QORLA: A MEASURE AWARE STATISTICAL SEQUENCE OPERATOR WITH EXACT FAST EVALUATION AND INNOVATION MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer based sequence models rely on dense pairwise interactions whose computational cost grows rapidly with context length. QORLA replaces dense query key scoring with measure aware statistical estimation over learned scalar geometry, while preserving full vector values through Laplacian localization and local linear correction. Its structured formulation admits exact efficient evaluation without materializing the full pairwise interaction matrix, and innovation memory preserves information beyond the statistical predictor. Across diverse sequence modeling domains, QORLA shows competitive predictive performance, strong transfer across architectures, and favorable long context computational scaling. These properties position QORLA as an efficient statistical sequence modeling primitive for scalable alternatives to dense pairwise attention.

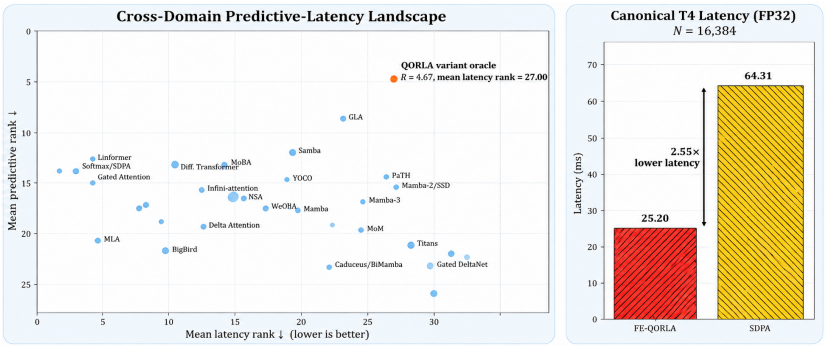


Figure 1: **QORLA across domains and context lengths. Left:** Mean predictive rank versus mean latency rank across the 33 mechanisms in Table 1; lower is better on both axes, and marker area indicates mean peak memory. The highlighted point denotes the domain specific QORLA variant oracle ($\bar{R} = 4.67$). **Right:** Canonical Tesla T4 FP32 latency at $N = 16{,}384$: FE-QORLA takes 25.20 ms versus 64.31 ms for SDPA, giving a $2.55\times$ lower latency.

1 INTRODUCTION

Transformer models have become a standard architecture for modern sequence modeling because self attention provides a powerful way to capture global dependencies across long inputs (Vaswani et al., 2017). However, standard self attention computes pairwise interactions between tokens, leading to quadratic computation and memory growth with sequence length. This cost becomes increasingly restrictive as context windows expand. Considerable effort has therefore been devoted to efficient sequence modeling. FlashAttention and its successors improve the hardware execution of exact dense attention (Dao et al., 2022; Dao, 2024; Shah et al., 2024). Linformer compresses the interaction through low rank projection (Wang et al., 2020), Performer uses random feature

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.

approximations (Choromanski et al., 2020), Nyströmformer reconstructs attention from landmark based approximations (Xiong et al., 2021) and cosFormer introduces a kernel formulation with structured reweighting (Qin et al., 2022). Sparse approaches such as BigBird and Longformer reduce the number of explicit token interactions (Zaheer et al., 2020; Beltagy et al., 2020), while Native Sparse Attention and MoBA extend this idea through adaptive or content dependent routing (Yuan et al., 2025; Lu et al., 2025).

A different family of methods avoids dense pairwise interaction altogether. RetNet introduces retention based sequence processing (Sun et al., 2023), while Gated Linear Attention and Gated DeltaNet use recurrent or fast weight states to summarize sequence information efficiently (Yang et al., 2023; 2025a). Mamba and Mamba 2 use selective state space dynamics to model long sequences without constructing an attention matrix (Gu and Dao, 2023; Dao and Gu, 2024). Persistent memory mechanisms provide another alternative by storing compressed information beyond the immediate interaction step (Munkhdalai et al., 2024; Behrouz et al., 2025). These methods substantially improve scalability but generally obtain efficiency by accelerating dense attention, approximating it, restricting its interaction pattern or replacing it with a compressed state. QORLA follows a different direction by designing the sequence operator itself so that efficient evaluation follows directly from its mathematical structure.

QORLA is a measure aware statistical sequence operator that combines learned scalar geometry, Laplacian localization, a positive learned measure and local linear correction while preserving full vector valued information. This statistical structure admits an exact ordered realization through prefix and suffix sufficient statistics, allowing the dense estimator to be evaluated without materializing the full pairwise interaction matrix. Dense and ordered forms therefore compute the same operator up to numerical precision. Unlike methods based on low rank projection (Wang et al., 2020), kernel approximation (Choromanski et al., 2020; Xiong et al., 2021), sparse interaction (Zaheer et al., 2020; Beltagy et al., 2020; Yuan et al., 2025; Lu et al., 2025) or recurrent and state based compression (Yang et al., 2023; 2025a; Gu and Dao, 2023; Dao and Gu, 2024), QORLA derives efficient evaluation directly from the mathematical structure of the estimator itself.

Figure 1 summarizes QORLA across predictive quality, computational efficiency, memory use and context length, with performance on the vertical axis, efficiency on the horizontal axis and marker size indicating memory footprint. Across this landscape, QORLA demonstrates strong cross-domain competitiveness together with exact nonquadratic evaluation, favorable long-context computational scaling, and substantially reduced dependence on dense pairwise interaction storage. Further evaluations examine long-context prediction, statistical correction, exact forward and gradient agreement, systems scaling, and pretrained-model transfer. Unlike low-rank, sparse, or recurrent alternatives (Wang et al., 2020; Zaheer et al., 2020; Yang et al., 2025a; Dao and Gu, 2024), QORLA derives these advantages from a statistical interaction rule whose structure directly admits exact efficient evaluation.

2 Measure-Aware Scalar Statistical Geometry

Let $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times D}$ with H heads and $d = D/H$. QORLA retains learned $Q/K/V$ projections (Vaswani et al., 2017) but replaces pairwise softmax scoring by measure-aware statistical estimation. Suppressing the head index,

$$z_{0,i} = \frac{\langle q_i, c_q \rangle}{\sqrt{d}}, \qquad z_j = \frac{\langle k_j, c_k \rangle}{\sqrt{d}}, \qquad \mu_j = \mathrm{softplus}(u^\top x_j + \beta_\mu) + \varepsilon_\mu > 0, \tag{1}$$

$$w_{ij} = \mu_j \exp\left(-\frac{|z_j - z_{0,i}|}{\lambda}\right), \qquad \lambda > 0. \tag{2}$$

Here λ is the Laplacian *bandwidth*, so the logarithmic decay rate with scalar-coordinate distance is $-1/\lambda$. Scalar geometry controls relevance only; $v_j \in \mathbb{R}^d$ remains fully vector valued.

