FROM PROPOSAL COVERAGE TO MINT: A UNIFIED ANALYSIS OF RANDOM-FEATURE SOFTMAX ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient softmax approximation is important across deep learning, particularly for computing attention in large language models. Random-feature methods approximate softmax attention at linear cost, but can miss regions that contribute strongly to query–key interactions. Reducing kernel error or large-sample output variance need not reduce output error at a fixed feature count. To precisely characterize the approximation error in attention computation and make related work comparable, we unify standard positive features and their geometry-adapted variants by relating kernel error to proposal coverage: how well sampling covers contributing regions. Attention normalizes these kernel estimates before combining values, so output error depends on changes in relative weights and differences between values. Following these errors through both operations yields an exact expression for expected squared output error at a fixed feature count under independent sampling. This leads to Mixture Importance-Tilted random features (MINT), which covers multiple contributing regions with a Gaussian mixture. MINT# additionally adapts the component shapes. Mixture weights are fitted using a numerical approximation of this expected squared output error at the chosen feature count. Importance correction preserves the softmax kernel in expectation. With component locations and shapes fixed, this weight calibration reduces output error compared with calibration based on coverage or large-sample output variance. At matched feature counts, MINT and MINT# improve next-token accuracy by 4.14–7.44 and 2.86–5.37 percentage points, respectively, over FAVOR# when replacing 25% of Qwen3-4B-Base's query heads. On OLMo 3, MINT# lowers mean error relative to FAVOR# in all 15 attention, logit, and decoding comparisons using six fixed heads. Code at: `https://anonymous.4open.science/r/mint-submission`.

1 INTRODUCTION

Softmax attention combines sequence information by weighting value vectors with query–key similarities, at quadratic cost in sequence length. Linear attention reuses key–value summaries for linear cost and fixed-size decoding states at fixed dimensions (Katharopoulos et al., 2020; Yang et al., 2025b). Optimized implementations demonstrate training and generation speedups (Yang et al., 2024; Zhang et al., 2025). Random-feature methods obtain this structure by approximating the exponential query–key similarity, or *softmax kernel*, with feature inner products (Choromanski et al., 2021; Peng et al., 2021). The sampled set of M features is a *feature bank*, and M is its *feature budget*. Keeping this bank fixed lets causal decoding update the summaries as tokens arrive.

At a fixed feature budget, a central goal of random-feature design is to improve the feature inner products that estimate the softmax kernel. Starting from FAVOR+'s positive exponential features, FAVOR++ and FAVOR# optimize scalar and matrix parameters to reduce kernel variance (Likhosherstov et al., 2022; 2023). Structured draws further reduce the estimation error by coordinating the sampled vectors (Reid et al., 2023). Importance sampling adapts the distribution of each draw, called a *proposal*, and corrects each feature's contribution to preserve the kernel in expectation. Sernau et al. (2024) derive proposals that minimize input-averaged kernel variance. These designs assess the accuracy of unnormalized kernel estimates. Attention, however, divides each estimate by

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

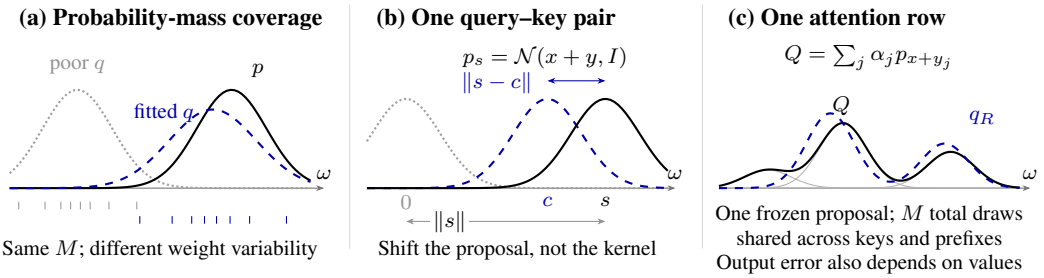

Figure 1: **Why sampling must cover attention's contribution regions.** Target densities show where kernel integrals receive their mass. (a) Poor overlap between proposal q and target p produces variable importance weights; rugs show $M = 8$ illustrative feature draws ω. (b) Query x and key y give target center $s = x + y$; c is a proposal center, distances are Euclidean, and I is the identity covariance. (c) Attention weights α_j combine pair targets into Q; output error also depends on values. Solid curves are targets, dashed curves fitted proposals, and dotted curves origin/poor references. MINT's frozen q_R has R components and shares M total draws across rows.

a shared *normalizer*, the sum over visible keys, to obtain weights. Kernel-level objectives therefore do not directly describe how approximation errors propagate through normalization to the output.

This normalization changes which errors matter: a common rescaling of all kernel estimates cancels, while errors in their relative weights affect the output through differences between values. RA/LARA incorporates both effects through self-normalized importance sampling with an attention-weighted Gaussian mixture target (Zheng et al., 2022, Section 3.1 and Appendix D). Its first-order output-variance approximation does not fully capture the interaction between the value-weighted sum and random normalizer at finite budgets. Designing a shared proposal therefore requires understanding how the allocation of sampling probability affects errors in the value-weighted sum and its normalizer, and how these errors jointly determine the output.

To guide the design of shared proposals using finite-budget output error, we develop a unified proposal-coverage analysis (Section 3). Coverage measures sampling probability in regions contributing to the kernel integral. Figure 1 contrasts coverage at equal budgets (a), shows each pair's shifted Gaussian target (b), and combines these targets across keys (c). With independent and identically distributed (iid) draws, attention-weighted mean squared relative kernel error separates into normalizer error and keywise variation. Value differences weight these deviations to give the output residual before division and a finite-budget output bound. Retaining random division gives an exact output-risk representation for calibration, whose scaled large-budget limit is the value-centered variance coefficient. The framework also represents FAVOR+, FAVOR++, and FAVOR#'s frozen symmetric dense-exponential random features (SDERF) as Gaussian proposals for the same softmax kernel, allowing their kernel, normalizer, and output errors to be compared at matched feature budgets (Table 1).

This view motivates Mixture Importance-Tilted random features (MINT), which address location mismatch with multiple centers and allocate probability using finite-budget output risk. Origin-centered Gaussians adjust spread; mixtures can also place peaks near separated contribution regions. Their coverage bound depends on center distances and component probabilities. MINT fits centers with a coverage-based clustering surrogate. Its variant MINT# additionally fits a shared covariance to the center residuals. With geometry fixed, we fit weights to numerically approximated output risk at budget M, accounting for values and random normalization (Section 4).

We make three contributions: 1. **Unified error analysis.** To connect kernel approximation quality to attention output accuracy, we develop a unified coverage-based analysis that follows errors through the attention computation. It yields an upper bound and an exact expression for finite-budget output risk under independent sampling, supporting matched-budget comparisons of random-feature methods (Section 3). 2. **MINT and MINT#.** Guided by the analysis, we propose Mixture Importance-Tilted random features (MINT) and its variant MINT#. Multiple centers cover separated contribution regions, and output-risk calibration allocates sampling probability at the deployment budget. MINT# further adapts component shapes (Section 4). 3. **Synthetic and LLM experiments.** Syn-

thetic experiments validate our theoretical analysis. LLM experiments demonstrate gains in next-token prediction accuracy over FAVOR baselines at matched feature budgets (Section 5).

2 Related Work

Random features and attention adaptation. Beyond kernel random features (Rahimi & Recht, 2007) and LARA's input-adaptive proposals, EVA uses value-dependent control variates to combine local exact attention with grouped approximations, including a causal variant (Zheng et al., 2023); FlashEVA develops efficient implementations (Kostelec & Guo, 2025). Learned kernels (Pal Chowdhury et al., 2022; Farzam et al., 2026), fitted feature distributions (Badiei Khuzani et al., 2026), and output-matched linearization (Zhang et al., 2025) adapt attention during training. MINT preserves the softmax kernel and freezes its fitted features before evaluation.

Importance sampling. Importance-weight second moments give finite-sample bounds (Agapiou et al., 2017), while adaptive self-normalized samplers target value-centered asymptotic variance (Branchini & Elvira, 2025). At the kernel level, Sernau et al. (2024) establish representation-independent optimal variance for independent inputs from the same distribution. We retain the random normalization in attention's finite-budget output risk. An established Laplace representation (de la Peña et al., 2025) makes this risk numerically accessible for calibrating a shared proposal.

3 From Proposal Coverage to Finite-Budget Output Risk

Proposal coverage links sampling choices to attention approximation error. Sections 3.1–3.3 trace how kernel errors combine in the shared normalizer and propagate through normalization and value aggregation to finite-budget output risk. Section 3.4 then compares feature constructions through their proposal geometry: covariance adaptation changes spread, while multiple centers can target separated contribution regions. Together, these results guide MINT's geometry and its allocation of sampling probability at the deployment budget.

We fix the attention inputs and draw a bank of M feature vectors $\omega_m \in \mathbb{R}^d$ independently from a proposal density q. Queries x and keys y include attention scaling, and d is their head dimension. Expectations $\mathbb{E}$ average over banks conditional on the inputs and fitted proposal; $\mathbb{E}_q$ averages over one draw. The norm $\|\cdot\|$ is Euclidean. The variance identities require $q > 0$ wherever the targets have mass and finite second moments; proofs and further moment conditions appear in Appendix A.

3.1 From a pair to a shared denominator

Standard positive features express the softmax kernel $K(x, y) = e^{x^\top y}$ through the normalized Gaussian contribution density $p_s = \mathcal{N}(s, I_d)$, where $s = x + y$ and I_d is the identity covariance. This density shows where features contribute to the kernel integral, while q determines where features are drawn (Figure 1b). Importance correction preserves the expected kernel value, but poorly covered regions produce rare, large contributions and high variance.

At a fixed feature budget, coverage determines a kernel estimate's relative mean-squared error. For a target density p, the mismatch is $\chi^2(p\|q) = \int p(\omega)^2/q(\omega)\, d\omega - 1$. The ratio penalizes regions that matter to the target but receive too little sampling probability. For an estimator $\widehat{T}$ of a nonzero scalar T, define relative mean-squared error as $\mathrm{RelMSE}(\widehat{T}) = \mathbb{E}[(\widehat{T} - T)^2]/T^2$. The importance-corrected estimate $\widehat{K}_{q,M}$ then satisfies $\mathrm{RelMSE}(\widehat{K}_{q,M}) = \chi^2(p_s\|q)/M$. For the origin proposal $\gamma = \mathcal{N}(0, I_d)$, the coefficient $e^{\|s\|^2} - 1$ grows exponentially with squared target distance.

Shared normalization obeys the same coverage law, with a target that averages contributions across visible keys. For L keys indexed by j, let $K_j = K(x, y_j)$, $D = \sum_j K_j$, $\alpha_j = K_j/D$, and $p_j = \mathcal{N}(x + y_j, I_d)$. The shared target Q and attention-weighted pairwise coefficient C_{pair} are

$$Q = \sum_j \alpha_j p_j, \qquad C_{\mathrm{pair}}(q) = \sum_j \alpha_j \chi^2(p_j\|q). \tag{1}$$

Thus keys with larger exact attention weights contribute more to Q. Write $\widehat{K}_j = \widehat{K}_{q,M}(x, y_j)$ and $\widehat{D} = \sum_j \widehat{K}_j$.

Theorem 1 (Shared-denominator coverage and error decomposition)**.** *For a fixed row, assume $q > 0$ wherever $Q > 0$, up to null sets, and $C_{\mathrm{pair}}(q) < \infty$. A bank of M iid draws shared by all keys satisfies*

$$\mathrm{RelMSE}(\widehat{D}) = \frac{\chi^2(Q\|q)}{M} \le \frac{C_{\mathrm{pair}}(q)}{M}. \tag{2}$$

The pairwise coefficient separates exactly into normalizer error and variation across keys:

$$C_{\mathrm{pair}}(q) = \chi^2(Q\|q) + \mathbb{E}_q \sum_j \alpha_j \left(\frac{p_j - Q}{q}\right)^2. \tag{3}$$

After division by M, the first term measures error in the shared sum. The second records differences between individual keys' contributions and their weighted average; these differences matter when attention combines values. The proof is in Appendix A.1.

3.2 How normalization and values shape output error

Normalization cancels common kernel errors, and value differences determine the effect of errors that remain. Doubling every kernel estimate leaves attention weights unchanged. Likewise, if every value vector is identical, changing the weights cannot change the output. These cancellations explain why pairwise or normalizer error alone does not determine output error.

Output error is the ratio of a centered residual to a random normalizer, both computed from the same bank. For values $v_j \in \mathbb{R}^{d_v}$, where d_v is the value dimension, the exact output is $o = \sum_j \alpha_j v_j$ and its estimate is $\widehat{o} = (\sum_j \widehat{K}_j v_j)/\widehat{D}$. Let $\widehat{d} = \widehat{D}/D$ be the normalizer relative to its exact value, and $\widetilde{e} = (\sum_j \widehat{K}_j(v_j - o))/D$ the centered residual before division. The output error and its finite-budget risk are

$$\widehat{o} - o = \frac{\widetilde{e}}{\widehat{d}}, \qquad R_M(q) = \mathbb{E}\|\widehat{o} - o\|^2 = \mathbb{E}\frac{\|\widetilde{e}\|^2}{\widehat{d}^2}. \tag{4}$$

The risk therefore depends on how these quantities fluctuate together.

