes in Figure 10; the prespecified success rule is documented in Appendix D. Variation across counts also cautions against interpreting these fits as an estimate of an asymptotic exponent.

Sampled RMSE averages errors under the chosen query law, whereas the grid maximum tests the worst sampled location across the whole interval. Energy covariates at $N = 128$ show the same endpoint reversal: median paired reductions are $+30.25\%$ for test RMSE and -73.65% for grid maximum. These are different questions, and a favorable distributional score alone does not validate the theorem's uniform-error premise.

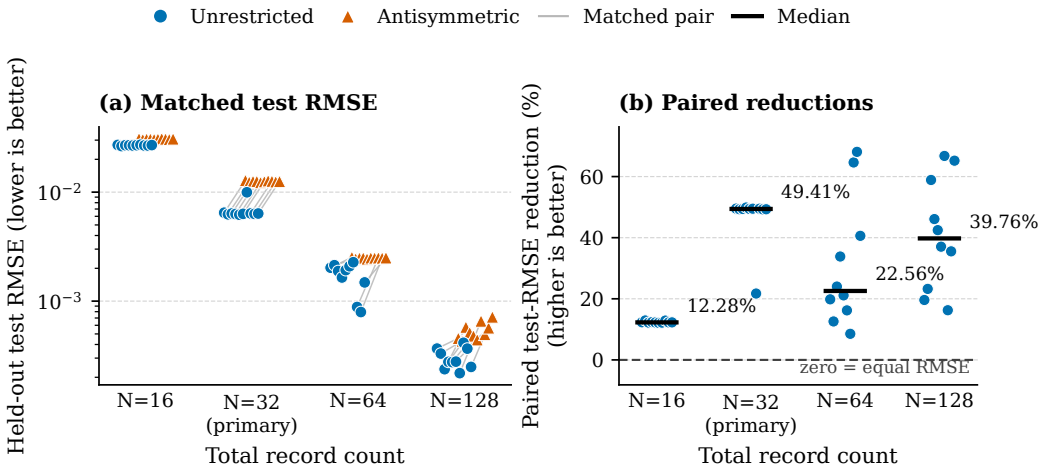


Figure 10: (a) Every held-out RMSE, on a logarithmic axis: blue circles are unrestricted, orange triangles antisymmetric, and a gray line joins each matched pair; lower is better. (b) Every paired reduction $100(1 - \mathrm{RMSE}_U/\mathrm{RMSE}_A)$; higher is better. Black bars and labels mark medians, not confidence intervals; the gray zero line marks equal errors. Jitter only separates points. The primary $N = 32$ mean-log-ratio bootstrap interval is $[0.5503, 0.6838]$; it concerns a different statistic from the percent medians shown.

O CALIBRATION-MATCHED VALUE-ROUTE DIAGNOSTIC

The primary comparison changes both the basis rule and its use of calibration contexts. After reading all primary outcomes, we specified one adapted diagnostic to isolate a narrower change: omit only the direct key-derived outer product from the local derivative used to construct Joint. The new value-route Jacobian basis (VJ) uses

$$D^{\mathrm{VJ}}_{ti} = \sum_h a_{thi} V_h O_h,$$

with exactly the same 32 calibration texts, eligible query positions, pair-specific Z_0, natural-plus-prefix softmax probabilities, output maps and cross-head summation as Joint in Appendix M. VJ is still attention-dependent; it is not a no-retrieval control. Both bases have rank 16. We fit only the 12 new VJ units using the original data order, common frozen head, norm constraint and 500-update optimizer protocol, and reuse all 12 saved Joint checkpoints. No original arm is refit or retuned.

Fresh evaluation and protocol. The adapted diagnostic uses 2,000 AG News examples (500 per class) and 1,000 SST root sentences (500 per class), drawn from unused normalized-exact-text groups in their original training sources. We exclude every prior fitting, calibration, development, smoke and official-evaluation text group. IDs, labels, source hashes and deterministic group rankings were frozen before VJ fitting. All 12 fits and checkpoint identities were verified before evaluation. These fresh training-source holdouts differ from the original official evaluation in Table 9; their absolute scores must not be pooled or compared as if the examples were identical. This is an outcome-informed, prospectively frozen follow-up, not independent confirmation of the original test results.