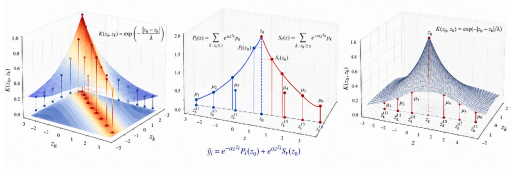


Figure 2: QORLA geometry and exact realization. Learned measures μ_j weight vector values around z_0, while ordered statistics reconstruct Base+R without a dense interaction matrix.

Define

$$M_r = \sum_j w_{ij} z_j^r \ (r = 0, 1, 2), \qquad U_r = \sum_j w_{ij} z_j^r v_j \ (r = 0, 1), \tag{3}$$

and

$$B = \frac{U_0}{M_0}, \quad m = \frac{M_1}{M_0}, \quad \bar{z} = m - z_{0,i}, \quad Q = \frac{M_2}{M_0} - m^2, \quad C = \frac{U_1}{M_0} - mB. \tag{4}$$

QORLA outputs

$$Y_i = B_i - \frac{\bar{z}_i C_i}{Q_i + \rho} = B_i + R_i, \qquad \rho \geq 0. \tag{5}$$

B is a learned measure-aware kernel estimator (Nadaraya, 1964; Watson, 1964), while R is its local-linear correction (Fan and Gijbels, 1996). With $\pi_j = w_{ij}/M_0$ and $V = \sum_j \pi_j \|v_j - B\|_2^2$,

$$\|C\|_2^2 \leq QV, \qquad \|R\|_2 \leq \frac{|\bar{z}|\sqrt{QV}}{Q + \rho} \leq \frac{|\bar{z}|\sqrt{V}}{2\sqrt{\rho}}, \qquad \rho > 0, \tag{6}$$

which controls the correction under nearly degenerate local support.

For an affine field $v_j = c_\star + g_\star z_j$,

$$Y_i - (c_\star + g_\star z_{0,i}) = g_\star \bar{z}_i \frac{\rho}{Q + \rho}. \tag{7}$$

Thus $\rho = 0$, $Q > 0$ exactly reproduce the affine target, removing Base's first-order error under asymmetric support. Moreover, for

$$z' = sz + c, \qquad \lambda' = s\lambda, \qquad \rho' = s^2\rho, \qquad s > 0,$$

the normalized distance $|z_j - z_0|/\lambda$ and hence the output are unchanged. The noncausal operator is also permutation equivariant.

Exact ordered realization. The Laplacian kernel factorizes exactly:

$$\sum_j e^{-|z_j - z_0|/\lambda} f_j = e^{-z_0/\lambda} \sum_{z_j \leq z_0} e^{z_j/\lambda} f_j + e^{z_0/\lambda} \sum_{z_j > z_0} e^{-z_j/\lambda} f_j. \tag{8}$$

For $r \in \{0, 1, 2\}$,

$$P_r^f(z) = \sum_{z_j \leq z} e^{z_j/\lambda} \mu_j z_j^r f_j, \qquad S_r^f(z) = \sum_{z_j > z} e^{-z_j/\lambda} \mu_j z_j^r f_j, \tag{9}$$

so

$$T_r^f(z_0) = e^{-z_0/\lambda} P_r^f(z_0) + e^{z_0/\lambda} S_r^f(z_0), \qquad M_r = T_r^1, \quad U_r = T_r^v. \tag{10}$$

Hence the ordered evaluator reconstructs exactly the same five sufficient statistics as dense QORLA, without approximation or an $N \times N$ interaction matrix.

Theorem 1 (Exactness and complexity)**.** *For nonempty support with $\lambda > 0$ and either $\rho > 0$ or $Q > 0$,*

$$Y^{\mathrm{ord}}(X) = Y^{\mathrm{dense}}(X)$$

in exact arithmetic, with identical gradients wherever differentiable. Sorting and scans require

$$O(HN \log N + ND) \subseteq O(ND \log N)$$

sequence-dependent time and $O(ND)$ auxiliary memory.

Coordinate ties and the Laplacian cusp form measure-zero nondifferentiable boundaries and do not affect forward exactness. Strict causal evaluation inserts only past keys into the same ordered statistics, giving exact causal evaluation in $O(ND \log N)$ time.

QORLA variants. The statistical estimator defines

$$\underbrace{y_t = B_t + R_t}_{\text{QORLA}}, \qquad \underbrace{y_t = B_t + R_t + r_t^{\mathrm{IDM}}}_{\text{QORLA-IDM}}, \qquad \underbrace{y_t = B_t + R_t + r_t^{\mathrm{EIDM}}}_{\text{QORLA-EIDM}}. \tag{11}$$

IDM adds persistent fast-weight storage (Schlag et al., 2021; Yang et al., 2025a); EIDM instead stores the innovation unexplained by the causal predictor:

$$\widehat{v}_t^{\mathrm{past}} = B_t + R_t, \qquad \delta_t = v_t - \widehat{v}_t^{\mathrm{past}}. \tag{12}$$

Both use

$$M_t^h = \gamma_h M_{t-1}^h + \beta_t^h \bar{k}_{t-1}^h s_t^\top, \qquad r_t^h = (\bar{q}_t^h)^\top M_t^h, \tag{13}$$

with

$$s_t = \begin{cases} v_t, & \text{QORLA-IDM}, \\ \delta_t, & \text{QORLA-EIDM}. \end{cases} \tag{14}$$

If $\widehat{v}_t^{\mathrm{past}} = \mathbb{E}[v_t \mid \mathcal{F}_t]$, then

$$\mathbb{E}[\delta_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\delta_t^\top g(\mathcal{F}_t)] = 0 \tag{15}$$

for every square-integrable $\mathcal{F}_t$-measurable vector-valued g, providing the statistical basis for innovation-directed storage.

The exact $O(HN \log N + ND)$ guarantee applies to Base+R. IDM and EIDM are optional memory extensions; full-rank EIDM adds $O(ND^2/H)$ update cost and $O(D^2/H)$ persistent state. Thus QORLA separates exact statistical estimation from optional persistent and innovation-directed memory.

3 Empirical Evidence

A comprehensive empirical study examines QORLA from six complementary perspectives spanning predictive breadth, statistical validity, data-scale behavior, pretrained-model compatibility and realized computational efficiency. A cross-domain benchmark first compares the QORLA family with more than 30 sequence mechanisms across text, streaming, genomics, time series, multimodal learning and vision. A controlled long-context study then compares QORLA with FLA Gated Linear Attention, Mamba-2 and a matched Transformer++-style baseline from 1K to 16K context under smooth, asymmetric and heteroskedastic support. Targeted ablations further isolate the contributions of the learned measure μ and local linear correction R, directly linking empirical behavior to the statistical structure of the operator.