To bound output risk, we first quantify the residual before division. Pairing the keywise deviations in (3) with value differences gives the centered contribution G and its second-moment coefficient J_o:

$$G = \sum_j \alpha_j (v_j - o)(p_j - Q), \qquad J_o(q) = \int \frac{\|G(\omega)\|^2}{q(\omega)}\, d\omega. \tag{5}$$

Then $\mathbb{E}\|\widetilde{e}\|^2 = J_o(q)/M$.

Value dispersion links the keywise variation in (3) to output risk. Let $\sigma_v^2 = \sum_j \alpha_j \|v_j - o\|^2$ be the attention-weighted value dispersion and $\rho_v = \max_j \|v_j - o\|$ the maximum value deviation.

Proposition 1 (Finite-budget output error bound)**.** *Suppose $Q/q > 0$ q-almost surely and $J_o(q) + \chi^2(Q\|q) < \infty$. If $C_{\mathrm{pair}}(q) < \infty$, then*

$$J_o(q) \le \sigma_v^2 \big[C_{\mathrm{pair}}(q) - \chi^2(Q\|q)\big]. \tag{6}$$

Together with normalizer variability, J_o bounds the iid output risk for every $M \ge 1$:

$$R_M(q) \le \frac{\left(\sqrt{J_o(q)} + \rho_v \sqrt{\chi^2(Q\|q)}\right)^2}{M}. \tag{7}$$

Pairwise coverage thus controls output risk through both the value-weighted residual and shared-normalizer fluctuations (proof in Appendix A.2).

3.3 A FINITE-BUDGET OBJECTIVE FOR SHARED PROPOSALS

The bound controls residual and normalizer variability separately. Calibration with R_M retains their joint fluctuations through the random division. A one-dimensional integral makes this risk numerically accessible. For one draw $\omega \sim q$, let $W = Q(\omega)/q(\omega) > 0$ and $Z = G(\omega)/q(\omega)$ be its normalizer and centered residual contributions. For $t > 0$, define the Laplace expectations

$$a(t) = \mathbb{E}_q[e^{-tW}], \qquad b(t) = \mathbb{E}_q[Ze^{-tW}], \qquad c(t) = \mathbb{E}_q[\|Z\|^2 e^{-tW}].$$

A Laplace representation (de la Peña et al., 2025) then gives, for M iid draws,

$$R_M(q) = \int_0^\infty t\left[Mc(t)a(t)^{M-1} + M(M-1)\|b(t)\|^2 a(t)^{M-2}\right] dt. \tag{8}$$

The expectations couple Z and W from the same draw. Shared integration samples approximate them for numerical calibration at the deployment budget. For $M = 1$, omit the second term; Proposition 4 (Appendix A.3) gives the proof.

A shared proposal must balance output error across the rows it serves. Rows sharing one frozen proposal and bank form a *reuse group*, such as one head across prefixes and documents. For calibration rows i, let $R_{M,i}$ and $J_{o,i}$ be their risks and residual coefficients. Fixed nonnegative weights β_i normalize calibration output energy (Section 4.2), giving the shared objective

$$\mathcal{F}_M(q) = M \sum_i \beta_i R_{M,i}(q). \tag{9}$$

Multiplying by M leaves the minimizer unchanged; sharing a bank across rows preserves this sum by linearity of expectation.

The large-budget limit recovers the value-centered variance objective. For fixed q, the fourth-moment conditions in Appendix A.2 give $\mathcal{F}_M(q) \to \sum_i \beta_i J_{o,i}(q)$ as $M \to \infty$. At finite M, $\mathcal{F}_M$ retains the effect of random normalization.

3.4 A COMMON GAUSSIAN DESIGN SPACE

Gaussian proposals expose complementary ways to improve coverage at a fixed budget: covariance adjusts spread, while multiple centers can place probability peaks near separated contribution regions. For the same softmax kernel, FAVOR+, FAVOR++, and frozen SDERF in FAVOR# correspond to origin-centered Gaussians with unit, scalar, and full covariance, respectively (Appendix B). The following coverage laws quantify these choices and motivate MINT's multiple centers.

With unit covariance, moving a proposal center closer to the target center improves pairwise coverage. Define log coverage error $\ell_q(s) = \log(1 + \chi^2(p_s\|q))$; lower values mean lower pairwise RelMSE at a fixed budget. For a unit-covariance Gaussian centered at c,

$$\ell_{\mathcal{N}(c,I_d)}(s) = \|s - c\|^2. \tag{10}$$

A general covariance Σ changes the distance metric and adds a determinant penalty. For the offset $z = s - c$, define $h_\Sigma(z) = \ell_{\mathcal{N}(0,\Sigma)}(z)$. When $2\Sigma - I_d \succ 0$ (positive definite),

$$h_\Sigma(z) = \log\det\Sigma - \frac{1}{2}\log\det(2\Sigma - I_d) + z^\top(2\Sigma - I_d)^{-1}z. \tag{11}$$

The first two terms form the determinant penalty; the last measures the squared offset using the metric $(2\Sigma - I_d)^{-1}$. Table 1 uses this expression for each covariance choice (derivation in Appendix A.4).

Multiple centers can cover separated contribution regions that a single origin-centered Gaussian can only reach by changing its spread. For component centers c_r and weights π_r, the mixture bounds in Table 1 balance distance against the penalty $-\log\pi_r$: a nearby center helps only if its component receives enough probability. For MINT, a component bound below the origin value $\|s\|^2$ certifies better pairwise coverage (Lemma 1).

Table 1: **Gaussian proposals unify feature geometry and coverage.** Origin-centered Gaussians adapt covariance; mixtures also fit centers and weights. Lower log coverage error $\ell_q(s)$ is better; $=$ denotes exact expressions and $\leq$ upper bounds. See (11) for h_Σ and Appendix B for conditions.

Method	Proposal / geometry	Coverage dependence: $\ell_q(s)$	Causal
Frozen proposals and feature maps			
FAVOR+	$\mathcal{N}(0, I_d)$	$= \|s\|^2$	Yes
FAVOR++	$\mathcal{N}(0, \sigma^2 I_d)$	$= h_{\sigma^2 I_d}(s)$	Yes
FAVOR# (SDERF)	$\mathcal{N}(0, \Sigma)$	$= h_\Sigma(s)$	Yes
MINT	$\sum_r \pi_r \mathcal{N}(c_r, I_d)$	$\leq \min_r\{\|s - c_r\|^2 - \log \pi_r\}$	**Yes**
MINT#	$\sum_r \pi_r \mathcal{N}(c_r, \Sigma_{\text{res}})$	$\leq \min_r\{h_{\Sigma_{\text{res}}}(s - c_r) - \log \pi_r\}$	**Yes**
Input-adaptive proposals (outside the frozen-bank setting)			
RA	Query-specific proposal	—	No†
LARA	Shared adaptive proposals	—	No†

†Original versions lack causal implementations; dashes mark input-adaptive proposals (Appendix D.1).

4 MINT: From Coverage to Reusable Proposals

MINT calibrates shared proposals using the finite-budget output objective $\mathcal{F}_M$ from Section 3.3. Its Gaussian-mixture family follows the coverage analysis in Section 3.4: multiple centers target separated contribution regions, and covariance adapts their shapes. Calibration fits geometry with a coverage surrogate, then optimizes weights using output risk at the deployment budget.

4.1 Coverage-based component construction

Geometry fitting targets the contribution centers $s = x+y$ from Section 3.1; (10) turns their coverage into a squared-distance cost. For R Gaussian components with centers $c_r \in \mathbb{R}^d$ and mixture weights $\pi_r > 0$ summing to one, set $q = q_R$ and define

$$q_R(\omega) = \sum_{r=1}^{R} \pi_r \mathcal{N}(\omega; c_r, I_d), \qquad \phi_q(x;\omega) = \sqrt{\frac{\gamma(\omega)}{q(\omega)}} e^{\omega^\top x - \|x\|^2/2}, \quad \omega \sim q. \tag{12}$$

$R = 1$ gives MINT-1; multiple centers give MINT-R. Correction by the *full mixture density* preserves the kernel: $\mathbb{E}_q[\phi_q(x;\omega)\phi_q(y;\omega)] = e^{x^\top y}$.

The component law (10) gives this bound (proof and output implications in Appendix C).

Lemma 1 (Mixture coverage)**.** *For q_R in* (12)*,*

$$\ell_{q_R}(s) \leq \min_{r \leq R}\{\|s - c_r\|^2 - \log \pi_r\}. \tag{13}$$

Location fitting. For calibration targets $s_t = x_t + y_t$ and nonnegative weights w_t summing to one, averaging (13) gives

$$\sum_t w_t \ell_{q_R}(s_t) \leq \sum_t w_t \min_r\{\|s_t - c_r\|^2 - \log \pi_r\}.$$

For uniform $\pi_r = 1/R$, the right-hand side is weighted k-means plus $\log R$. Clustering initializes centers and positive frequency weights. As in log-variance optimization (Likhosherstov et al., 2023), the surrogate uses a logarithmic scale; mean MSE instead depends exponentially on distances.

After fitting centers, the same bound calibrates covariance from residuals $s_t - c_{a_t}$, where a_t is the nearest-center index. Replacing squared distances by h_Σ from (11) leaves the covariance objective $\sum_t w_t h_\Sigma(s_t - c_{a_t})$ at fixed centers, assignments and weights. Its closed-form minimizer over $2\Sigma - I_d \succ 0$, Σ_{res}, defines MINT#, a covariance-adapted variant of MINT, with $q(\omega) = \sum_r \pi_r \mathcal{N}(\omega; c_r, \Sigma_{\text{res}})$ (derivation in Appendix C.1). With these locations and shapes fixed, probability allocation must still account for values and random normalization.

4.2 MIXTURE WEIGHTS FROM FINITE-BUDGET OUTPUT RISK

For the components g_r constructed in Section 4.1, write $q_\pi = \sum_r \pi_r g_r$. Holding geometry fixed, we minimize a numerical approximation to $\mathcal{F}_M(q_\pi)$ in (9) over positive weights summing to one. Row weights β_i normalize each calibration window's squared output energy, so fitting targets window-relative squared error.

A fixed set of integration samples, the *pilot*, caches component and target densities and row residuals to approximate $a(t)$, $b(t)$, and $c(t)$ for each candidate π. Trapezoidal quadrature on a $\log t$ grid evaluates (8), and automatic differentiation of this discrete objective supplies weight gradients. Optimization starts from the lowest pilot risk among nine candidate fits, with a positive weight floor. An independent pilot audits the fit (Appendix C.2). Weights fit the iid risk objective; LLM main comparisons evaluate frozen proposals with Gaussian-marginal orthogonal random features (ORF).

Frozen causal evaluation. Geometry and weights are frozen before drawing the bank and building causal states (Algorithm 1, Appendix C). Unit-covariance density evaluation costs $O(RMd)$ once per bank; processing L tokens costs $O(LM(d + d_v))$, with $O(M(d_v + 1))$ prefix storage. Dense geometry preserves linear dependence on L. The budget M counts features across all components.

5 EXPERIMENTS

Synthetic and OLMo diagnostics test coverage and its relation to output error. Qwen and OLMo comparisons evaluate the complete proposals, followed by a controlled weight-fitting experiment that isolates the benefit of finite-budget risk calibration.

5.1 PROTOCOL

The main LLM comparisons use $M \in \{64, 128, 256, 512, 1024\}$ total features and $R = 16$ for both MINT variants. Qwen3-4B-Base (Yang et al., 2025a) and OLMo 3 7B base (Team Olmo et al., 2025; Groeneveld et al., 2024) use eight WikiText-103 calibration windows and eight PG-19 test books (Merity et al., 2017; Rae et al., 2020), with native attention scaling.

Calibration samples four stratified rows per window, jointly covering 32 position bins over the full sequence range. With calibrated centers and covariance fixed, weights minimize the numerical risk objective $\mathcal{F}_M$ at each budget. Paired intervals use a crossed bootstrap over eight books and four test banks, conditional on the fixed heads and calibration. Synthetic tests and Figure 3 use iid draws; both LLM main comparisons use ORF (details in Appendix D).

5.2 FROM COVERAGE TO PROPOSAL CALIBRATION

Figure 2 tests the coverage picture in Figure 1 on independent synthetic rows. Varying values separates normalizer and output effects: two-center MINT improves both in all six settings, while one center can improve the normalizer yet worsen output. Appendices E.2 and E.1 give budget extensions and exact-law checks.