Complete result and decision. Table 12 reports every task/count mean, and Table 13 every seed pair. Joint has lower CE in all 12 pairs, including when calibration is matched. Accuracy moves in the other direction in three of four task/count means: by 0.033 percentage points on AG News at $N = 32$, and 0.8 and 0.9 points on SST at $N = 8, 32$. The prespecified composite support pattern required four favorable mean CE comparisons, at least 9/12 favorable CE pairs, and no more than one mean-accuracy reversal. The accuracy guard fails, so the composite pattern is *not met*. We report the loss-specific advantage together with the accuracy reversals, not broad task-quality superiority. Lower CE does not by itself establish improved calibration. All 12 VJ development-selected checkpoints

lie at step 500, reinforcing the fixed-budget interpretation. These observations concern the specified local basis procedure, not a general causal role of retrieval or an asymptotic count rate.

Table 12: Fresh adapted diagnostic, all three-seed means. Positive $\Delta\mathrm{CE} = \mathrm{CE}_{\mathrm{VJ}} - \mathrm{CE}_{\mathrm{Joint}}$ favors Joint. Accuracy is in percent; its difference is in percentage points. No seed confidence interval is claimed.

Task	N	CE Joint	CE VJ	ΔCE	Acc. Joint	Acc. VJ	ΔAcc.
AG News	8	0.334735	0.340617	+0.005882	89.383	89.250	+0.133
AG News	32	0.333234	0.335384	+0.002150	89.500	89.533	-0.033
SST	8	0.398717	0.409755	+0.011038	84.467	85.267	-0.800
SST	32	0.392784	0.403979	+0.011195	84.500	85.400	-0.900

Table 13: All 12 matched pairs on the same task-specific fresh holdout. CE differences are VJ minus Joint; accuracy differences are Joint minus VJ in percentage points.

Task	N	Seed	CE Joint	CE VJ	ΔCE	ΔAcc.	VJ step
AG News	8	4101	0.336559	0.340489	+0.003930	-0.05	500
AG News	8	4102	0.334962	0.340591	+0.005630	+0.15	500
AG News	8	4103	0.332686	0.340772	+0.008086	+0.30	500
AG News	32	4101	0.332669	0.335366	+0.002697	+0.05	500
AG News	32	4102	0.333552	0.335073	+0.001521	-0.05	500
AG News	32	4103	0.333482	0.335715	+0.002233	-0.10	500
SST	8	4101	0.398268	0.410385	+0.012117	-0.40	500
SST	8	4102	0.399658	0.409823	+0.010165	-1.10	500
SST	8	4103	0.398225	0.409058	+0.010833	-0.90	500
SST	32	4101	0.395342	0.404467	+0.009125	-1.00	500
SST	32	4102	0.390047	0.403516	+0.013468	-0.70	500
SST	32	4103	0.392963	0.403955	+0.010992	-1.00	500

Independent metric checks. All 24 saved-logit evaluations passed the identity/constraint audit and a separate NumPy recomputation of CE, accuracy and exact ID/label sets. The fitting protocol, bases and checkpoints remained frozen throughout evaluation; source identities and complete execution receipts are retained with the diagnostic record.

P CERTIFIED FINITE-AMPLITUDE SEPARATION

Theorems 2–3 are asymptotic, and the literal palette recipe in the proof is far from practical (Appendix D). This appendix shows that the separation already holds, rigorously, at moderate amplitude. For $Q = 1$ and $\epsilon = 1/4$ we exhibit explicit unrestricted admissible libraries with integer records and certified uniform error, and compare their counts with the finite restricted converse of Theorem 3 (Appendix B, Step 9), which holds at every m:

$$\log N_{\mathrm{anti}}(m,\delta) \;\geq\; \frac{m}{A} J\left(\left[t - 2\delta - \frac{A-1}{m}\right]_+\right).$$

Whenever an explicit unrestricted library has fewer records than the right side allows, $N_*(m,\delta) < N_{\mathrm{anti}}(m,\delta)$ holds at that m.

Libraries. Each library follows the structure of Appendix I. It has a shifted truncated-binomial diagonal base, with cutoff $\beta = (t-(A+1)d)/A$ for a slack $d < \delta$ and an integer shift $s = \mathrm{round}(dm)$. It adds a palette of sign/exchange orbits $\{(a,y),(y,a),(-a,-y),(-y,-a)\}$ whose keys sit at small integer offsets from the base's leading key, and the two designated anchors. The integer palette multiplicities are chosen by a mixed-integer program (HiGHS) that minimizes the palette count subject to the accuracy constraints on a query grid, with an iterative-refinement step that was available but never needed. The program is only a search heuristic: we claim no optimality of the counts, and a configuration that fails says nothing about existence.