The study is extended to large-scale language modeling, pretrained-model conversion and cross-hardware systems behavior. QORLA, Transformer++, Mamba-2 and Gated DeltaNet are trained at closely matched parameter scale for more than two billion prediction tokens on both FineWeb-Edu

and SlimPajama. Pretrained compatibility is examined through controlled QORLA replacement and adaptation in Gemma 3 270M, including conversion depth, representation fidelity and recovery through QORLA-specific adaptation. Finally, the exact ordered evaluator is tested independently on NVIDIA T4, L4 and A100 GPUs through dense-to-exact latency, memory, numerical precision, forward–backward and long-context scaling measurements. Together these experiments connect predictive quality, statistical mechanism, data-scale training, pretrained-model transfer and exact computational realization within a single empirical framework.

3.1 Cross-Domain Evaluation Across Sequence Mechanisms

The broadest evaluation examines whether QORLA's statistical interaction rule transfers across substantially different sequence-learning regimes. More than 30 mechanisms are compared across text, streaming, genomics, time series, multimodal learning and vision using predictive rank, task performance, latency and peak memory. The reported $\bar{R}$ denotes the mean predictive rank across the six domains and is kept distinct from systems measurements rather than folded into a composite score.

Table 1: Cross-domain predictive rank, task performance, latency and peak memory across six sequence-modeling domains. A denotes accuracy (%)

Mechanism	$\bar{R}$	Text				Stream				Genomics				Time Series				Multimodal				Vision			
		R	A	T	S	R	A	T	S	R	A	T	S	R	A	T	S	R	A	T	S	R	A	T	S
QORLA Variant Oracle	**4.67**	**1**[B]	**70.4**	7.6	61.3	**1**[E]	**4.0**	1575	34.9	**2**[E]	**78.6**	3377	52.9	**4**[B]	**70.45**	8.5	56.1	**5**[E]	**12.7**	766	27.9	**15**[I]	**32.6**	87.8	44.2
GLA (Yang et al., 2023)	8.67	3	67.6	67.3	44.4	3	3.5	67.7	28.8	**1**	79.2	142	38.7	10	69.75	129	38.6	13	11.3	34.4	23.8	22	31.6	35.3	23.8
Samba (Ren et al., 2025)	12.00	16	65.4	23.9	59.5	7	3.0	25.4	43.9	12	73.6	56.6	100.8	3	69.55	47.3	100.8	21	10.5	14.3	27.5	13	33.2	13.9	27.4
Linformer (Wang et al., 2020)	12.67	26	61.8	1.6	45.0	20	2.67	1.9	29.3	3	76.1	1.7	39.3	**1**	82.95	2.0	39.3	25	10.7	1.7	23.8	**1**	39.7	1.7	23.8
Diff. Transformer (Ye et al., 2025)	13.17	2	68.5	2.2	81.5	14	2.92	2.8	65.9	18	74.1	6.9	192.7	27	69.25	6.8	192.7	9	11.9	1.8	32.7	9	33.6	1.9	32.4
MoBA (Lu et al., 2025)	13.33	11	66.1	3.5	60.3	13	2.83	4.5	44.6	21	73.7	13.4	111.4	5	70.45	13.2	111.4	4	12.9	5.5	28.5	26	31.6	3.2	27.1
FlashAttention-3 (Shah et al., 2024)	13.83	12	66.1	0.9	44.4	26	2.5	1.3	28.8	15	73.9	1.9	38.7	23	69.2	2.1	38.6	**1**	13.0	1.3	23.8	6	34.4	1.0	23.8
FlashAttention-2 (Dao, 2024)	13.83	12	66.1	1.0	44.4	26	2.5	1.5	28.8	15	73.9	2.0	38.7	23	69.2	2.2	38.6	**1**	13.0	1.3	23.8	6	34.4	1.0	23.8
Softmax/SDPA (Vaswani et al., 2017)	13.83	12	66.1	0.9	44.4	26	2.5	1.4	28.8	15	73.9	2.0	38.7	23	69.2	2.3	38.6	**1**	13.0	1.3	23.8	6	34.4	1.1	23.8
PaTH (Yang et al., 2025b)	14.33	33	52.0	88.3	44.8	18	2.83	91.8	29.2	19	73.9	198	39.1	2	75.4	185	39.0	10	11.9	50.9	24.2	4	34.6	48.3	24.2
YOCO (Sun et al., 2024)	14.67	22	64.6	22.6	44.5	21	2.5	25.4	28.8	9	73.9	50.0	38.7	6	70.4	50.4	38.7	28	10.5	13.0	23.9	2	35.7	12.9	23.9
Gated Attn. (Qiu et al., 2025)	15.00	6	67.1	1.1	44.4	24	2.58	1.3	28.8	23	73.9	2.1	38.7	13	69.8	2.3	38.6	8	12.0	1.4	23.8	16	33.4	1.2	23.8
Mamba-2/SSD (Dao and Gu, 2024)	15.33	4	67.2	95.7	44.5	23	2.67	92.2	28.8	5	75.7	200	38.7	20	67.35	186	38.7	30	9.4	45.5	23.9	10	33.5	48.4	23.9
Infini-attn. (Munkhdalai et al., 2024)	15.67	8	67.2	3.1	58.4	17	2.83	3.8	42.8	14	73.8	5.4	98.7	16	69.65	5.3	98.7	27	10.6	3.0	26.9	12	33.6	3.4	26.8
HashAttention (Desai et al., 2025)	16.33	5	67.0	5.6	199.7	6	3.17	6.2	184.0	33	72.6	21.7	669.2	30	68.65	21.3	669.1	7	12.3	3.1	63.6	17	33.2	2.8	61.4
NSA (Yuan et al., 2025)	16.50	29	60.1	5.5	60.3	8	3.17	6.4	44.6	13	73.9	17.6	111.4	33	65.4	17.5	111.4	11	11.7	5.6	28.5	5	34.6	5.6	27.1
Mamba-3 (Lahoti et al., 2026)	16.83	9	67.2	75.1	44.4	9	3.08	76.6	28.8	6	74.5	157	38.6	21	68.5	151	38.6	29	10.2	39.6	23.8	27	30.9	40.9	23.8
Longformer (Beltagy et al., 2020)	17.17	20	64.5	1.8	57.4	16	2.83	2.0	41.8	26	73.3	4.8	96.7	7	69.95	4.7	96.7	20	11.0	2.0	26.4	14	33.3	1.9	26.3
Sliding-Window (Beltagy et al., 2020)	17.50	23	63.8	1.9	57.4	12	2.83	2.0	41.8	24	73.5	4.7	96.7	9	69.95	4.7	96.7	16	11.2	1.8	26.4	21	31.9	1.8	26.3
VMoBA (Wu et al., 2026)	17.50	10	66.3	9.5	61.8	5	3.08	12.2	46.2	32	73.1	39.6	114.1	28	69.05	39.9	114.0	6	12.6	9.0	29.6	24	31.6	7.8	28.0
Mamba (Gu and Dao, 2023)	17.67	15	65.5	25.1	44.4	2	3.67	27.6	28.8	7	74.7	49.1	38.6	31	68.25	49.7	38.6	26	10.6	14.3	23.8	25	31.6	14.2	23.8
JetBlock (Gu et al., 2025)	18.83	17	65.5	2.4	44.4	32	2.33	2.7	28.8	20	73.8	3.1	38.7	11	69.7	2.6	38.6	22	10.7	2.8	23.8	11	33.7	2.6	23.8
RetNet (Sun et al., 2023)	19.17	28	60.5	57.1	44.4	10	3.0	60.8	28.8	11	73.8	120	38.7	14	68.4	130	38.6	33	9.5	34.3	23.8	19	32.1	32.5	23.8
Delta Attn. (Willette et al., 2025)	19.33	21	64.5	3.1	57.4	15	2.92	3.4	41.8	26	73.3	5.3	96.7	8	68.45	5.3	96.7	23	10.9	3.2	26.4	23	31.7	3.5	26.3
MoM (Du et al., 2026)	19.67	32	55.6	66.0	44.4	31	2.25	79.2	28.8	8	74.0	159	38.7	12	69.8	145	38.6	32	9.6	42.1	24.0	3	35.5	37.0	24.0
MLA (Liu et al., 2024)	20.67	18	65.3	1.2	44.4	11	2.92	1.6	28.7	22	73.9	2.0	38.6	22	69.4	2.3	38.6	31	9.9	1.4	23.8	20	31.9	1.2	23.8
Titans (Behrouz et al., 2025)	21.17	27	60.7	120	93.3	4	3.25	126	77.6	25	73.8	236	134.5	26	69.25	244	134.5	15	11.3	65.6	61.3	30	26.7	64.2	61.1
BigBird (Zaheer et al., 2020)	21.67	25	62.1	2.1	57.4	19	2.83	2.7	41.8	29	73.1	5.0	96.7	15	68.95	5.0	96.7	24	10.9	2.3	26.4	18	33.3	2.2	26.3
Gated DeltaNet-2 (Hatamizadeh et al., 2026)	22.00	19	64.9	175	78.6	25	2.33	179	62.9	28	73.1	346	72.7	17	69.55	355	72.7	12	11.5	91.4	58.0	31	26.5	86.0	57.9
Log-Linear (Guo et al., 2026)	22.33	31	58.0	231	53.5	22	2.58	246	37.9	4	75.7	536	56.8	32	65.45	511	56.8	17	11.0	132	29.5	28	30.1	128	29.4
Gated DeltaNet (Yang et al., 2025a)	23.17	7	67.3	147	78.5	33	2.17	151	62.7	30	73.0	282	72.6	19	69.5	284	72.6	18	11.1	73.9	57.9	32	24.9	73.5	57.7
Caduceus/BiMamba (Schiff et al., 2024)	23.33	24	63.4	43.6	44.4	29	2.5	49.4	28.8	10	73.6	198	38.7	29	67.3	87.4	38.6	19	11.1	23.4	23.8	29	27.7	26.1	23.8
KDA (Team et al., 2025)	25.83	30	58.8	151	78.5	30	2.25	147	62.8	30	73.0	297	72.7	18	69.6	289	72.6	14	11.2	80.1	57.9	33	24.3	73.4	57.8