Coverage on LLM rows. Figure 3 tests whether the proposal geometries in Table 1 improve shared-target coverage and output accuracy on OLMo, pairing a target-density projection with a full-dimensional coverage certificate. Let $\overline{C}_{\mathrm{pair}}$ be the analytic component bound on C_{pair} (Appendix E.4) and q_F the FAVOR# proposal. Define

$$\Gamma(q) = \frac{1 + \overline{C}_{\mathrm{pair}}(q)}{1 + \chi^2(Q\|q_F)}, \tag{14}$$

By (2), $\Gamma(q) < 1$ certifies lower shared-target coverage error than FAVOR#; output measurements test whether both errors decrease together.

5.3 LLM PREDICTION AND APPROXIMATION ACCURACY

Qwen next-token prediction. With eight query heads replaced per layer, both MINT variants improve next-token accuracy over all three FAVOR baselines at every budget (Table 2). Accuracy

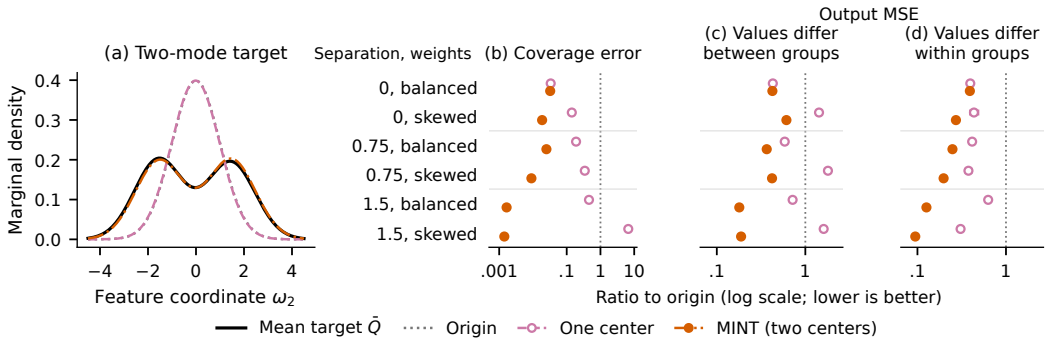

Figure 2: **Two-center calibration improves coverage and output accuracy.** Six two-dimensional synthetic settings vary key-group separation and attention-weight balance. Proposals use unit covariances and frequency weights fitted on 64 rows, with 16 independent test rows and $M = 64$ total features. (a) $\bar{Q}$ averages the test rows' contribution densities for attention's normalizing sum, projected onto ω_2 (separation 1.5, balanced attention weights). MINT follows both modes; origin $\mathcal{N}(0, I)$ and one-center nearly overlap. (b) Coverage error averages $\chi^2(Q_i \| q)$ in both dimensions. (c,d) Output mean-squared error (MSE) with values differing between/within key groups. All ratios use origin as reference; left of the dotted line is better. Error bars: pointwise 95% paired bootstrap intervals over 4,096 banks, conditional on the fixed rows.

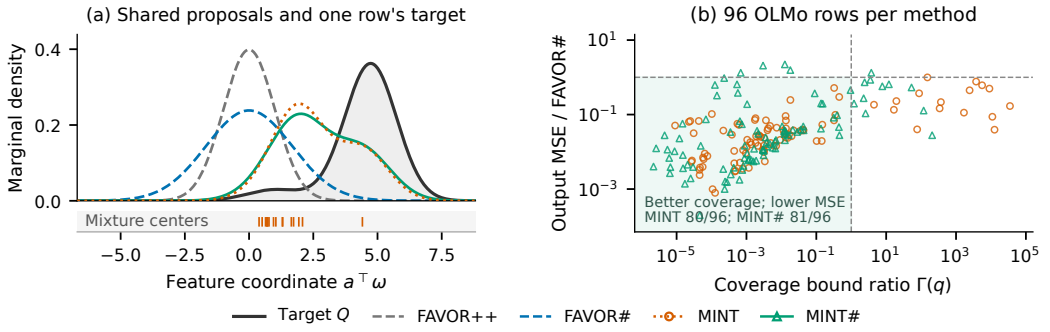

Figure 3: **Shared calibration improves coverage and output accuracy on most tested OLMo rows.** (a) Q is one row's contribution density for attention's normalizer. Shared proposals q and Q have exact marginals along calibration direction a; ticks mark centers. (b) Each point is a test row (6 heads $\times$ 8 books $\times$ 2 positions), with $M = 256$ features per bank. Horizontally, Γ bounds the ratio of $1 + \chi^2(Q\|q)$ to its FAVOR# value. $\Gamma < 1$ certifies lower full-dimensional coverage error; $\Gamma \geq 1$ is inconclusive. Vertically, mean-squared output error is relative to FAVOR#, averaged over 16 iid banks; below one is better. Both improve for 80/96 MINT rows and 81/96 MINT# rows (shaded).

gains over FAVOR# are 4.14–7.44 percentage points for MINT and 2.86–5.37 for MINT#; all five MINT# Bonferroni-adjusted intervals are positive (Table 4). MINT# has 1.28–2.51 points lower mean accuracy than MINT; exact softmax reaches 47.75%.

OLMo approximation probes. Figure 4 traces error from attention outputs to vocabulary distributions during prompt processing and decoding, replacing the same six heads. MINT# lowers mean error against FAVOR# in all fifteen comparisons: 65.7–81.1% for attention, 51.1–73.0% for prompt logits, and 55.0–77.6% for KV decoding. Fourteen Bonferroni-adjusted intervals remain positive (Table 8). Plain MINT has lower attention error at $M = 512$ and 1024.

5.4 What finite-budget risk calibration adds

With centers and covariance fixed, changing the weight objective isolates the benefit of output-risk calibration. Fitting $\mathcal{F}_M$ lowers all six ORF mean errors against denominator-only and shared-target KL fitting (Table 3). Six denominator and four KL Bonferroni-adjusted intervals remain positive.

Table 2: Qwen3-4B-Base top-1 next-token accuracy (%; ↑), replacing eight query heads per layer (25%). M counts features per head. WikiText calibration and PG-19 test windows contain 4096 tokens. Means use eight test books and four ORF banks with the same head mask across methods; exact softmax scores 47.753%.

M	FAVOR+	FAVOR++	FAVOR#	MINT	MINT#
64	10.316	8.661	21.687	28.674	27.053
128	13.694	15.236	22.498	29.933	27.421
256	11.179	13.070	23.170	29.755	28.411
512	9.280	14.093	25.414	30.130	28.403
1024	12.372	17.430	27.045	31.183	29.904

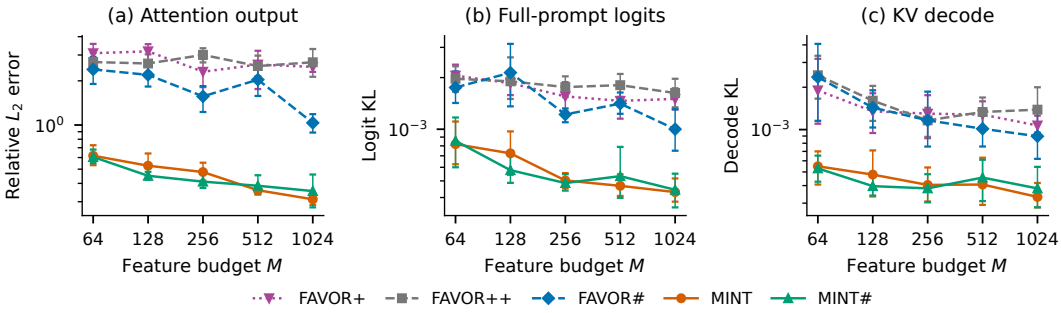

Figure 4: **MINT proposals reduce approximation error on OLMo 3.** Each of six fixed heads uses M ORF features. (a) Attention-output error on 1,024 tokens: Frobenius error norm divided by the exact output norm. (b,c) Mean token-level Kullback–Leibler (KL) divergence between exact and approximate vocabulary distributions, for 4,096-token prompts and 32 teacher-forced key–value (KV) decoding steps after 1,024-token exact prefill. Lower is better. Means and pointwise 95% crossed-bootstrap intervals use eight PG-19 books and four banks (paired improvements: Table 8).

Under iid draws, the fit also lowers all six means against pairwise and J_o-only objectives; per-head validation selection remains competitive (Appendix E.5).

Table 3: **Finite-budget risk calibration improves ORF output accuracy.** Within each family, centers and covariance stay fixed; weights fit the iid finite-budget output risk $\mathcal{F}_M$ at M features per head. Entries give test relative-MSE reduction (%) with pointwise 95% paired intervals: six fixed OLMo heads; 16 PG-19 books and 128 ORF banks resampled.

	MINT		MINT#	
M	vs. denominator	vs. shared KL	vs. denominator	vs. shared KL
64	40.8 [27.9, 52.2]	16.7 [1.8, 30.0]	59.9 [49.2, 69.6]	21.2 [14.7, 27.4]
256	41.9 [33.4, 50.0]	22.4 [14.6, 32.1]	53.2 [42.4, 63.0]	19.3 [9.4, 27.3]
1024	31.7 [20.3, 42.2]	12.3 [1.8, 20.0]	57.7 [49.9, 64.4]	23.5 [15.8, 29.7]

Denominator minimizes value-weighted normalizer error; shared KL fits a weighted mixture of row-contribution densities by KL divergence. Reduction is $100(1 - E_{\text{risk}}/E_{\text{base}})$, where E averages $\|\widehat{O} - O\|_F^2/\|O\|_F^2$ for approximate/exact attention-output tensors $\widehat{O}, O$. Positive values favor risk fitting.

6 Conclusion

Our unified proposal-coverage analysis traces kernel approximation errors through shared normalization and value aggregation to the attention output. It yields an upper bound and an exact expression for finite-budget output risk under independent sampling. Guided by this analysis, MINT and MINT# combine multiple centers with budget-aware weight calibration. Synthetic experiments validate the analysis, while LLM experiments demonstrate improved prediction and approximation accuracy over FAVOR baselines at matched feature budgets.

AI USE STATEMENT

Generative AI assisted with coding, manuscript restructuring and editing, literature cross-checking, mathematical formulation and proof revision, experiment-design feedback, execution of synthetic and LLM calibration experiments, interpretation of results, LaTeX tables and schematic figures, and scripts for reanalyzing and plotting saved experimental data.

REPRODUCIBILITY STATEMENT

Appendices A–C give assumptions, proofs, coordinate conventions, and fitting objectives. Appendix D records the available LLM and synthetic protocols, including geometry fitting and budget-dependent weight calibration. The accompanying supplementary package contains self-contained code and full generated data for the synthetic calibration experiment, plus saved inputs and provenance for the experiment figures and tables, the norm-probe and paired-summary analyses in Appendix E, and numerical checks of the theoretical identities. Implementations and fixed reproduction recipes are available at `https://anonymous.4open.science/r/mint-submission`. Public model checkpoints and datasets are downloaded separately using the provided download and verification scripts. The calibration recipes generate captures, integration pilots, and fitted proposals locally.

REFERENCES

Cathrine Aeckerle-Willems, Ilja Klebanov, and Simon Weissmann. Error bounds for importance sampling with estimated proposal distributions. *arXiv preprint arXiv:2605.19989*, 2026. URL https://arxiv.org/abs/2605.19989.

Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017. doi: 10.1214/17-STS611. URL https://arxiv.org/abs/1511.06196.

Masoud Badiei Khuzani, Sharath Honnaiah, Atiq Islam, Alex Cozzi, and Abraham Bagherjeiran. A Coulomb particle model for learning kernel attention in transformers. In *ICML Workshop on High-dimensional Learning Dynamics (HiLD)*, 2026. URL https://arxiv.org/abs/2607.23869.

Nicola Branchini and Víctor Elvira. Towards adaptive self-normalized importance samplers. In *IEEE Statistical Signal Processing Workshop (SSP)*, 2025. doi: 10.1109/SSP64130.2025.11073204. URL https://arxiv.org/abs/2505.00372v2.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Victor H. de la Peña, Heyuan Yao, and Haolin Zou. A scalable formula for the moments of a family of self-normalized statistics. *arXiv preprint arXiv:2509.14428*, 2025. URL https://arxiv.org/abs/2509.14428.

Amirhossein Farzam, Hossein Mobahi, Nolan Andrew Miller, and Luke Sernau. Data-aware random feature kernel for transformers. In *2nd DeLTa Workshop at ICLR*, 2026. URL https://arxiv.org/abs/2603.04127.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, 2024. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Juan Gabriel Kostelec and Qinghai Guo. FlashEVA: Accelerating LLM inference via efficient attention. *arXiv preprint arXiv:2511.00576*, 2025. URL https://arxiv.org/abs/2511.00576.

Valerii Likhosherstov, Krzysztof Choromanski, Avinava Dubey, Frederick Liu, Tamás Sarlós, and Adrian Weller. Chefs' random tables: Non-trigonometric random features. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34559–34573, 2022. doi: 10.52202/068431-2504. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/df2d62b96a4003203450cf89cd338bb7-Abstract-Conference.html.