Certificate. Admissibility is certified for the saved integer library, independently of how it was found. We evaluate the general two-head output, with no symmetry assumed, on the dyadic grid $q_j = j2^{-21}$ over $[-Q, Q]$. Between grid points, each head's derivative is a covariance under that head's attention, so Popoviciu's inequality gives $|F'| \le (\mathrm{range}_x \mathrm{range}_{v_1} + \mathrm{range}_y \mathrm{range}_{v_2})/8$, and $|\frac{d}{dq} m \tanh q| \le m$. The certified bound is the maximum grid error plus an explicit floating-point bound plus the Lipschitz term. The floating-point bound assumes relative error at most 8 ulp for $\exp$ and $\tanh$, with a safety factor of four, and is at most 2×10^{-11}. A 60-digit decimal re-evaluation at the worst and at uniformly spaced grid points agrees with the float64 values to within 10^{-13}. Every library also satisfies the unrestricted converse of Theorem 1, as it must. A separate verifier, written without access to the search code, re-checked every library on a grid twice as fine, with both heads over $[-Q, Q]$, its own Lipschitz and curvature bounds, and 60-digit interval arithmetic at the critical points. It agreed with all saved values; its certified bounds differ from the saved ones by at most 10^{-9}, owing to a more cautious floating-point allowance.

Result. At $\delta = 0.05$ the separation is certified for every tested m from 70 to 200 (Table 15, Figure 11). The restricted lower bound exceeds the explicit unrestricted count by a factor of 1.4 at $m = 70$, 20 at $m = 120$ and 450 at $m = 200$. At $m = 60$ our library is not small enough, so no separation is certified there. At $\delta = 0.1$ the separation is certified for every tested m from 60 to 200. Every certified error lies within $0.0022m$ of the tolerance. The palettes use at most twelve orbits, and the count is dominated by the binomial base. The conservative constants of the existence proof are therefore not needed at these amplitudes. They are upper bounds on N_*, not optimal counts.

Further parameter settings. A second study, with its own design fixed in writing before its main run, repeats the search and certificate at $(Q, \epsilon) = (2, 1/4)$, $(1, 1/2)$ and $(1, 1/8)$ with a runner generalized to these parameters (Table 14). Each of its 46 cells has a certified admissible unrestricted library, and 36 are certified separations. Separation holds at every tested amplitude from $m = 40$ for $(2, 1/4)$; from $m = 50$ ($\delta = 0.05$) and $m = 70$ ($\delta = 0.1$) for $(1, 1/2)$; and from $m = 120$ and $m = 80$ for $(1, 1/8)$, where the asymptotic gaps are smallest (0.049 and 0.039) and the ratio stays near one up to $m = 120$ at $\delta = 0.1$. At smaller tested m the libraries are admissible but not small enough, which again says nothing about N_{anti} there. A copy of the verifier, extended to read Q and ϵ after its author had read the generalized runner, agreed with every verdict and with 3,633 of 3,634 compared fields. The exception is a certified bound at $(1, 1/8)$, $\delta = 0.05$, $m = 200$ that differs by $1.04 \times 10^{-5} m$, just above its comparison tolerance, because the verifier's finer grid halves the Lipschitz term; both bounds are below δm.

Table 14: Certified finite separation at three further settings. "Ratio range" is $\exp\{\text{restricted bound}\}/N$ over the separated amplitudes, rounded down; every separated range is contiguous among the tested m.

(Q, ϵ)	δ	Tested m (cells)	Separated m	Ratio range	Max. orbits
$(2, 1/4)$	0.05	40–120 (7)	40–120	3.21–1,857	7
$(2, 1/4)$	0.1	40–120 (7)	40–120	1.81–449	5
$(1, 1/2)$	0.05	40–200 (9)	50–200	1.02–3,432	10
$(1, 1/2)$	0.1	40–200 (9)	70–200	1.48–954	6
$(1, 1/8)$	0.05	60–200 (7)	120–200	1.50–13.7	5
$(1, 1/8)$	0.1	60–200 (7)	80–200	1.09–22.9	7

Q RECORD-COUNT STUDY IN FROZEN PRETRAINED ENCODERS

This study asks the paper's resource question in pretrained models: at a fixed number N of shared key/value records, how much does confining the records to a low-dimensional subspace cost relative to unrestricted records? The main grid was fixed in writing before any fit, checks (i)–(iii) before their own fits, and check (iv), added after an audit showed restricted arms still improving, before its fits; all after the earlier studies. Every cell mean and paired count is reported below; per-unit rows are in the supplement.