R = predictive rank; A = task performance (%); T = latency (ms); S = peak memory (MB). $\bar{R}$ is the mean predictive rank across the six domains. [B] Base, [I] IDM and [E] EIDM. Time-series values are means over ECG5000 and FordA.

The comparison exposes a deliberately broad operating regime rather than a single task-specific advantage. The QORLA variant oracle obtains the lowest mean cross-domain predictive rank among the mechanisms shown, while performance varies across domains, with vision remaining comparatively weaker. The superscripts identify the QORLA family member used in each domain and prevent the family result from being interpreted as a single universal configuration. Complete dataset definitions, per-domain architectures, training protocols, hyperparameters, per-seed results, additional metrics and domain-specific analyses are reported in Appendix B.

3.2 Long-Context Quality and Statistical Geometry

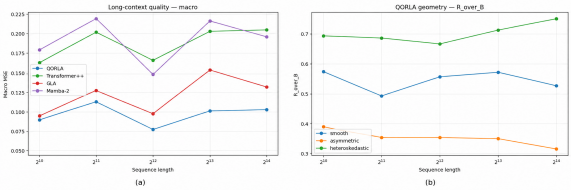
We evaluate length extrapolation after training at $N = 1{,}024$, extending evaluation to $N = 16{,}384$ across smooth, asymmetric, and heteroskedastic regimes. The efficient-mixer comparisons use three seeds and paired evaluation examples. Dense attention is evaluated separately under the powered scope-control protocol reported in Appendix F.5.

Table 2: Long-context quality against efficient sequence mixers and QORLA geometry. Negative Δ indicates lower QORLA MSE.

N	Regime	ΔT++	ΔGLA	ΔMamba-2	N_{eff}/N	$\|R\|/\|B\|$
8K	Smooth	-0.1069	-0.0451	$\mathbf{-0.1203}^*$	0.846	0.572
	Asymmetric	$\mathbf{-0.0905}^*$	$\mathbf{-0.0853}^*$	$\mathbf{-0.1134}^*$	0.893	0.350
	Heteroskedastic	$\mathbf{-0.1078}^*$	-0.0259	$\mathbf{-0.1109}^*$	0.856	0.713
16K	Smooth	$\mathbf{-0.0966}^*$	-0.0123	$\mathbf{-0.0740}^*$	0.856	0.527
	Asymmetric	$\mathbf{-0.0906}^*$	$\mathbf{-0.0455}^*$	$\mathbf{-0.0937}^*$	0.906	0.316
	Heteroskedastic	$\mathbf{-0.1191}^*$	-0.0298	-0.1125	0.854	0.750

* 95% paired CI supports lower QORLA MSE.

Figure 3: Long-context quality and geometry. Macro MSE and relative QORLA correction magnitude.



Analysis. Despite training only at 1K, QORLA remains stable through 16K and maintains lower macro MSE than Transformer++, GLA, and Mamba-2 across the tested length range. At 16K, paired inference supports QORLA over Transformer++ in all three regimes, over Mamba-2 on smooth and asymmetric data, and over GLA under asymmetry. Meanwhile, N_{eff}/N remains high (0.854–0.906) and $\|R\|/\|B\|$ remains substantial (0.316–0.750), indicating that long-context extrapolation preserves broad effective support together with an active, regime-dependent local-linear correction.