Valerii Likhosherstov, Krzysztof Choromanski, Avinava Dubey, Frederick Liu, Tamás Sarlós, and Adrian Weller. Dense-exponential random features: Sharp positive estimators of the Gaussian kernel. In *Advances in Neural Information Processing Systems*, volume 36, pp. 957–985, 2023. doi: 10.52202/075280-0046. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/02dec8877fb7c6aa9a79f81661baca7c-Abstract-Conference.html.

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

Sankalan Pal Chowdhury, Adamos Solomou, Avinava Dubey, and Mrinmaya Sachan. Learning the transformer kernel. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=tLIBAEYjcv.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. Random feature attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QtTKTdVrFBB.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Isaac Reid, Krzysztof Choromanski, Valerii Likhosherstov, and Adrian Weller. Simplex random features. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28864–28888, 2023. URL https://proceedings.mlr.press/v202/reid23a.html.

Luke Sernau, Silvano Bonacina, and Rif A. Saurous. All random features representations are equivalent. *arXiv preprint arXiv:2406.18802*, 2024. URL https://arxiv.org/abs/2406.18802.

Team Olmo, Allyson Ettinger, Amanda Bertsch, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. URL https://arxiv.org/abs/2512.13961.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523, 2024. URL https://proceedings.mlr.press/v235/yang24ab.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4904fad153f6434a7bcf04465d4be2cc-Abstract-Conference.html.

Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré. LoLCATs: On low-rank linearizing of large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/72163d1c3c1726f1c29157d06e9e93c1-Abstract-Conference.html.

Lin Zheng, Chong Wang, and Lingpeng Kong. Linear complexity randomized self-attention mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27011–27041, 2022. URL https://proceedings.mlr.press/v162/zheng22b.html.

Lin Zheng, Jianbo Yuan, Chong Wang, and Lingpeng Kong. Efficient attention via control variates. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=G-uNfHKrj46.

A Proofs and Error Structure

All results condition on a fixed finite row and a fixed proposal, including all calibration-dependent choices. Integrals use Lebesgue measure. Features and normalization are exact unless a numerical implementation is explicitly discussed.

A.1 Pair and shared-row identities

Let $\gamma = \mathcal{N}(0, I_d)$ and $\phi_0(x;\omega) = \exp(\omega^\top x - \|x\|^2/2)$. For a target density p, the support condition $p \ll q$ requires $q > 0$ wherever p has mass, up to null sets. Its coverage mismatch is

$$\chi^2(p\|q) = \int p(\omega)^2/q(\omega)\, d\omega - 1. \tag{15}$$

For $s = x + y$ and $p_s = \mathcal{N}(s, I_d)$, the feature product and importance-corrected estimator satisfy

$$\phi_0(x;\omega)\phi_0(y;\omega) = K(x,y)\frac{p_s(\omega)}{\gamma(\omega)}, \qquad \widehat{K}_{q,M}(x,y) = \frac{K(x,y)}{M}\sum_{m=1}^{M}\frac{p_s(\omega_m)}{q(\omega_m)}. \tag{16}$$

Proposition 2 (Pairwise coverage)**.** *If $p_s \ll q$ and $\chi^2(p_s\|q) < \infty$, then $\widehat{K}_{q,M}$ is unbiased and* $\mathrm{RelMSE}(\widehat{K}_{q,M}) = \chi^2(p_s\|q)/M$.

The likelihood ratio satisfies $\log(p_s/\gamma) = \omega^\top s - \|s\|^2/2$. Since $\|x+y\|^2/2 - x^\top y = (\|x\|^2 + \|y\|^2)/2$, this proves (16). Under q, $Z_s = p_s/q$ has mean one and second moment $1 + \chi^2(p_s\|q)$. Averaging M iid copies proves Proposition 2. Infinite χ^2 gives infinite pairwise MSE even though the nonnegative estimator remains unbiased.

Proof of Theorem 1. For the row quantities in (1), write $\widehat{D}_{q,M} = \sum_j \widehat{K}_{q,M}(x, y_j) = \widehat{D}$. For one shared bank, $\widehat{D}/D = M^{-1}\sum_m Q(\omega_m)/q(\omega_m)$, and the same argument proves the equality in (2). Set $Z_j = p_j/q$ and $Z_Q = Q/q = \sum_j \alpha_j Z_j$. Expanding the squares gives

$$\begin{aligned} C_{\mathrm{pair}}(q) - \chi^2(Q\|q) &= \mathbb{E}_q\left[\sum_j \alpha_j (Z_j - 1)^2 - (Z_Q - 1)^2\right] \\ &= \mathbb{E}_q \sum_j \alpha_j (Z_j - Z_Q)^2 \geq 0. \end{aligned} \tag{17}$$

This proves (3); nonnegativity also yields the inequality in (2). With independent banks of size M for each key, all cross-pair covariances disappear, leaving $M^{-1}\sum_j \alpha_j^2 \chi^2(p_j\|q)$. That comparison consumes LM draws; it is not an equal-total-budget alternative.

A.2 Output identity, limit, and finite-budget bound

For the fixed values v_j, the exact and estimated outputs are

$$o = \sum_j \alpha_j v_j, \qquad \widehat{o} = \frac{\sum_j \widehat{K}_j v_j}{\sum_j \widehat{K}_j}. \tag{18}$$

Define the value-weighted integrand H, centered residual G, and its second-moment coefficient J_o by

$$H = \sum_j \alpha_j v_j p_j, \qquad G = H - oQ, \qquad J_o(q) = \int \frac{\|G(\omega)\|^2}{q(\omega)}\, d\omega. \tag{19}$$

Rescaling the estimated numerator and denominator by D gives

$$\widehat{n} = \frac{1}{M}\sum_m \frac{H(\omega_m)}{q(\omega_m)}, \quad \widehat{d} = \frac{\widehat{D}_{q,M}}{D} = \frac{1}{M}\sum_m \frac{Q(\omega_m)}{q(\omega_m)}, \quad \widetilde{e} = \widehat{n} - o\widehat{d}, \quad \widehat{o} - o = \frac{\widetilde{e}}{\widehat{d}}. \tag{20}$$

Here $\rho_v = \max_j \|v_j - o\|$ bounds the deviation from the exact output.

Proposition 3 (Output error before and after normalization)**.** *Suppose $Q/q > 0$ q-almost surely and $\mathbb{E}_q[(Q/q)^2 + \|G/q\|^2] < \infty$. Then*

$$\mathbb{E}\|\widetilde{e}\|^2 = \frac{J_o(q)}{M}. \tag{21}$$

If additionally $\mathbb{E}_q[(Q/q)^4 + \|G/q\|^4] < \infty$, then $MR_M(q) \to J_o(q)$ for fixed q.

Write $Z_Q = Q/q$ and $Z_G = G/q$. Their means are one and zero, respectively. Since $\widetilde{e} = M^{-1}\sum_m Z_G(\omega_m)$, independence and centering give

$$\mathbb{E}[\widetilde{e}\widetilde{e}^\top] = \frac{1}{M}\int \frac{GG^\top}{q}, \qquad \mathbb{E}\|\widetilde{e}\|^2 = J_o/M. \tag{22}$$

The law of large numbers gives $\widehat{d} \to 1$ in probability. The multivariate central limit theorem for Z_G and Slutsky's theorem applied to $\widetilde{e}/\widehat{d}$ give

$$\sqrt{M}(\widehat{o} - o) \Rightarrow \mathcal{N}(0, \Sigma_o), \qquad \Sigma_o = \int GG^\top/q, \qquad \operatorname{tr}\Sigma_o = J_o. \tag{23}$$

Here $\Rightarrow$ denotes convergence in distribution and tr the matrix trace; singular covariance cases are included.

Why centering is essential. Because $\sum_j \alpha_j(v_j - o) = 0$,

$$\frac{G}{q} = \sum_j \alpha_j(v_j - o)\left(\frac{p_j}{q} - 1\right). \tag{24}$$

Jensen's inequality for the squared Euclidean norm, followed by expectation under q, gives

$$\begin{aligned} J_o &\le \sum_j \alpha_j \|v_j - o\|^2 \mathbb{E}_q\left[\left(\frac{p_j}{q} - 1\right)^2\right] = \sum_j \alpha_j \|v_j - o\|^2 \chi^2(p_j\|q) \\ &\le \rho_v^2 C_{\mathrm{pair}}(q). \end{aligned} \tag{25}$$

Without the subtraction of one, the direct bound would involve $1 + \chi^2$ rather than χ^2.

Proof of Proposition 1: value dispersion. Suppose $C_{\mathrm{pair}}(q) < \infty$. As in Theorem 1, set $Z_j = p_j/q$ and $Z_Q = Q/q = \sum_j \alpha_j Z_j$. Let $\sigma_v^2 = \sum_j \alpha_j \|v_j - o\|^2 \le \rho_v^2$ be the attention-weighted value dispersion. Because $\sum_j \alpha_j(v_j - o) = 0$, centering at Z_Q instead of one yields

$$\frac{G}{q} = \sum_j \alpha_j(v_j - o)(Z_j - Z_Q), \qquad \|G/q\|^2 \le \sigma_v^2 \sum_j \alpha_j (Z_j - Z_Q)^2. \tag{26}$$

The inequality is weighted Cauchy–Schwarz. Integrating and using (17) proves (6). This supplements the value-weighted bound (25): neither needs to dominate the other for every row.

Proof of Proposition 1: finite-budget normalization. Define the estimated attention weights $\widehat{\alpha}_j = \widehat{K}_j / \sum_k \widehat{K}_k$. Positive features imply $\widehat{o} = \sum_j \widehat{\alpha}_j v_j$, $\widehat{\alpha}_j \ge 0$, and $\sum_j \widehat{\alpha}_j = 1$. Therefore

$$\|\widehat{o} - o\| \le \sum_j \widehat{\alpha}_j \|v_j - o\| \le \rho_v. \tag{27}$$

Let $e = \widehat{o} - o$. From $\widetilde{e} = \widehat{d}e$ we have the samplewise identity

$$e = \widetilde{e} + (1 - \widehat{d})e. \tag{28}$$

For a random vector X, write $\|X\|_{L^2} = (\mathbb{E}\|X\|^2)^{1/2}$. The triangle inequality and (27) give

$$\begin{aligned} \|e\|_{L^2} &\le \|\widetilde{e}\|_{L^2} + \|(1 - \widehat{d})e\|_{L^2} \\ &\le \sqrt{\frac{J_o(q)}{M}} + \rho_v \sqrt{\frac{\chi^2(Q\|q)}{M}}. \end{aligned} \tag{29}$$

Here $\mathbb{E}\|\widetilde{e}\|^2 = J_o(q)/M$ and $\mathbb{E}(1-\widehat{d})^2 = \chi^2(Q\|q)/M$ follow from the centered iid sample means. Squaring proves (7). Bounding the product pointwise avoids any independence assumption between e and $\widehat{d}$. Positivity ensures $\widehat{d} > 0$ almost surely, but no deterministic positive lower bound or inverse moment is needed. The decomposition parallels the bounded-estimator argument for self-normalized importance sampling in Agapiou et al. (2017, proof of Theorem 2.1); here the two terms retain the distinct output and shared-denominator coefficients.

For the pairwise certificate, set $B = \chi^2(Q\|q)$ and $C = C_{\mathrm{pair}}(q) < \infty$. Equation (6) gives $J_o(q) \le \sigma_v^2(C-B)$, with $0 \le B \le C$. A two-dimensional Cauchy–Schwarz inequality gives

$$\begin{aligned}\left(\sqrt{J_o(q)} + \rho_v\sqrt{B}\right)^2 &\le \left(\sigma_v\sqrt{C-B} + \rho_v\sqrt{B}\right)^2 \\ &\le (\sigma_v^2 + \rho_v^2)C \le 2\rho_v^2 C. \end{aligned} \tag{30}$$

Combining this with (7) yields the pairwise certificate

$$R_M(q) \le \frac{(\sigma_v^2+\rho_v^2)C_{\mathrm{pair}}(q)}{M} \le \frac{2\rho_v^2 C_{\mathrm{pair}}(q)}{M}. \tag{31}$$

Both finite-budget conclusions use only second moments. The fourth-moment condition below is a separate sufficient condition for convergence of the scaled MSE to $J_o(q)$; it is not required above.

Convergence of the mean squared error. Let $a = \widehat{d} - 1$, $b = \widetilde{e}$, and $e = \widehat{o} - o$. The fourth-moment assumption implies $\mathbb{E}a^4 = O(M^{-2})$ and $\mathbb{E}\|b\|^4 = O(M^{-2})$ for the centered iid sample means. On the event $A = \{|a| \le 1/2\}$, with indicator $\mathbf{1}_A$, $e - b = -ab/(1+a)$, so Cauchy–Schwarz gives

$$\mathbb{E}[\|e-b\|^2\mathbf{1}_A] \le 4(\mathbb{E}a^4)^{1/2}(\mathbb{E}\|b\|^4)^{1/2} = O(M^{-2}). \tag{32}$$

On A^c, $\Pr(A^c) \le 16\mathbb{E}a^4 = O(M^{-2})$ and

$$\mathbb{E}[\|e-b\|^2\mathbf{1}_{A^c}] \le 2\rho_v^2\Pr(A^c) + 2(\mathbb{E}\|b\|^4)^{1/2}\Pr(A^c)^{1/2} = O(M^{-2}). \tag{33}$$

Thus $\mathbb{E}\|e-b\|^2 = O(M^{-2})$. The difference between $\mathbb{E}\|e\|^2$ and $\mathbb{E}\|b\|^2$ is $O(M^{-3/2})$ by one more Cauchy–Schwarz inequality. Multiplication by M proves $M\mathbb{E}\|e\|^2 \to J_o$.