Table 15: Certified finite-amplitude separation at $Q = 1$, $\epsilon = 1/4$. "Thm 3 bound" is the finite restricted converse $\exp\{(m/A)J([t - 2\delta - (A-1)/m]_+)\}$; "certified N" is the count of an explicit unrestricted library whose certified uniform error is at most δm. A ratio above one certifies $N_*(m, \delta) < N_{\mathrm{anti}}(m, \delta)$; † marks the one tested case without certified separation. The last column gives palette orbits and records.

m	Thm 3 restricted bound	certified N	bound/N	certified error/m	orbits/records
$\delta = 0.05$					
60	1.070×10^4	15 476	0.691†	0.04876	4/32
70	5.247×10^4	37 254	1.40	0.04893	3/28
80	2.573×10^5	89 468	2.87	0.04897	4/180
90	1.261×10^6	213 444	5.91	0.04901	3/220
100	6.187×10^6	1.323×10^6	4.67	0.04900	3/156
120	1.488×10^8	7.275×10^6	20.4	0.04902	7/1 452
150	1.755×10^{10}	2.414×10^8	72.7	0.04902	12/6 804
200	4.980×10^{13}	1.106×10^{11}	450	0.04802	4/6 752
$\delta = 0.1$					
60	6.521×10^2	624	1.04	0.09795	3/60
70	1.988×10^3	1 396	1.42	0.09789	3/64
80	6.061×10^3	3 250	1.86	0.09801	4/84
90	1.848×10^4	7 634	2.42	0.09784	3/124
100	5.635×10^4	17 862	3.15	0.09785	2/100
120	5.240×10^5	99 164	5.28	0.09798	3/316
150	1.485×10^7	1.280×10^6	11.6	0.09799	3/148
200	3.918×10^9	4.051×10^7	96.7	0.09801	3/144

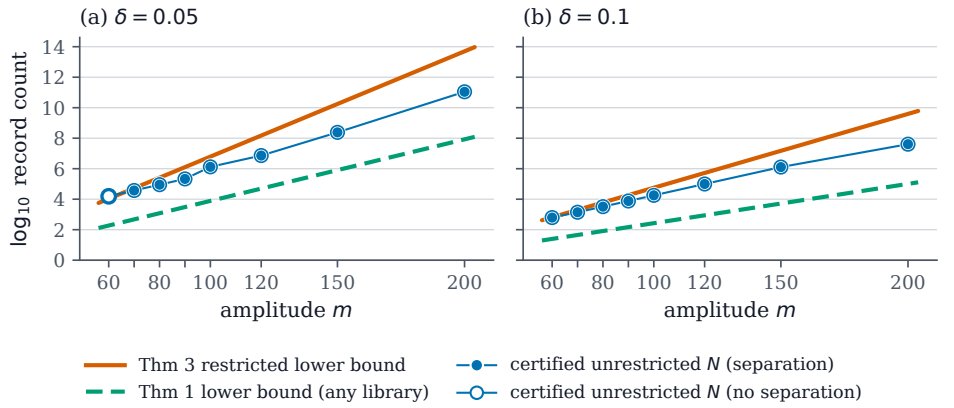


Figure 11: Record counts on a $\log_{10}$ scale: the finite restricted lower bound of Theorem 3 (orange), explicit certified unrestricted libraries (blue; filled where the separation is certified), and the lower bound of Theorem 1 for any library (green).

Interface. We use frozen RoBERTa-base (Liu et al., 2019) (the checkpoint and eager attention of Appendix M). At zero-based block 6, N record embeddings $Z = Z_0 + CB^\top$ supply extra keys and values through the block's frozen projections; they add no query or output positions, and natural context keys and values remain. Only C is trained. It starts at zero, so all arms begin from the identical function for each (N, seed); we checked this exactly (maximum output difference 0) at the smallest and largest N before training. After every update each row of C is projected into a common ball whose radius is the median calibration hidden-state norm. $Z_0 \sim \mathcal{N}(0, 0.02^2)$ uses the per-(N, seed) seeds of Appendix M. A frozen common linear head reads the first-token state, the head of Appendix M.