Dense attention is evaluated separately under the powered position-free regression scope control in Appendix F.5; its results are not pooled with the efficient-mixer comparison above. Full per-task efficient-mixer errors and confidence intervals are reported in Appendix I.1, with geometry trajectories and diagnostics in Appendix I.2.

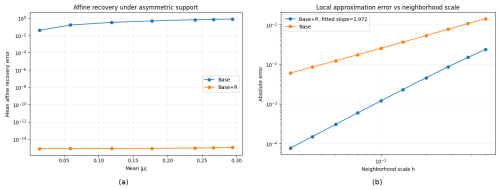
3.3 Empirical Identification of the Statistical Mechanism

We isolate the statistical mechanism underlying QORLA independently of downstream task performance, testing the weighted local-linear identity, affine reproduction, local approximation order, consistency, stability, structural invariances, and distant-token suppression predicted by the Base+R formulation.

Table 3: Statistical-mechanism audit. Numerical tests against the corresponding analytic prediction.

Test	Observed result	Reference
Weighted local-linear identity	max error 2.66×10^{-15}	roundoff
Affine reproduction	Base: $0.0397\to0.7979$; Base+$R\le 3.93\times10^{-15}$	0
Curved boundary	max error: $0.1843\to0.0245$	first-order removal
Second-order local error	slope 1.9716	2
Consistency	MSE $5.98\times10^{-4}\to3.2\times10^{-5}$; slope -0.7652	-0.8
Stability bounds	$\max\|C\|/\sqrt{QV}=1.000000$; $\max\|R\|/\text{bound}=0.999860$	≤ 1
Translation / scale / permutation	5.94×10^{-15} / 1.17×10^{-14} / 6.66×10^{-16}	0
Distant-token suppression	identity error 3.82×10^{-16}; slope -2.9997	$-1/\lambda=-3$ ($\lambda=1/3$)

Figure 4: Mechanism verification. (a) Affine recovery under asymmetric support. (b) Second-order local-error scaling.



Analysis. The controlled audit identifies the behavior predicted by QORLA's weighted local-linear interpretation. Direct ridge-WLS and Base+R agree to 2.66×10^{-15}, while increasing support asymmetry drives the local-constant Base error from 0.0397 to 0.7979 but leaves Base+R within 3.93×10^{-15} of the affine target, directly verifying first-order bias cancellation. For curved signals, the observed local-error exponent is 1.9716, closely matching the predicted quadratic rate, while Monte-Carlo MSE decreases from 5.98×10^{-4} to $\sim 3.2\times10^{-5}$ with tail slope -0.7652 toward the classical $-4/5$ local-linear rate. The Cauchy–Schwarz and ridge bounds remain satisfied under near-degenerate geometry, structural invariances hold to approximately 10^{-14} or better, and and distant-token influence decays with slope -2.9997 under decay rate 3, equivalently bandwidth $\lambda = 1/3$.Together, these results connect QORLA's empirical behavior directly to its proposed statistical mechanism rather than to an implementation-specific effect. Full boundary, ridge, invariance, Monte-Carlo, and distractor diagnostics are reported in Appendix E.2.

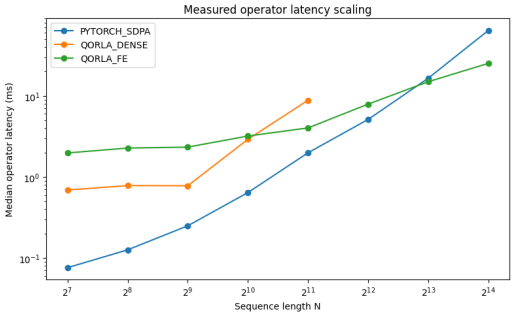
3.4 Exact Evaluation and Realized Long-Context Scaling

We evaluate whether QORLA's exact ordered realization preserves the dense Base+R operator while translating its nonquadratic sequence complexity into measurable long-context behavior. Experiments use a Tesla T4 with $D = 128$, $H = 4$, sequence lengths $N = 128$–16,384, and a controlled irregular-function regression task trained only at $N = 256$.

Table 4: Exactness and long-context systems scaling. Measured FP32 results on Tesla T4.

Evidence	QORLA / FE	Reference	Result
Forward exactness	3.21×10^{-14}	dense Base+R	roundoff
Gradient exactness	3.90×10^{-13}	dense Base+R	roundoff
Strict-online causal	3.11×10^{-15}	dense causal	exact ref.
$N = 2$K latency	4.02 ms	dense: 8.84 ms	2.20×
$N = 8$K latency	14.96 ms	SDPA: 16.52 ms	1.10×
$N = 16$K latency	**25.20** ms	SDPA: 64.31 ms	**2.55**×
Empirical log slope	0.785	SDPA: 1.635	slower growth
$N = 16$K peak alloc.	91.0 MB	SDPA: 8.0 MB	SDPA lower

Figure 5: Long-context systems scaling. Realized operator latency for exact QORLA and the tested dense baselines across increasing context length.



Analysis. QORLA's ordered evaluator reproduces dense Base+R to numerical precision, with worst forward and gradient discrepancies of 3.21×10^{-14} and 3.90×10^{-13}, while the causal Fenwick reference agrees within 1.47×10^{-14}. Realized T4 scaling then exhibits the expected systems transition: FE-QORLA overtakes dense QORLA at $N = 2$K and PyTorch SDPA at $N = 8$K, reaching 25.20 ms at 16K versus 64.31 ms for SDPA (2.55× lower latency), with empirical log-slopes of 0.785 and 1.635, respectively. The gain is specifically a latency result—optimized SDPA retains substantially lower measured allocation (8 MB versus 91 MB at 16K).

The length-extrapolation panel is retained as a numerical-stability diagnostic for checkpoints trained only at $N = 256$ and is not used to support a comparative predictive-quality claim against dense attention. The separate powered dense-attention scope control at 8K and 16K is reported in Appendix F.5, where predictive quality is evaluated with substantially larger paired test sets and uncertainty estimates.

Full forward and gradient exactness, strict-online causal evaluation, and canonical T4 systems measurements are reported in Appendices F.3 and F.4.