A.3 Exact finite-budget risk and its large-budget limit

We prove the integral representation in Section 3.3, using the single-draw contributions W, Z and Laplace expectations a, b, c defined there.

Proposition 4 (Finite-budget output risk)**.** *For a fixed finite attention row and M iid draws with $W > 0$ almost surely, the output risk satisfies* (8)*, with the second term omitted for $M = 1$.*

To prove Proposition 4, write $S_W = \sum_{m=1}^M W_m$ and $S_Z = \sum_{m=1}^M Z_m$. Positive features give $W_m > 0$, $\widehat{o} - o = S_Z/S_W$, and $\|Z_m\| \le \rho_v W_m$. The inverse-power identity $S_W^{-2} = \int_0^\infty te^{-tS_W}\,dt$ therefore yields

$$R_M(q) = \int_0^\infty t\,\mathbb{E}\left[\|S_Z\|^2 e^{-tS_W}\right]dt. \tag{34}$$

The equality follows by Tonelli's theorem. To separate the diagonal and cross terms, note that $\sum_{m,n}|\langle Z_m, Z_n\rangle|/S_W^2 \le \rho_v^2$; thus each term is absolutely integrable. Independence factors each diagonal term as $c(t)a(t)^{M-1}$ and each distinct-index term as $\|b(t)\|^2 a(t)^{M-2}$, giving (8). Exponential damping makes these expectations finite for every $t > 0$, even when the undamped second moments are infinite. For $M = 1$ the cross term is absent. This derivation applies the established inverse-power technique for self-normalized statistics (de la Peña et al., 2025); the design target here is a shared attention proposal at a specified feature budget.

Under Proposition 3's fourth-moment conditions, $MR_{M,i}(q) \to J_{o,i}(q)$ for each row. Summing with weights β_i in (9) gives the large-budget limit in Section 3.3. This limit concerns the population objective; with a fixed finite integration pilot, discrete approximation bias can persist as M grows.

A.4 GAUSSIAN INTEGRAL AND MOMENT CONDITIONS

For a Gaussian proposal $q = \mathcal{N}(c, \Sigma)$ with $2\Sigma - I_d \succ 0$, the coverage coefficient is

$$1 + \chi^2(p_s \| q) = \frac{\det \Sigma}{\sqrt{\det(2\Sigma - I_d)}} \exp\big((s-c)^\top (2\Sigma - I_d)^{-1}(s-c)\big). \tag{35}$$

In particular, $\Sigma = I_d$ gives (10). To derive the general expression, set $u = \omega - c$ and $h = s - c$. Then

$$\int \frac{p_s^2}{q} = (2\pi)^{-d/2} (\det \Sigma)^{1/2} \int \exp\left[-\tfrac{1}{2} u^\top (2I - \Sigma^{-1}) u + 2h^\top u - \|h\|^2\right] du. \tag{36}$$

The quadratic integral is finite when $2\Sigma - I \succ 0$. Completing the square and using $2(2I - \Sigma^{-1})^{-1} - I = (2\Sigma - I)^{-1}$ proves (35). If $2\Sigma - I$ is not positive definite, the integral diverges. For a mixture, a single positive-weight component with suitable tails suffices for a finite moment, since $q \geq \pi_r q_r$ for its component density q_r.

For moment order $k > 1$, $\int p_s^k / q^{k-1}$ is finite for a Gaussian component when $k\Sigma - (k-1)I \succ 0$. Hence $4\Sigma - 3I \succ 0$ for one positive-weight component is a convenient sufficient fourth-moment condition. Unit-covariance location mixtures satisfy it for every fixed finite set of target centers. The finite-row mixtures Q and G inherit these moments by convexity and finite sums.

A.5 TWO ANALYTICAL EXAMPLES SEPARATING THE ERROR OBJECTS

Example 1 (Poor coverage with exact output). For any row, set $v_j = v_\star$ for every key. Then $o = v_\star$, $G = 0$, and $\widehat{o} = v_\star$ for every positive-feature bank. To make coverage arbitrarily poor, take scalar $x = 0$, $y_j = a$ and $q = \mathcal{N}(0, 1)$. Both the pair and shared-denominator coefficients are $e^{a^2} - 1$, while output error is identically zero.

Example 2 (Exact denominator with noisy output). Take scalar $x = 0$, $y_1 = a$, $y_2 = -a$ with $a \neq 0$, and $v_1 = 1$, $v_2 = -1$. Both exact kernels are one, $\alpha_1 = \alpha_2 = 1/2$, and $o = 0$. Let $q = Q = \frac{1}{2}\mathcal{N}(a, 1) + \frac{1}{2}\mathcal{N}(-a, 1)$. Then $\widehat{d} = 1$ for every bank, but

$$\frac{G}{Q} = \frac{p_a - p_{-a}}{p_a + p_{-a}} = \tanh(a\omega), \qquad \mathbb{E}|\widehat{o} - o|^2 = \frac{\mathbb{E}_Q[\tanh^2(a\omega)]}{M} > 0. \tag{37}$$

Here $\sigma_v^2 = \rho_v^2 = 1$ and $C_{\mathrm{pair}}(Q) = J_o(Q)$, so both (6) and (7) hold with equality. The row-specific oracle $q = Q$ serves to separate the coefficients; it is not the frozen proposal used by MINT. Because $o = 0$, this is an absolute, not relative, output-error example.

B GAUSSIAN PROPOSAL REPRESENTATIONS

A change of coordinates expresses frozen symmetric DERF features (Likhosherstov et al., 2023) as Gaussian importance proposals. This makes the location and covariance choices in Table 1 directly comparable.

Let $u \sim \gamma = \mathcal{N}(0, I_d)$, $A = A^\top$, $P = I - 4A \succ 0$, and let an invertible B satisfy $B^\top P^{-1} B = I$. Write $\theta = (A, B)$ for the frozen geometry parameters. With $D_\theta = (\det P)^{1/4}$, write the symmetric feature as

$$\psi_\theta(z; u) = D_\theta \exp\left(u^\top A u + u^\top B z - \tfrac{1}{2}\|z\|^2\right). \tag{38}$$

The Gaussian moment identity yields $\mathbb{E}_\gamma[\psi_\theta(x; u)\psi_\theta(y; u)] = e^{x^\top y}$. The normalized pair target in these *natural coordinates* is

$$p^u_{x,y}(u) = \frac{\gamma(u)\psi_\theta(x; u)\psi_\theta(y; u)}{e^{x^\top y}} = \mathcal{N}(u; P^{-1}Bs, P^{-1}), \qquad s = x + y. \tag{39}$$

Evaluator coordinates. Set $w = B^\top u$ and $\Sigma = B^\top B$. The base draw law becomes $r_w = \mathcal{N}(0, \Sigma)$, while (39) becomes $p_s = \mathcal{N}(s, I)$ because $B^\top P^{-1} B = I$. In fact, at feature level,

$$\psi_\theta(z; B^{-\top} w) = \sqrt{\frac{\gamma(w)}{r_w(w)}} \exp\left(w^\top z - \tfrac{1}{2}\|z\|^2\right). \tag{40}$$

To verify the prefactor, use $P = BB^\top$, the Jacobian $du = |\det B|^{-1}dw$, and $r_w(w)D_\theta^2 \exp(2u^\top Au) = \gamma(w)$. Thus the FAVOR# (SDERF) row of Table 1 is an exact Gaussian-proposal representation for this frozen feature. Its finite pairwise variance condition is $2\Sigma - I \succ 0$, equivalently $I - 8A \succ 0$. A sufficient fourth-moment condition is $4\Sigma - 3I \succ 0$.

Adding location in the natural coordinates. For $q_u(u) = \sum_r \pi_r \mathcal{N}(u; c_r^u, I)$, draw $u \sim q_u$ and use $\sqrt{\gamma(u)/q_u(u)}\psi_\theta(z;u)$. Its pushforward law is

$$q_w(w) = \sum_r \pi_r \mathcal{N}(w; B^\top c_r^u, \Sigma), \qquad q_w(w) = \frac{q_u(B^{-\top}w)}{|\det B|}. \tag{41}$$

Since $\gamma(u)/q_u(u) = r_w(w)/q_w(w)$, the fully corrected feature becomes $\sqrt{\gamma(w)/q_w(w)}\phi_0(z;w)$. The common-covariance bound (42) therefore applies with centers $c_r^w = B^\top c_r^u$. A calibration target mean in natural coordinates is $P^{-1}Bs = B^{-\top}s$; its covariance is P^{-1}, not generally I. The fitted unit-covariance components can therefore differ from the natural targets.

Coverage expressions and comparisons. The coordinate transformation applies jointly to centers, covariances, draws, and full mixture densities. Table 1 compares these proposals using $\ell_q(s) = \log(1+\chi^2(p_s\|q))$ for $s = x+y$ and $p_s = \mathcal{N}(s, I_d)$. At M iid features, pairwise RelMSE is $(e^{\ell_q(s)} - 1)/M$. For $2\Sigma - I_d \succ 0$, define $B_\Sigma = (2\Sigma - I_d)^{-1}$, $a_\Sigma = \log\det\Sigma - \frac{1}{2}\log\det(2\Sigma - I_d)$, and $\|h\|_{B_\Sigma}^2 = h^\top B_\Sigma h$. The shorthand in Table 1 is $h_\Sigma(z) = a_\Sigma + \|z\|_{B_\Sigma}^2$. For $\Sigma = \sigma^2 I_d$, the finite second-moment condition is $\sigma^2 > 1/2$.

Equation (35) gives exact values for single Gaussians, including $\ell_{\mathcal{N}(c,I_d)}(s) = \|s-c\|^2$ for MINT-1. For a mixture $q = \sum_{r=1}^R \pi_r g_r$ with common-covariance components $g_r = \mathcal{N}(c_r, \Sigma)$, positive weights π_r summing to one, and $2\Sigma - I_d \succ 0$, component domination $q \geq \pi_r g_r$ gives

$$\ell_q(s) \leq \min_{r \leq R}\{h_\Sigma(s - c_r) - \log \pi_r\}. \tag{42}$$

Taking $\Sigma = I_d$ or $\Sigma = \Sigma_{\text{res}}$ yields the MINT or MINT# bound in Table 1. A mixture upper bound below an exact baseline value certifies better pairwise coverage at the same target and budget; a larger bound is inconclusive. For fixed Σ, the bound uses the B_Σ metric in evaluator coordinates. Our implementation fits centers using Euclidean k-means and then fits covariance to their residuals (Appendix C).

In Table 1, σ^2 is scalar variance, Σ is the full covariance fitted from global inputs for FAVOR#, and Σ_{res} is the covariance fitted around mixture centers for MINT#. Mixture indices run over $r = 1, \ldots, R$, with centers c_r and positive weights π_r summing to one. The weights are calibrated using $\mathcal{F}_M$; full importance correction preserves the softmax kernel.

B.1 One-center translation equivalence

For $q_c = \mathcal{N}(c, I)$ and $\omega = c + \epsilon$ with $\epsilon \sim \gamma$, the corrected pair product is

$$\phi_{q_c}(x;\omega)\phi_{q_c}(y;\omega) = \exp\left[c^\top(x+y) + \epsilon^\top(x+y-c) - \tfrac{1}{2}\|c\|^2 - \tfrac{1}{2}\|x\|^2 - \tfrac{1}{2}\|y\|^2\right] \tag{43}$$

$$= e^{c^\top x}\phi_0(x;\epsilon)\phi_0(y-c;\epsilon). \tag{44}$$

Consequently, for the same base draws, $\widehat{K}_{q_c}(x,y) = e^{c^\top x}\widehat{K}_\gamma(x, y-c)$. The query-only factor cancels in normalized attention, exactly as it does for exact softmax. This gives a useful samplewise implementation check and a centered-FAVOR control for MINT-1. A full multi-center mixture correction generally does not reduce to one common key translation.

C MINT Calibration: Geometry and Output Risk

The mixture coverage bound yields a clustering objective for proposal location. For calibration centers $s_t = x_t + y_t$ and nonnegative weights w_t summing to one, define

$$U(c,\pi) = \sum_t w_t \min_r\{\|s_t - c_r\|^2 - \log\pi_r\}, \qquad \sum_t w_t \ell_{q_R}(s_t) \leq U(c,\pi). \tag{45}$$

For uniform $\pi_r = 1/R$, this is weighted k-means plus $\log R$. At a fixed assignment, each nonempty cluster's minimizing center is its weighted mean. With uniform π, nearest-center reassignment and mean updates give Lloyd descent; with nonuniform π, reassignment also includes $-\log \pi_r$. For one center, mean log coverage averages squared distances, whereas mean MSE depends on their exponentials, giving distant targets greater influence. For a fixed common covariance Σ, replace squared Euclidean distances by the B_Σ metric and add a_Σ; weighted centroids remain arithmetic means.