Arms. *Unrestricted*: $B = I_{768}$, so each record has 768 free coordinates. *Value-k*: B holds the top-k eigenvectors of the static value/output Gram $\sum_h (V_h O_h)(V_h O_h)^\top$, the pretrained analog of the value map's highest-gain eigenspace. *Random-k*: B is a Haar-random k-dimensional orthonormal basis. We use $k = 16, 64$. At equal N the unrestricted arm has more coordinates; that removal of freedom is exactly what the restriction does, as in the theorem. We therefore also compare arms at similar coefficient counts.

What a restriction constrains. In the prefix formalization of Petrov et al. (2024b, Sec. 4), records added at a block cannot change the relative attention among the natural tokens there; each head gains attention mass on the records and a term in the span of their values $z_j V_h$. With $Z = Z_0 + CB^\top$ and Z_0 fixed, the trainable part of both the records' keys and their values therefore lies in the images of $\mathrm{col}(B)$ under the key and value projections. A restricted arm thus confines where records can draw attention and which directions they can return; this study does not separate the two effects.

Training and evaluation. AdamW (learning rate 0.01, zero weight decay), batch 16, maximum length 128, dropout disabled, 2000 updates. Every 100 updates the internal development CE (500 train-source examples) is recorded, and the checkpoint minimizing it is selected. Records $N \in \{1, 2, 4, 8, 16, 32, 64\}$ and paired seeds 4101–4103 give 105 units per task. Held-out sets (AG News: the 7,600 official test examples; SST-2: the 872 official binary development sentences) were evaluated once per unit, only after every unit of a study had finished training and its selected checkpoints were fixed. These evaluation sets were also used by the earlier diagnostics of Appendices M–O, which studied other arms and fixed the block, learning rate and head reused here; no fit, checkpoint selection or stopping rule in this study used them. We also record the training-split CE (2,000 examples) at the final state, as a capacity measure.

Validity checks. (i) Learning rate: the unrestricted and Value-16 arms at $N \in \{1, 8, 64\}$ are refit with learning rates 0.003 and 0.03 (one seed). (ii) Larger N: the restricted arms are extended to $N \in \{128, 256\}$ (three seeds). (iii) Intervention block: blocks 3 and 9 for the unrestricted, Value-16 and Random-16 arms at $N \in \{1, 4, 16, 64\}$ (three seeds). (iv) Training budget: one unrestricted record, Value-64 at $N \in \{1, 64, 256\}$ and Random-64 at $N \in \{64, 256\}$ are refit for 10,000 updates with development evaluation every 250 (three seeds). A separate audit program, written without the training code, rechecks every unit: saved checkpoints, coefficient counts and ranks, row norms against the radius, development selection, and held-out metrics recomputed from saved per-example predictions against the evaluation labels.

Results. Table 16 lists the mean held-out CE and accuracy of every RoBERTa-base cell, and Table 17 the smallest restricted N whose mean held-out CE reaches that of one unrestricted record. On SST-2 one unrestricted record reaches CE 0.260 and accuracy 88.8%, against 0.438 and 80.4% without records. Within the 2000-update budget, no 16- or 64-dimensional arm reaches that CE with up to 256 records. The 64-dimensional arms do reach similar accuracy at $N = 256$ (Value-64 88.7%, Random-64 88.8%). By seed, the unrestricted arm has lower CE than Value-16, Value-64 and Random-16 in all 21 pairs each and than Random-64 in 20 of 21. The restricted arms underfit: at $N = 256$ their training CE is 0.209 (Value-64) and 0.221 (Random-64), against 0.181 for one unrestricted record, and most restricted SST-2 units (Value-64: 20 of 21) select their final checkpoint, whereas no unrestricted unit does. At similar coefficient counts, one unrestricted record (768 coefficients) against 16 Random-64 or Value-64 records (1,024) gives 0.260 versus 0.288 and 0.321 on SST-2, and 0.324 versus 0.340 and 0.333 on AG News.

On AG News (no records: 0.351 and 88.7%) one unrestricted record gives 0.324 and 89.5%. The best restricted mean is 0.330 (Random-64 at $N = 256$ and Value-64 at $N = 4$), and at each N the arms differ by at most 0.021 nats. Unrestricted records overfit as N grows (training CE 0.034 and held-out CE 0.331 at $N = 64$). By seed, unrestricted CE is lower than Random-16 in 21 of 21 pairs, than Value-16 in 20, than Random-64 in 18 and than Value-64 in 14.