3.5 QORLA Inside Real Pretrained Decoders

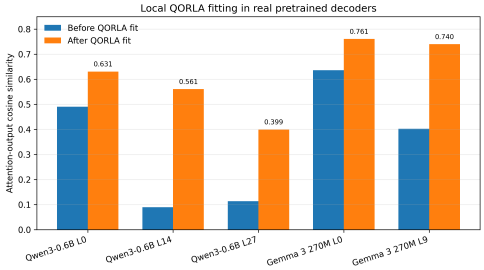
We test whether QORLA can approximate attention operators inside real pretrained decoder checkpoints without retraining the surrounding network. QORLA-specific parameters are fitted locally at selected layers of Qwen3-0.6B and Gemma 3 270M, after which we measure attention-output fidelity and, where numerically available, the language-model effect of a single-layer replacement.

Table 5: Local transfer into real pretrained decoders. Cosine and NRMSE are measured before and after QORLA fitting.

Model	Layer	Cos. before	Cos. after	NRMSE before	NRMSE after	Single-swap PPL
Qwen3-0.6B	0	0.490	**0.631**	0.883	**0.774**	10757.47
Qwen3-0.6B	14	0.090	**0.561**	6.640	**0.628**	**29.28**
Qwen3-0.6B	27	0.114	**0.399**	5.866	**1.950**	32.80
Gemma 3 270M	0	0.636	**0.761**	1.516	**0.748**	177.14
Gemma 3 270M	9	0.402	**0.740**	1.687	**0.889**	–

Qwen3 native PPL: 28.53. Gemma PPL entries with NaN evaluation are omitted.

Figure 6: Pretrained operator fidelity. Attention-output cosine before and after local QORLA fitting.



Analysis. Local QORLA fitting improves attention-output agreement at every numerically valid pretrained probe: cosine rises from 0.090 to 0.561 at Qwen layer 14 and from 0.402 to 0.740 at Gemma layer 9, with corresponding NRMSE reductions from 6.64 to 0.63 and 1.69 to 0.89. Most notably, replacing Qwen layer 14 yields PPL 29.28 versus native 28.53 ($\sim 1.03\times$), while layer 27 remains at 32.80, demonstrating that a scalar QORLA operator can be inserted locally into selected pretrained layers with limited network-level degradation. The effect is strongly layer-dependent:

Qwen layer 0 degrades severely, Gemma PPL evaluation is partially unresolved, and multi-layer replacement accumulates substantial error. We therefore interpret these experiments as evidence for *local pretrained-operator portability*, not complete attention replacement.Canonical cross-family single-layer transfer and the Gemma conversion-depth frontier are reported in Appendix I.6, with the matched 6/18 adaptation control in Appendix I.7. Earlier independent-layer distillation experiments are retained only as exploratory stress tests and are not used as the canonical pretrained-conversion evidence.

3.6 Large-Scale Language Modeling

We evaluate QORLA in controlled 2B-token language-model pretraining on FineWeb-Edu and SlimPajama against Transformer++, Mamba-2, and Gated DeltaNet. All models contain approximately 10.2M parameters and process the same number of prediction tokens, isolating differences in optimization quality and computational efficiency. Table 6 summarizes the final validation and systems measurements.

Table 6: 2B-token LLM pretraining. Matched model scale and token budget.

Corpus	Model	Params.	PPL ↓	tok/s ↑	T4 h ↓
FineWeb-Edu	**QORLA**	10.236M	45.11	144,020	7.72
	Transformer++	10.227M	33.06	229,124	4.85
	Mamba-2	10.255M	34.44	65,532	16.96
	Gated DeltaNet	10.237M	36.79	105,235	10.56
SlimPajama	**QORLA**	10.236M	54.14	143,891	7.72
	Transformer++	10.227M	34.29	224,706	4.94
	Mamba-2	10.255M	35.61	63,596	17.47
	Gated DeltaNet	10.237M	37.39	103,507	10.73

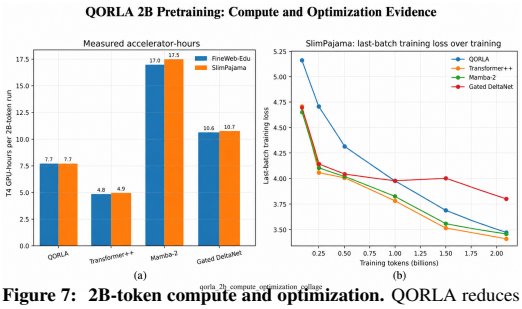


Figure 7: 2B-token compute and optimization. QORLA reduces compute relative to Mamba-2 and Gated DeltaNet while maintaining sustained optimization.

Compute and optimization. At matched $\sim$ 10.2 M-parameter and 2B-token budgets, **QORLA** delivers a strong compute–optimization trade-off across FineWeb-Edu and SlimPajama: it achieves 1.37–1.39$\times$ the throughput of Gated DeltaNet and 2.20–2.26$\times$ that of Mamba-2, while reducing T4 GPU-hours by approximately 27–28% and 54–56%, respectively. Despite starting from a higher-loss regime, QORLA improves steadily throughout pretraining, reaching $\sim$ 3.47 last-batch loss at 2B tokens and substantially narrowing the optimization gap without early saturation. Transformer++ remains stronger in absolute perplexity and raw throughput, but QORLA establishes a distinct efficiency regime among structured sequence models: substantially lower compute than Mamba-2 and Gated DeltaNet, stable cross-corpus throughput, and sustained billion-token optimization. Extended trajectories, validation statistics, throughput accounting, and training details are provided in Appendix F.7.

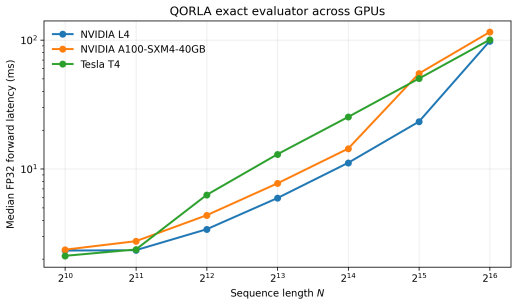
3.7 Cross-GPU Systems Scaling

We profile the identical QORLA exact evaluator across NVIDIA L4, A100-SXM4-40GB, and Tesla T4 GPUs using $B = 1$, $D = 128$, $H = 4$, and 66,316 parameters. FP32 evaluation spans $N = 1$K–65K, with additional dense-reference, backward, mixed-precision, and width-scaling measurements.

Table 7: Cross-GPU QORLA profile. FP32 exact evaluation across three NVIDIA GPUs.

			$N = 8$K Exact				$N = 16$K Train		$N = 65$K
GPU	VRAM (GB)	Cross.	Lat. (ms)	Thrpt. (M tok/s)	Speedup (×)	Memory (×)	F+B (ms)	Thrpt. (M tok/s)	FP32 (ms)
L4	22.03	2K	**5.94**	**1.378**	**24.19**	85.01	**24.81**	**0.660**	**97.91**
A100	39.49	2K	7.73	1.060	4.19	85.01	30.20	0.542	115.05
T4	14.56	1K	12.99	0.631	11.46	85.01	55.72	0.294	100.29
BF16 latency at $N = 65K$									
L4: 46.93 ms			A100: 113.25 ms			T4: 100.53 ms			

Figure 8: Cross-GPU exact scaling. Median FP32 latency across 1K–65K contexts.