Proof of Lemma 1. $q_R \geq \pi_r \mathcal{N}(c_r, I_d)$ implies $\int p_s^2/q_R \leq \pi_r^{-1} e^{\|s-c_r\|^2}$ for each r. Take logarithms and minimize. □

Corollary 1 (When location improves the pairwise law)**.** MINT-*1 improves on the origin proposal exactly when* $\|s-c\|^2 < \|s\|^2$. *For* MINT-*R, a sufficient condition is* $\min_r\{\|s-c_r\|^2 - \log \pi_r\} < \|s\|^2$.

The mixture expression bounds log coverage error; coincident centers recover a single Gaussian for any R. Combining the bound with (25) gives

$$J_o(q_R) \leq \sum_j \alpha_j \|v_j - o\|^2 \left[\min_r \frac{e^{\|x+y_j-c_r\|^2}}{\pi_r} - 1\right]. \tag{46}$$

Together with (7), this bounds the expected squared output error through location residuals and value deviations.

The synthetic experiment clusters all calibration pair centers with attention weights. The LLM fits give equal weight to each sampled pair: both models sample a query row uniformly, then a visible key uniformly, use Euclidean k-means, and set smoothed cluster-frequency weights. Appendix D specifies these protocols. Both MINT variants share their initial centers and frequency weights; MINT# then fits residual covariance. These location fits use the log-coverage bound. With geometry fixed, weights are calibrated to the numerical approximation of the shared finite-budget risk $\mathcal{F}_M$ in (9). All error laws condition on the fitted proposal; controlling proposal-estimation error requires additional assumptions (Aeckerle-Willems et al., 2026). Algorithm 1 specifies how the fitted distribution and total budget M are reused during causal evaluation.

For an input vector $z \in \mathbb{R}^d$, set $\Phi_q(z) = M^{-1/2}(\phi_q(z;\omega_1), \ldots, \phi_q(z;\omega_M))^\top \in \mathbb{R}^M$. Within one sequence, let i denote the current query/prefix position. The key–value state $S_i \in \mathbb{R}^{M\times d_v}$ and key-feature sum $z_i \in \mathbb{R}^M$ give

$$S_i = \sum_{j\leq i} \Phi_q(y_j) v_j^\top, \qquad z_i = \sum_{j\leq i} \Phi_q(y_j), \qquad \widehat{o}_i = \frac{S_i^\top \Phi_q(x_i)}{\Phi_q(x_i)^\top z_i}. \tag{47}$$

Algorithm 1 Offline calibration and fixed-bank causal evaluation

Require: Separate calibration and evaluation inputs; reuse groups; R centers; total budget M.

1: Form calibration target centers $s_t = x_t + y_t$ and fit centers $c_{1:R}$ and frequency weights $\pi^{(0)}$.
2: Set $\Sigma = I_d$ for MINT; for MINT#, fit $\Sigma = \Sigma_{\text{res}}$ using (49).
3: Hold centers and covariance fixed and fit π at budget M to the calibration-pilot approximation of $\mathcal{F}_M$, subject to $\pi_r \geq .01\pi_r^{(0)}$ and $\sum_r \pi_r = 1$.
4: Freeze geometry, centers, weights, and fitting choices for each reuse group.
5: Draw M total feature vectors using the prescribed iid or Gaussian-marginal ORF law; cache full-density corrections.
6: Initialize $S = 0$, $z = 0$; update them for each new key/value via (47).
7: Query the same states with $\Phi_q(x_i)$; evaluate only on disjoint inputs.

C.1 SHARED COVARIANCE FROM CALIBRATION RESIDUALS

For fixed assignments $a_t \in \{1, \dots, R\}$, let $\delta_t = s_t - c_{a_t}$ and $S_\delta = \sum_t w_t \delta_t \delta_t^\top$, with $w_t \geq 0$ and $\sum_t w_t = 1$. Component domination and the Gaussian second moment give

$$\begin{aligned} \sum_t w_t \ell_q(s_t) \leq\ & \log\det\Sigma - \tfrac{1}{2}\log\det(2\Sigma - I_d) \\ & + \operatorname{tr}((2\Sigma - I_d)^{-1} S_\delta) - \sum_t w_t \log \pi_{a_t}, \quad 2\Sigma - I_d \succ 0. \end{aligned} \tag{48}$$

Write $S_\delta = V \operatorname{diag}(\lambda_1, \dots, \lambda_d) V^\top$, where V is orthogonal and $\lambda_k \geq 0$. The covariance minimizing this bound at fixed centers, weights and assignments is

$$\Sigma_{\mathrm{res}} = V \operatorname{diag}(t(\lambda_1), \dots, t(\lambda_d)) V^\top, \qquad t(\lambda) = \frac{3 + 2\lambda + \sqrt{4\lambda^2 + 12\lambda + 1}}{4}. \tag{49}$$

To verify the optimum, set $H = (2\Sigma - I_d)^{-1} \succ 0$. The covariance-dependent objective is $\log\det(I_d + H) - \frac{1}{2}\log\det H + \operatorname{tr}(HS_\delta) - d\log 2$. Its stationarity equation implies that H commutes with S_δ. In each eigendirection, minimizing $\log t - \frac{1}{2}\log(2t-1) + \lambda/(2t-1)$ for $t > 1/2$ gives the unique admissible stationary root in (49). The objective diverges at the domain boundaries, establishing the minimum. In particular, $t(0) = 1$ and $t(\lambda) \geq 1$, so $\Sigma_{\mathrm{res}} \succeq I_d$ satisfies the Gaussian second- and fourth-moment conditions in Appendix A.4.

The implementation holds Euclidean nearest-center assignments fixed during covariance fitting and uses uniform weights over sampled calibration pairs. The log-second-moment spectral optimization is of the same form as symmetric DERF optimization (Likhosherstov et al., 2023); here its input is the residual second moment after mixture-center calibration. The covariance step minimizes the assigned-component log-coverage bound in (48). With centers and covariance fixed, MINT# then calibrates weights to the finite-budget risk $\mathcal{F}_M$. For the eigendecomposition $\Sigma_{\mathrm{res}} = V \operatorname{diag}(t_k) V^\top$, the exact symmetric feature representation uses $B = \operatorname{diag}(\sqrt{t_k}) V^\top$ and $A = \operatorname{diag}((1 - t_k)/4)$ in (38).

C.2 NUMERICAL FINITE-BUDGET RISK CALIBRATION

Let g_r denote the fixed Gaussian components and $r(\omega) = R^{-1} \sum_r g_r(\omega)$ the pilot density. We draw 65,536 points with equal quotas from the components, independently of deployment banks. For each point ω_s we cache $g_r(\omega_s)/r(\omega_s)$, $\log(Q_i(\omega_s)/r(\omega_s))$, and $\delta_i(\omega_s) = G_i(\omega_s)/Q_i(\omega_s)$ for each calibration row. Candidate mixture weights give discrete probabilities proportional to $q_\pi(\omega_s)/r(\omega_s)$; we normalize their empirical sum and separately record its deviation from one. Equation (8) then gives the exact risk for this discrete law up to quadrature, yielding a generally biased finite-pilot approximation to the continuous-proposal risk.

Set $x = \log t$ and $u = tW$. The integrand with respect to x becomes

$$M\mathbb{E}_q[\|\delta\|^2 u^2 e^{-u}] a^{M-1} + M(M-1)\|\mathbb{E}_q[\delta u e^{-u}]\|^2 a^{M-2}. \tag{50}$$

The functions ue^{-u} and $u^2 e^{-u}$ are bounded, avoiding explicit large undamped moments. Densities and quadrature use float64, with residual vectors stored in float32. We use trapezoidal integration with log-step 0.5 and audit step 0.25. Gradients include both the discrete sampling probabilities and importance weights as functions of π.

For each head and budget, SLSQP performs local optimization of this generally nonconvex objective, using at most 80 iterations, tolerance 10^{-6}, and lower bounds $.01\pi_r^{(0)}$. Initialization is the lowest training-pilot risk among nine fixed candidates: frequency, uniform, pairwise, denominator, J_o, shared-target KL, two denominator-constrained J_o fits, and one denominator-constrained pairwise fit. The constraint is $\mathcal{D}(q) \leq \min\{\mathcal{D}(q_{\pi^{(0)}}), (1+s)\mathcal{D}(q_{\mathrm{den}})\}$, where $\mathcal{D}$ is defined in Appendix E.5 and q_{den} is its fitted minimizer. The two output candidates use $s = .01$ and 1; the pairwise candidate uses .01. Failed or infeasible solves, or those increasing the training objective, retain their initialization. An independent 65,536-point pilot audits fitted weights without selecting them.

Both LLM studies use four stratified rows per calibration window, one in each quarter of the local position range. Across eight windows, the 32 rows occupy distinct equal-width bins: 128 tokens

for Qwen and 32 for OLMo. One position is drawn within each bin, and the eight bins in each quarter are randomly assigned to the windows. Each head fixes this draw across all budgets, both variants, and weight objectives. Let $\mathcal{I}_d$ denote the four selected positions in window d, and N the number of calibration windows. We set $\beta_i = [N \sum_{j \in \mathcal{I}_d} \|o_j\|^2]^{-1}$ for $i \in \mathcal{I}_d$, so the objective averages window-relative squared error on those positions. Main test banks use Gaussian-marginal ORF, independently of the integration pilots, with the numerical settings specified below. The direct-selection controls use all validation rows (Appendix E.5).

D Experimental Setup

The protocols below specify the comparison scope, calibration/evaluation splits, and sampling rules. Supplementary `README.md` gives the commands, input files, field mappings, and provenance needed to reproduce the analyses.

D.1 Comparison scope

Table 1 compares frozen positive-feature proposals for the original softmax kernel. Table 2 and Figure 4 report all three FAVOR baselines and both MINT variants on Qwen and OLMo, respectively; FAVOR# uses full-matrix SDERF. The comparisons match total feature count M; computation and memory costs remain method-dependent. ORF changes draw dependence while permitting fixed banks; the iid error identities require independent draws. RA uses query-specific key features and quadratic sequence cost; the original LARA construction does not support causal attention (Zheng et al., 2022, Section 3.2 and Appendix H.4). Table 1's "No" entries concern those original constructions. EVA's causal variant combines local exact scores with grouped approximations (Zheng et al., 2023, Appendix D), with additional budget parameters for the local-window size and group count.

D.2 LLM protocols

Data and geometry. Qwen3-4B-Base uses eight length-4,096 WikiText-103 calibration windows starting at token offsets $0, 1024, \ldots, 7168$; adjacent windows overlap by 3,072 tokens. OLMo 3 7B base uses eight length-1,024 chunks. Both retain native attention scaling, and each window starts at local position zero. Calibration samples 8,192 query–key pairs per head, with equal quotas across chunks: a query row is uniform, followed by a uniform visible key. Euclidean k-means uses random-point initialization and at most 25 iterations; an empty cluster retains its previous center. Adding 10^{-3} to each cluster count and normalizing gives the initial weights. Residual covariance uses Euclidean nearest-center assignments held fixed (Appendix C.1). FAVOR++ selects a zero-centered isotropic covariance τI by minimizing the calibration mean log second moment over $\tau \in \{.75, 1, 1.25, 1.5, 2, 3, 4, 6, 8, 12, 16\}$; FAVOR# fits full-matrix SDERF using all queries and keys from the same eight chunks.

Budget-dependent weight fitting. With centers and covariance fixed, each head's weights are fitted separately at $M \in \{64, 128, 256, 512, 1024\}$ to $\mathcal{F}_M$ on 32 stratified calibration rows. The position bins have width 128 for Qwen and 32 for OLMo; each window contributes one row per quarter. The fitted proposals freeze before test banks are drawn.

Replacement scope. Qwen replaces eight of 32 query heads in each of 36 layers, totaling 288/1,152 heads. A seeded mask (917202601) selects one query head from each group sharing a key–value head; this mask is common to all methods and budgets. The accompanying protocol records every selected index. OLMo uses six fixed zero-based layer:head indices: 7:18, 9:9, 13:16, 21:3, 23:15, and 25:12. All other heads retain exact attention; model weights are unchanged.