Learning rate. At learning rates 0.003 and 0.03 (seed 4101), the unrestricted arm has lower held-out CE than Value-16 at every tested N on SST-2. On AG News it does so except at learning rate 0.03 and $N = 64$, where Value-16 is lower (0.330 versus 0.333).

Intervention block. On SST-2 the ordering holds at blocks 3 and 9: unrestricted CE is lower than Value-16 and Random-16 in all 24 pairs at each block (block 3: 0.326–0.353 versus 0.410–0.432; block 9: 0.267–0.278 versus 0.364–0.405). On AG News block 3 agrees (0.324–0.326 versus 0.347–0.350), but at block 9 Value-16 has lower CE than unrestricted records for $N \geq 4$ (0.327 versus 0.332 at $N = 16$; unrestricted lower in only 4 of 12 pairs), although unrestricted accuracy is higher in 11 of 12 pairs.

Training budget. The cells of check (iv) were chosen after the main-grid held-out results were known, as the restricted arms closest to one unrestricted record, and their number was limited by a compute cap; the 16-dimensional arms were not refit. The unrestricted record selects early checkpoints (steps 1000–2000) and reaches CE 0.263 and accuracy 89.3% (0.260 and 88.8% with 2000 updates). Value-64 and Random-64 at $N = 64$ and 256 select steps 2250–5500, well before the end, and reach 0.287, 0.278, 0.271 and 0.290 with accuracy 89.0–89.5%. Paired by seed, the unrestricted record has lower CE in all 12 pairs and higher accuracy in 6. Value-64 at $N = 1$ is still improving at 10,000 updates (0.359 and 85.1%).

Table 16: RoBERTa-base, block 6: mean held-out CE/accuracy (%) over three paired seeds for every arm and record count ($N = 128, 256$ only for restricted arms). No records: AG News 0.351/88.7, SST-2 0.438/80.4.

Task	Arm	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
AG News	Unrestricted	0.324/89.5	0.328/89.5	0.327/89.3	0.333/89.3	0.332/89.3	0.332/89.4	0.331/89.4	–	–
	Value-16	0.343/89.0	0.342/89.0	0.342/89.1	0.340/89.1	0.338/89.2	0.337/89.2	0.334/89.2	0.331/89.3	0.331/89.3
	Random-16	0.341/89.0	0.347/88.7	0.347/88.7	0.344/88.9	0.345/88.8	0.346/88.8	0.339/89.0	0.343/88.9	0.341/88.9
	Value-64	0.331/89.3	0.331/89.3	0.330/89.3	0.332/89.3	0.333/89.2	0.333/89.2	0.334/89.2	0.334/89.2	0.333/89.2
	Random-64	0.332/89.1	0.334/89.3	0.335/89.1	0.335/89.2	0.340/88.9	0.339/88.9	0.334/89.1	0.333/89.1	0.330/89.2
SST-2	Unrestricted	0.260/88.8	0.254/89.2	0.254/89.6	0.265/89.6	0.253/89.6	0.261/89.3	0.263/90.1	–	–
	Value-16	0.403/82.6	0.398/83.0	0.391/83.3	0.384/83.8	0.380/83.8	0.379/83.7	0.362/82.9	0.361/82.9	0.357/83.9
	Random-16	0.426/81.2	0.418/81.9	0.416/82.1	0.404/82.5	0.410/82.4	0.402/82.5	0.408/82.7	0.386/83.3	0.383/83.3
	Value-64	0.363/84.2	0.359/84.1	0.336/85.5	0.321/86.7	0.321/86.7	0.315/86.8	0.306/87.2	0.298/87.3	0.281/88.7
	Random-64	0.345/84.8	0.332/86.1	0.313/86.3	0.315/86.4	0.288/88.5	0.293/87.8	0.281/88.4	0.279/88.4	0.282/88.8

Table 17: Records to match one unrestricted record: the smallest tested N at which the mean held-out CE of each restricted arm shown is at most that of the unrestricted arm at $N = 1$ ($>$ the largest tested N if never; – not run). The 10k row uses the budget check (iv).

Backbone	Task	Budget	Unrestricted $N = 1$	Value-16	Value-64	Random-64
RoBERTa-base	SST-2	2k	0.260	> 256	> 256	> 256
RoBERTa-base	AG News	2k	0.324	> 256	> 256	> 256
RoBERTa-base	SST-2	10k	0.263	–	> 256	> 256