Analysis. Across L4, T4, and A100, QORLA scales to $N = 65{,}536$, while the tested dense reference is limited to $N \leq 8{,}192$. At 8K, exact evaluation yields $24.19\times$, $11.46\times$, and $4.19\times$ speedups on L4, T4, and A100, respectively, with a consistent $85.01\times$ reduction in incremental peak allocation. L4 provides the strongest measured operating point, reaching 5.94 ms at 8K, 24.81 ms for 16K forward–backward evaluation, and 97.91 ms at 65K; BF16 further reduces the 65K latency to 46.93 ms. The cross-device variation indicates that realized QORLA efficiency depends on the evaluator's sort/search/prefix-statistics workload rather than nominal accelerator capability alone. Complete latency distributions, precision sweeps, memory profiles, width scaling, and accelerator diagnostics are reported in Appendix E.5.

4 Discussion

QORLA defines a measure-aware local-linear sequence operator over learned scalar geometry with an exact ordered Laplacian realization, supported by affine reproduction, structural invariances, dense/ordered forward and gradient agreement, and strict-online causal exactness. Full-data AG News shows classification quality comparable to dense attention; in the memory-focused IMDB protocol, EIDM improves Base QORLA with negligible parameter overhead, although an independent matched control does not reproduce the gain, making the benefit protocol-dependent. Long-context evaluation shows a measured latency crossover over optimized SDPA while avoiding dense-QORLA's quadratic interaction allocation, and pretrained conversion supports local architectural compatibility across multiple decoder families. The evidence also defines clear limits: short-context causal LM and powered 8K–16K regression favor SDPA in predictive quality, vision remains the weakest tested domain for the current one-dimensional geometry, optimized SDPA can use less realized memory, and full-network conversion is less faithful than local replacement. Accordingly, predictive, systems, and transfer claims are interpreted as protocol- and hardware-scoped rather than universal.

5 Conclusion

QORLA demonstrates that efficient sequence interaction can be formulated as exact statistical estimation rather than as an approximation to a dense attention matrix. Learned scalar geometry, positive token measure, local-linear correction, exact ordered sufficient statistics, and innovation-directed memory form a coherent sequence-modeling primitive with a precise computational contract. The exact $O(ND \log N)$, $O(ND)$ realization applies to scan-compatible Base+R, while EIDM introduces additional full-rank memory cost and dense-anchor IDM does not inherit the same bound. Across classification, long-context systems tests, causal language modeling, memory evaluation, cross-domain studies, and pretrained decoders, the resulting evidence supports QORLA as a principled and scalable alternative for regimes where dense pairwise interaction is unnecessary or computationally restrictive, while clearly identifying the settings in which current dense or specialized mechanisms remain stronger.

REPRODUCIBILITY STATEMENT

The paper separates algebraic claims from empirical and hardware-specific claims. Complete proofs and statistical identities are given in Appendix A; the controlled cross-domain protocol and per-domain results are in Appendix B; numerical forward/gradient exactness, strict-online causal evaluation, precision audits, systems scaling, long-context quality, maintained-reference causal language modeling, innovation-memory ablations, and pretrained conversion are documented in Appendices E and F. Failed, qualified, or confounded evidence is explicitly retained in Appendix F.11.

AI ASSISTANCE STATEMENT

AI tools were used to assist with language polishing and literature retrieval/discovery. All mathematical statements, citations, experimental configurations, numerical results, interpretations, and final manuscript content were reviewed and verified by the authors; AI tools were not treated as an independent scientific authority.

REFERENCES

A. Behrouz, P. Zhong, and V. Mirrokni. Titans: Learning to memorize at test time. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 113506–113543. Curran Associates, Inc., 2025. doi: 10.52202/085713-3786. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4ca07aa108036f80cbb5b82285fd4b1-Paper-Conference.pdf.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 35549–35562, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/98ed250b203d1ac6b24bbcf263e3d4a7-Paper-Conference.pdf.

T. Dao and A. Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16344–16359. Curran Associates, Inc., 2022. doi: 10.52202/068431-1189. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/67d57c32e20fd0a7a302cb81d36e40d5-Paper-Conference.pdf.

H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

A. Desai, S. Yang, A. Cuadron, M. Zaharia, J. E. Gonzalez, and I. Stoica. HashAttention: Semantic sparsity for faster inference. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 13402–13418.

PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/desai25a.html.

J. Du, W. Sun, D. Lan, J. Hu, Z. Tao, and Y. Cheng. Mom: Linear sequence modeling with mixture-of-memories. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 106613–106631, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ae2dd45fb3d7ce1dce45e5de4bdd9ba0-Paper-Conference.pdf.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1996.

K. Gresova, V. Martinek, D. Cechak, P. Simecek, and P. Alexiou. Genomic benchmarks: A collection of datasets for genomic sequence classification. *bioRxiv*, 2022. doi: 10.1101/2022.06.08.495248. URL https://www.biorxiv.org/content/early/2022/06/10/2022.06.08.495248.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Y. Gu, Q. Hu, H. Xi, J. Chen, S. Yang, S. Han, and H. Cai. Jet-nemotron: Efficient language model with post neural architecture search. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 47191–47218. Curran Associates, Inc., 2025. doi: 10.52202/085713-1572. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4364fbe773a77fb373df5c7f1d4cbb53-Paper-Conference.pdf.

G. Guo, S. Yang, T. Goel, E. P. Xing, T. Dao, and Y. Kim. Log-linear attention. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 91296–91317, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/939a9a2604f13729a4ec36ad25d10c60-Paper-Conference.pdf.

A. Hatamizadeh, Y. Choi, and J. Kautz. Gated deltanet-2: Decoupling erase and write in linear attention. *arXiv preprint arXiv:2605.22791*, 2026.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

A. S. Lahoti, K. Li, B. Chen, C. Wang, A. Bick, Z. Kolter, T. Dao, and A. Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, volume 2026, pages 86173–86202, 2026.

A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

E. Lu, Z. Jiang, J. Liu, Y. Du, T. Jiang, C. Hong, S. Liu, W. He, E. Yuan, Y. Wang, Z. Huang, H. Yuan, S. Xu, X. Xu, G. Lai, Y. Chen, H. Zheng, J. Yan, J. Su, Y. Wu, Y. Zhang, Z. Yang, X. Zhou, M. Zhang, and J. Qiu. MoBA: Mixture of Block Attention for Long-Context LLMs. In *Advances in Neural Information Processing Systems*, 2025. URL https://mlanthology.org/neurips/2025/lu2025neurips-moba/.