Test probes and uncertainty. Evaluation uses PG-19 test books 88–92 and 94–96, with seeds 9131701, 9132701, 9133701, and 9134701. Qwen evaluates 4,095 next-token predictions per 4,096-token book window, with all selected heads replaced jointly. Accuracy scores the observed tokens, whereas prediction agreement compares approximate and exact top-ranked tokens. Cross-entropy averages negative log probabilities over scored tokens; pooled PPL is its exponential. OLMo attention error uses all 1,024 rows; full-prompt logits use 4,096 tokens; KV decoding uses exact prefill

Table 4: Qwen3-4B-Base at 25% head replacement: pooled perplexity (PPL; lower is better) and MINT# minus FAVOR# accuracy in percentage points (pp). M counts features per head. Evaluation uses eight PG-19 books and four ORF banks with 4096-token windows. Intervals use 10000 paired crossed book–bank bootstrap resamples: pointwise 95% and Bonferroni-adjusted across the five accuracy comparisons. Exact-softmax PPL is 12.225. Intervals condition on the fixed head mask and calibrated proposals.

	PPL ↓			Accuracy gain (pp) ↑		
M	FAVOR#	MINT	MINT#	Mean	Pointwise interval	Adjusted interval
64	142.041	61.433	72.406	5.366	$[3.528, 7.144]$	$[2.907, 7.595]$
128	128.537	54.178	69.025	4.924	$[3.531, 6.464]$	$[3.105, 7.091]$
256	113.880	53.774	62.214	5.241	$[4.112, 6.265]$	$[3.762, 6.589]$
512	88.074	51.698	62.234	2.989	$[1.177, 4.371]$	$[0.294, 4.669]$
1024	70.659	46.732	51.597	2.859	$[1.606, 4.183]$	$[1.361, 4.560]$

PPL is the exponential of pooled mean negative log probability of observed next tokens. Accuracy counts top-ranked predictions matching those tokens; the gain and both intervals compare MINT# with FAVOR#.

of 1,024 tokens followed by 32 teacher-forced steps. Its six heads are replaced jointly in the latter two probes, with proposals fixed across probe types. All five methods use Gaussian-marginal ORF: Haar orthogonal directions have independent chi-distributed radii, with independently drawn mixture labels (`ortho_scaling=0`). Each bank remains fixed throughout its sequence. Qwen captures native normalized, rotary-encoded query/key tensors. Its offline and native evaluators share the same feature construction; rows with very small denominators in the floating-point feature gauge are recomputed from the same log features using FP64 log-sum-exp.

Reported means aggregate eight books and four banks. OLMo attention cells first average the six heads. Mean and paired improvement intervals use 10,000 crossed bootstrap resamples of books and banks (seed 93132026). Qwen comparisons pair methods within a budget; budget-dependent seeds produce distinct draws across budgets. Qwen accuracy gains are percentage-point differences; OLMo error reductions are $(E_{\mathrm{FAVOR}} - E_{\mathrm{MINT}})/E_{\mathrm{FAVOR}}$ for mean error E. Alongside pointwise 95% intervals, Bonferroni intervals cover five Qwen accuracy contrasts or fifteen OLMo stage–budget contrasts of MINT# against FAVOR#. Inference conditions on fixed heads and fitted proposals. The saved Qwen data contain all 800 method–budget–book–bank cells.

D.3 Synthetic calibration design

Six two-dimensional reuse groups combine separation $\delta \in \{0, .75, 1.5\}$ and base weights $b = (.25, .25, .25, .25)$ (B) or $(.4, .4, .1, .1)$ (S). Each contains 64 calibration rows and 16 independently generated test rows. For row i, set $x_i = (a_i, 0)$ with $a_i \sim \mathrm{Uniform}(.7, .9)$, and let $\ell_{ij} = \log b_j + \epsilon_{ij}$, $\epsilon_{ij} \sim \mathcal{N}(0, .1^2)$, and $\alpha_{ij} = e^{\ell_{ij}} / \sum_k e^{\ell_{ik}}$. Four keys are given by

$$y_{ij,1} = \frac{\ell_{ij} - \sum_k \alpha_{ik}\ell_{ik}}{a_i}, \qquad y_{ij,2} = \delta t_j + u_i + \eta_{ij}, \qquad t = (-1, -1, 1, 1), \tag{51}$$

where $u_i \sim \mathcal{N}(0, .15^2)$ and $\eta_{ij} \sim \mathcal{N}(0, .1^2)$; all construction draws are independent. Thus α_{ij} are exact softmax weights and $Q_i = \sum_j \alpha_{ij}\mathcal{N}(x_i + y_{ij}, I_2)$.

The proposals are $\mathcal{N}(0, I_2)$, a single Gaussian at the calibration mean weighted by $\alpha_{ij}/64$, and a two-center mixture fitted by weighted k-means. One k-means++ initialization is used, with at most 100 Lloyd steps, 10^{-12} displacement tolerance, and cluster weights floored at .001 then renormalized. Values and test rows do not enter fitting. Proposals remain fixed across $M \in \{16, 64, 256\}$. Each of 4,096 independent banks serves every test row; methods share base noise and budgets use nested prefixes. Full-density correction uses float64 without feature floors or denominator clamps. All scenarios, seeds, budgets, and figure selections were fixed before execution.

Values across keys are $(-1, -1, 1, 1)$ (between groups), $(-1, 1, -1, 1)$ (within groups), or $(1, 1, 1, 1)$ (constant). Metrics average $\chi^2(Q_i \| q)$, $(\widehat{D}_i/D_i - 1)^2$, and $(\widehat{o}_i - o_i)^2$ over test rows. Coverage uses full-dimensional Gauss–Hermite quadrature, converged within $10^{-9} + 10^{-5}|\chi^2|$ by order 96 and checked against analytic Gaussian moments. The 2,000 paired bootstrap resamples re-

tain whole banks across methods, budgets, values, and rows. Pointwise intervals condition on these calibration/test rows; they do not measure variability across newly generated datasets.

E Additional Empirical Results

The following results test the coverage laws, extend the calibration comparison across budgets, and trace coverage and output behavior in saved LLM computations.

E.1 Validation of the coverage laws

The pairwise and shared-row simulations follow the predicted $1/M$ laws (Figure 5). Across the pairwise grid, the median empirical/theoretical RelMSE ratio is .991; for shared rows it is .9990, with range $[.9750, 1.0124]$ and log–log slopes $[-1.0035, -.9950]$. All 18 row settings satisfy the pairwise upper bound. The simulations use float64, iid features and exact correction, without floors or clamps.

The pairwise grid uses $d = 16$, $\|s\| \in \{0, 2, 4\}$, $\|s - c\| \in \{0, 1, 2, 3\}$, variances $\tau \in \{1, 1.25, 2\}$, $M \in \{16, 32, 64, 128, 256, 512, 1024\}$, and 4,096 banks. The $d = 8$ shared-row grid combines separations $\{.25, .625, 1\}$, variances $\{1, 1.5\}$, and weights $(.25, .25, .25, .25)$, $(.55, .20, .15, .10)$, or $(.85, .05, .05, .05)$, using 65,536 banks. Query-aligned key coordinates realize centered log weights at $\|x\| = 4$; separation is the largest orthogonal-key norm and $c = \sum_j \alpha_j(x + y_j) = x$. An independent-per-pair control follows $\sum_j \alpha_j^2 \chi^2(p_j\|q)/M$ while consuming LM draws.

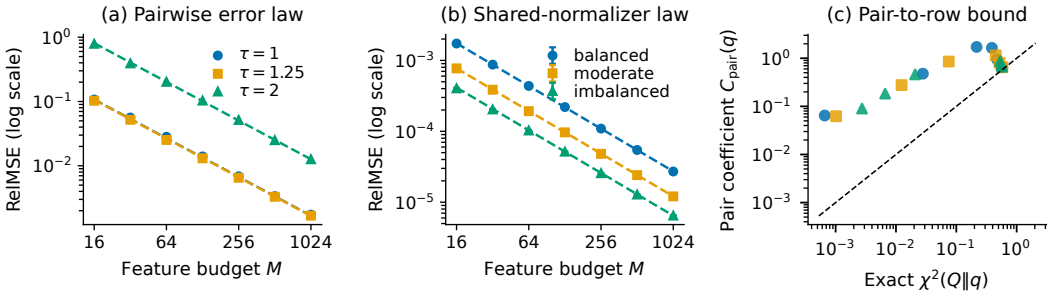

Figure 5: **Simulations reproduce the $1/M$ error laws and pair-to-row bound.** Here M counts iid features and $\text{RelMSE} = \mathbb{E}[(\widehat{T}/T - 1)^2]$ for an estimate $\widehat{T}$ of a scalar T; lower is better. In (a,b), points are empirical errors and dashed curves are exact predictions. (a) T is a query–key kernel value; $s = x + y$ is its target center and c is the proposal center, with $\|s\| = 2$, $\|s - c\| = 1$ and proposal covariance τI. (b) T is the attention normalizer; profiles vary attention-weight concentration at key separation .625 and $\tau = 1$. Bars are 95% Monte Carlo intervals. (c) Each point is one of 18 synthetic row settings, with profile colors as in (b). All lie above equality: the attention-weighted pair coefficient C_{pair} upper-bounds shared-target coverage error $\chi^2(Q\|q)$ for row target Q and proposal q.

E.2 Calibration across feature budgets

The held-out mixture gains in Figure 2 extend across all three tested budgets (Table 5), using the frozen proposals in Appendix D.3. The table covers all six settings and both nonconstant value profiles. Constant values give exact output up to roundoff (squared error below 10^{-28}). The density panel in Figure 2 shows the exact second-coordinate marginal of the mean test target $\bar{Q}$ at 1.5/B; origin and single-center curves nearly coincide there because their main displacement is in the first coordinate.

At $M = 64$, the extra mixture benefit over a single center is about 1.1% in 0/B. At .75/S, the single center reduces denominator MSE to .320 of origin while increasing between-group output MSE to 1.801 times origin. At 1.5/S it also worsens coverage and exhibits a tail-sensitive denominator moment: the analytic MSE is 3.437, versus an observed 1.088 with bootstrap interval $[.191, 2.763]$.

Table 5: **Two-center geometry calibration improves output MSE across feature budgets.** Six two-dimensional synthetic attention settings vary key-group separation and weight balance; values differ between or within groups (12 comparisons per budget). Centers and frequency weights fit 64 calibration rows and freeze across budgets. At M total features, 4,096 iid banks serve 16 independent test rows. Ratios divide two-center output MSE by that of origin $\mathcal{N}(0, I_2)$ or a fitted single center. Below one is better; ranges span the 12 comparisons. Counts give pointwise 95% paired confidence intervals (CI) entirely below one, conditional on the fixed rows.

	Reference: origin		Reference: single center	
M	Ratio range	Upper CI < 1	Ratio range	Upper CI < 1
16	$[0.1968, 0.5862]$	12/12	$[0.3227, 0.9901]$	12/12
64	$[0.0945, 0.6112]$	12/12	$[0.1167, 0.9890]$	12/12
256	$[0.0505, 0.5953]$	12/12	$[0.0449, 0.9914]$	12/12

Resampling cannot recover unsampled tail mass. These fixed-frequency-weight tests isolate geometry calibration; Appendix E.5 separately tests risk-based weight fitting.

E.3 Coverage in Qwen attention tensors

A separate coverage probe examines six fixed Qwen3-4B-Base heads (Table 6). It includes all causal pairs from eight length-1,024 WikiText-103 windows, totaling 25,190,400 pairs. The standard Gaussian poorly covers these pair-contribution regions. With $d = 128$, vectors use standard scaling $x = d^{-1/4} q_{\mathrm{raw}}$ and $y = d^{-1/4} k_{\mathrm{raw}}$, where the native query and key vectors include the model's head normalization and rotary position embedding.

For $s = x + y$, let $a = \|s\| > 0$ and let $F_{\mathcal{N}}$ be the standard normal cumulative distribution function. Projection onto s/a gives

$$A_s = \{\omega : s^\top \omega \geq a^2/2\}, \qquad p_s(A_s) = F_{\mathcal{N}}(a/2), \quad \gamma(A_s) = F_{\mathcal{N}}(-a/2). \tag{52}$$

Every recorded pair has $a > 5.20$, so its region A_s contributes over 99.53% of the kernel integral while receiving under $.464\%$ origin-proposal mass. These pair-specific regions quantify origin-proposal coverage over the fixed calibration heads and chunks.

Table 6: **The standard Gaussian poorly covers Qwen3-4B-Base's kernel-contribution regions.** Each pair's region contains at least 99.53% of its kernel integral but receives at most 0.464% of the standard Gaussian's probability mass. Each row pools all causal pairs of one fixed head across eight length-1,024 WikiText-103 calibration chunks with standard softmax scaling.

	Query–key target norm $\|x + y\|_2$					Origin mass bound
Layer:head	Min.	1st pct.	Median	99th pct.	Max.	$\max \gamma(A_s)$ (%)
7:18	5.204	6.212	7.125	8.182	10.208	0.4638
9:9	5.971	7.062	8.170	9.432	11.630	0.1416
13:16	5.606	6.573	7.642	8.877	10.298	0.2533
21:3	7.235	8.633	9.487	10.445	11.922	0.0149
23:15	6.274	7.693	8.964	10.390	12.844	0.0853
25:12	7.891	8.947	9.995	11.108	12.559	0.0040

For scaled query x and key y, let $s = x+y$. The pair-specific half-space $A_s = \{\omega : s^\top \omega \geq \|s\|_2^2/2\}$ has target mass $p_s(A_s)$ under the contribution density $p_s = \mathcal{N}(s, I)$ and sampling probability $\gamma(A_s)$ under $\gamma = \mathcal{N}(0, I)$. The last column is an upper bound over each head's pairs, rounded upward; pct. denotes percentile.

E.4 LLM coverage and output

Figure 3 uses the frozen proposals from the main comparison on 96 OLMo rows (six heads, eight books, two positions). Its illustrative marginal uses layer 7, head 18, book 88, row 384. The unit direction a is proportional to $\sum_r \pi_r^{(0)} c_r$, using the initial cluster-frequency weights. Projection gives

exact one-dimensional Gaussian-mixture marginals of Q and each proposal, showing component locations and spread relative to the target. The following analytic bound assesses coverage in the full space.

For Gaussian components g_r with $q = \sum_r \pi_r g_r$, define the evaluated pairwise surrogate

$$\overline{C}_{\mathrm{pair}}(q) = \sum_j \alpha_j \left[\min_r \frac{1 + \chi^2(p_j \| g_r)}{\pi_r} - 1 \right] \geq C_{\mathrm{pair}}(q). \tag{53}$$

The inequality follows from $q \geq \pi_r g_r$. For q_F, the origin-centered full-covariance FAVOR# proposal, $\chi^2(Q\|q_F)$ is available analytically from the Gaussian cross moments. Since $\chi^2(Q\|q) \leq C_{\mathrm{pair}}(q) \leq \overline{C}_{\mathrm{pair}}(q)$, $\Gamma(q) < 1$ in (14) certifies an improvement in the actual shared-target divergence. A ratio at least one is inconclusive. The full-dimensional analytic quantities are evaluated on all 96 rows. At $M = (64, 128, 256, 512, 1024)$, the numbers of certified rows are $(80, 80, 80, 81, 80)$ for MINT and $(81, 81, 85, 93, 93)$ for MINT#. These descriptive counts compare complete fitted proposals on the fixed row set, with rows grouped by head and book and sharing feature banks. Supplementary data provide all coverage ratios and output errors under both iid and ORF sampling.

Table 7 separates numerator, value-scaled denominator and cross-term contributions, then applies the estimated normalizer. Entries are empirical means over 16 iid banks and positions 384 and 896 in the six heads. The five methods use the same frozen proposals and eight test books as Figure 4, with independent feature seeds; no denominator clamps occurred. For each row, the pre-ratio term $\|\widetilde{e}\|^2$ has population expectation J_o/M. The accompanying data include the same decomposition under ORF draws. There the cross term is negative in all 25 cells.

Table 8 reports all fifteen comparisons of the budget-calibrated MINT# against FAVOR#, complementing Figure 4. All fifteen mean reductions and pointwise intervals are positive; fourteen Bonferroni-adjusted intervals exclude zero. The adjusted interval crosses zero for KV decoding at $M = 1024$.

E.5 Controlled finite-budget risk calibration

This experiment holds centers and covariance fixed to compare calibration by the finite-budget risk $\mathcal{F}_M$ with coverage and asymptotic-variance objectives. It uses the same six OLMo heads and eight WikiText calibration chunks as the main comparison, at budgets $M \in \{64, 256, 1024\}$. Within each family, all weight objectives use the same components and positive floor. Pairwise fitting minimizes the uniform row average of $C_{\mathrm{pair},i}$; J_o fitting minimizes $\sum_i \beta_i J_{o,i}$ with the document-normalized weights in Appendix C.2. Define each row's radius as $\rho_i = \max_j \|v_{ij} - o_i\|$. Denominator-only and shared-target KL fitting respectively minimize

$$\mathcal{D}(q) = \sum_i \beta_i \rho_i^2 \chi^2(Q_i \| q), \qquad \mathcal{K}(q) = \sum_i \beta_i \rho_i^2 \mathrm{KL}(Q_i \| q). \tag{54}$$

The latter is equivalent, up to a positive scale and a constant, to fitting the normalized mixture $\sum_i \beta_i \rho_i^2 Q_i / \sum_i \beta_i \rho_i^2$ by KL. Thus both controls use value radii and reference-output normalization; the risk objective $\mathcal{F}_M$ additionally retains vector residuals and the random denominator. The main comparison uses the same geometry and risk-based weight calibration, as specified in Appendix D.

The fitting and independent audit pilots each contain 65,536 points per head/geometry, with four stratified rows per chunk as specified in Appendix C.2. The nine starting candidates use a common 262,144-point pilot for their respective objectives, mixing the frequency-weight proposal, mean row target, and normalized assigned-component second-moment envelope in proportions $1/4$, $1/4$, and $1/2$. Full pilot-density correction is used for every candidate. Validation selection uses PG-19 books 28–35 and 32 banks; test uses books 50–65 and 128 banks. The test metric is the document tensor-relative squared error over all 1,024 rows, averaged equally across fixed heads, books and banks. This is relative MSE, whereas the main attention comparisons report relative L_2 error.

Table 9 reports the iid results. Mean error falls against pairwise, denominator, shared-target KL and J_o-only fitting in all six family–budget settings. All six denominator contrasts and all six shared-KL contrasts retain positive intervals after Bonferroni adjustment over these twelve primary comparisons. A direct selector that chooses among the nine candidates independently for each head by validation output error is competitive: the budget fit improves four of six means, and all six pointwise intervals include zero. Adjusted intervals account for 18 exploratory selector contrasts. The

Table 7: **Cross-term cancellation and normalization shape finite-bank output error.** Terms are computed per bank and then averaged over 96 OLMo rows (six heads, eight books, two positions) and 16 iid banks, using M features per bank and absolute squared-L_2 units. The cross term is negative in 24 of 25 cells; normalization increases nine mean errors and decreases sixteen. The first three terms sum to Pre-ratio.

Proposal	M	Numerator $\|u\|^2$	Den. value $\|z\|^2$	Signed cross $-2\langle u, z\rangle$	Pre-ratio $\|u-z\|^2$	Output $\|u-z\|^2/\widehat{d}^2$
FAVOR+	64	3.059	2.112	−4.731	0.441	6.326
	128	11.824	2.765	−9.213	5.375	5.865
	256	3.524	2.000	−4.207	1.316	5.595
	512	2.731	1.842	−3.906	0.667	4.064
	1024	10.802	43.564	−30.755	23.611	3.903
FAVOR++	64	14.706	16.405	−26.206	4.904	12.349
	128	3.157	2.059	−4.641	0.575	6.393
	256	129.636	7.479	−36.927	100.187	8.551
	512	41.615	26.055	12.508	80.179	5.301
	1024	12.398	5.736	−6.160	11.973	4.234
FAVOR#	64	3.015	2.481	−4.750	0.745	4.506
	128	5.187	2.154	−3.579	3.762	4.102
	256	3.187	4.820	−4.917	3.091	3.465
	512	14.580	18.889	−7.888	25.581	2.441
	1024	3.383	2.889	−2.753	3.519	2.404
MINT	64	0.925	0.389	−0.385	0.929	0.613
	128	3.923	0.204	−0.933	3.193	0.914
	256	0.706	0.135	−0.180	0.661	0.291
	512	0.409	0.142	−0.098	0.453	0.301
	1024	0.969	0.046	−0.128	0.887	0.182
MINT#	64	1.637	0.615	−0.679	1.573	0.789
	128	2.439	0.264	−0.570	2.133	0.308
	256	6.762	0.227	−0.655	6.334	1.841
	512	9.937	0.976	−5.095	5.817	0.253
	1024	0.631	0.087	−0.135	0.583	0.173

Let o be the exact attention output and $\widehat{n}, \widehat{d}$ the estimated numerator and denominator, each divided by the exact normalizer. Then $u = \widehat{n} - o$, $z = o(\widehat{d} - 1)$, and the output error is $(u - z)/\widehat{d}$. Den. value denotes the value-scaled denominator contribution. All norms are Euclidean.

Table 8: **MINT# lowers mean OLMo 3 error relative to FAVOR# in all fifteen comparisons.** Each of six fixed heads uses M ORF features. Entries give percentage error reduction $100(1 - E_{\text{MINT\#}}/E_{\text{FAVOR\#}})$, where E averages the indicated error over eight PG-19 books and four banks, with pointwise 95% paired crossed-bootstrap intervals. Positive values favor MINT#.

M	Attention relative L_2	Full-prompt logit KL	KV decode KL
64	74.7 [67.4, 80.3]	51.1 [32.4, 67.1]	77.6 [50.3, 88.0]
128	79.5 [74.3, 83.2]	73.0 [56.4, 84.2]	72.3 [63.3, 78.8]
256	73.8 [66.9, 78.8]	60.2 [53.8, 64.7]	67.2 [42.6, 80.3]
512	81.1 [76.6, 84.7]	62.2 [44.7, 72.3]	55.0 [31.1, 73.9]
1024	65.7 [53.7, 74.1]	55.7 [36.7, 69.9]	57.4 [25.1, 71.8]

All errors use exact softmax as reference. Attention measures the Frobenius norm of output error divided by the exact output norm on 1,024 tokens. Logit and KV errors are mean token-level vocabulary-distribution KL divergences on 4,096-token prompts and 32 teacher-forced decoding steps after 1,024-token exact prefill, respectively.

selector evaluates discrete candidate selection, while the numerical diagnostics below measure the incremental effect of continuous optimization from the initialization chosen by training-pilot risk.

For ORF deployment (Table 3), the same fitted proposals are frozen and evaluated with 128 new banks on the same 16 test books. Thus the comparison tests transfer across sampling laws, con-

Table 9: **Weight-calibration controls with iid deployment.** Each column names a comparator for mixture-weight fitting at M features per head. Entries give test relative-MSE reduction (%) from fitting the finite-budget output risk $\mathcal{F}_M$. Six OLMo heads are fixed; pointwise 95% paired crossed-bootstrap intervals resample 16 PG-19 books and 128 banks. Within each family, all objectives share centers, covariance and positive weight constraints. All means improve against the first four controls; all validation-selector intervals include zero.

Family	M	vs. pairwise	vs. J_o	vs. denominator	vs. shared KL	vs. validation
MINT	64	48.0 [38.7, 56.5]	50.8 [42.4, 58.6]	49.3 [40.3, 57.4]	18.7 [11.2, 26.2]	1.6 [−2.8, 5.9]
MINT	256	39.5 [28.7, 50.1]	41.4 [31.2, 51.4]	42.3 [31.9, 52.2]	18.5 [10.4, 25.8]	−1.7 [−9.5, 4.9]
MINT	1024	29.2 [20.8, 37.8]	31.6 [22.9, 40.2]	35.1 [27.8, 42.7]	23.2 [17.7, 28.7]	7.6 [−0.3, 15.0]
MINT#	64	51.6 [41.1, 61.0]	56.7 [48.0, 64.4]	52.5 [42.0, 61.6]	13.8 [5.5, 21.5]	−1.6 [−10.5, 5.7]
MINT#	256	52.3 [43.3, 60.4]	53.4 [44.5, 61.2]	54.2 [45.3, 62.0]	21.5 [13.2, 29.4]	4.3 [−3.5, 11.8]
MINT#	1024	54.1 [45.8, 62.1]	53.9 [45.4, 61.9]	56.0 [48.3, 63.5]	23.4 [15.9, 30.0]	7.1 [−2.3, 15.4]

Pairwise fits an average kernel-error coefficient; J_o fits the output residual second moment; denominator fits value-weighted normalizer error; shared KL fits a weighted mixture of row-contribution densities by KL divergence. Validation chooses among nine candidates separately for each head by validation output error; these selector contrasts are exploratory. Reduction is $100(1 - E_{\mathrm{risk}}/E_{\mathrm{base}})$, where E averages $\|\widehat{O} - O\|_F^2/\|O\|_F^2$ for approximate/exact attention-output tensors $\widehat{O}, O$. Positive values favor risk fitting.

ditional on these documents. All six contrasts against denominator-only and four of six against shared-target KL remain positive after adjustment over twelve primary ORF comparisons. Both experiments use 10,000 paired crossed-bootstrap resamples of books and bank indices, keeping heads fixed. The analysis includes all prescribed heads, books and banks; saved data include all candidate and per-head results.

Numerical accuracy. In the iid controlled study, the budget-risk pilot ranks the nine candidates against 256 independent calibration banks with mean within-head/budget Spearman correlation 0.829 (independent audit: 0.850). Independent 65,536-point estimates of J_o/M and the pairwise coefficient yield 0.476 and 0.406 on this candidate set. The independent budget-risk pilot's median absolute relative difference from measured risk is 9.2%, and its 90th percentile is 31.5%. Halving the log-quadrature step changes checked risks by at most 2.63×10^{-7} relatively. All 36 optimizations decrease training risk; 32 decrease audit risk and 23 decrease independent-bank calibration error relative to their initialization. Thus the objective ranks candidates well, while improvements from subsequent continuous optimization transfer to independent pilots and banks in a subset of settings.