T. Munkhdalai, M. Faruqui, and S. Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020. URL https://doi.org/10.1137/1109020.

Z. Qin, W. Sun, H. Deng, D. Li, Y. Wei, B. Lv, J. Yan, L. Kong, and Y. Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022.

Z. Qiu, Z. Wang, B. Zheng, Z. Huang, K. Wen, S. Yang, R. Men, L. Yu, F. Huang, S. Huang, D. Liu, J. Zhou, and J. Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 100092–100118. Curran Associates, Inc., 2025. doi: 10.52202/085713-3345. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/904e89bb4e632e75fb47f093b620b257-Paper-Conference.pdf.

L. Ren, Y. Liu, Y. Lu, y. shen, C. Liang, and W. Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 53551–53575, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/84a7fc24ed52e8eff514c33e8ac76ea3-Paper-Conference.pdf.

Y. Schiff, C.-H. Kao, A. Gokaslan, T. Dao, A. Gu, and V. Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of machine learning research*, 235: 43632, 2024.

I. Schlag, K. Irie, and J. Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pages 9355–9366. PMLR, 2021.

J. Shah, G. Bikshandi, Y. Zhang, V. Thakkar, P. Ramani, and T. Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 68658–68685. Curran Associates, Inc., 2024. doi: 10.52202/079017-2193. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7ede97c3e082c6df10a8d6103a2eebd2-Paper-Conference.pdf.

Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Y. Sun, L. Dong, Y. Zhu, S. Huang, W. Wang, S. Ma, Q. Zhang, J. Wang, and F. Wei. You only cache once: Decoder-decoder architectures for language models. *Advances in Neural Information Processing Systems*, 37:7339–7361, 2024.

K. Team, Y. Zhang, Z. Lin, X. Yao, J. Hu, F. Meng, C. Liu, X. Men, S. Yang, Z. Li, et al. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964. ISSN 0581572X. URL http://www.jstor.org/stable/25049340.

J. Willette, H. Lee, and S. J. Hwang. Delta attention: Fast and accurate sparse attention inference by delta correction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 12052–12080. Curran Associates, Inc., 2025. doi: 10.52202/085713-0403. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/11cc3abb01080339ceed66eff7c93920-Paper-Conference.pdf.

J. Wu, L. Hou, H. Yang, Y. Tian, P. Wan, D. ZHANG, and Y. Tong. Vmoba: Mixture-of-block attention for video diffusion models. In *International Conference on Learning Representations*, volume 2026, pages 132534–132548, 2026.

Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14138–14148, 2021.

S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

S. Yang, J. Kautz, and A. Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pages 29687–29707, 2025a.

S. Yang, Y. Shen, K. Wen, S. Tan, M. Mishra, L. Ren, R. Panda, and Y. Kim. Path attention: Position encoding via accumulating householder transformations. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 62220–62247. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2080. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/59c27bf8d56d3d50c7aeaf7535dee975-Paper-Conference.pdf.

T. Ye, L. Dong, Y. Xia, Y. Sun, Y. Zhu, G. Huang, and F. Wei. Differential transformer. In *International Conference on Learning Representations*, volume 2025, pages 144–164, 2025.

J. Yuan, H. Gao, D. Dai, J. Luo, L. Zhao, Z. Zhang, Z. Xie, Y. Wei, L. Wang, Z. Xiao, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23078–23097, 2025.

M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

APPENDIX

A MATHEMATICAL FOUNDATIONS AND CLAIM TRACEABILITY

We give compact proofs of the statistical, algorithmic, stability, and memory properties of QORLA. Consider one head with query coordinate $z_0 \in \mathbb{R}$, scalar key coordinates z_j, positive measures μ_j, and values $v_j \in \mathbb{R}^d$. Define

$$\Delta z_j = z_j - z_0, \qquad w_j = \mu_j \exp\left(-\frac{|\Delta z_j|}{\lambda}\right), \qquad \lambda > 0,$$

$$M_0 = \sum_j w_j, \qquad p_j = \frac{w_j}{M_0}, \qquad m = \sum_j p_j z_j, \qquad \bar{z} = m - z_0.$$

Define

$$B = \sum_j p_j v_j, \qquad Q = \sum_j p_j (z_j - m)^2,$$

$$C = \sum_j p_j (z_j - m)(v_j - B) = \sum_j p_j (z_j - m) v_j,$$

and, for total ridge/stabilization $\rho \geq 0$,

$$R = -\frac{\bar{z}\, C}{Q + \rho}, \qquad Y = B + R.]$$

A.1 STATISTICAL INTERPRETATION AND APPROXIMATION THEORY

Theorem A.1 (Weighted local-linear characterization)**.** *If $Q + \rho > 0$, QORLA is exactly the intercept at the query coordinate of*

$$\min_{c, g \in \mathbb{R}^d} \sum_j p_j \, \|v_j - c - g \Delta z_j\|_2^2 + \rho \|g\|_2^2.$$

Specifically,

$$g^\star = \frac{C}{Q + \rho}, \qquad c^\star = B - \bar{z}\, g^\star = Y.$$

Proof. Stationarity in c gives $c = B - g\bar{z}$. Substituting into the objective yields

$$\sum_j p_j \, \|(v_j - B) - g(\Delta z_j - \bar{z})\|_2^2 + \rho \|g\|_2^2.$$

Since

$$\Delta z_j - \bar{z} = z_j - m,$$

stationarity in g gives

$$(Q + \rho) g = C.$$

Hence

$$g^\star = \frac{C}{Q + \rho}, \qquad c^\star = B - \frac{\bar{z}\, C}{Q + \rho} = Y.$$

□

Theorem A.2 (Affine reproduction and first-order bias cancellation)**.** *If locally*

$$v_j = c_\star + g_\star z_j,$$

then

$$B = c_\star + g_\star m, \qquad C = g_\star Q,$$

and

$$Y - (c_\star + g_\star z_0) = g_\star \bar{z} \frac{\rho}{Q + \rho}. \tag{A.2}$$

Thus, if $\rho = 0$ and $Q > 0$,

$$\boxed{Y = c_\star + g_\star z_0},$$

whereas Base alone has first-order error

$$B - (c_\star + g_\star z_0) = g_\star \bar{z}.$$

For $\rho > 0$,

$$\|Y - (c_\star + g_\star z_0)\|_2 = \|g_\star\|_2 |\bar{z}| \frac{\rho}{Q + \rho}.$$

Proof. Substitution gives

$$B = \sum_j p_j (c_\star + g_\star z_j) = c_\star + g_\star m.$$

Moreover,

$$C = \sum_j p_j (z_j - m) g_\star (z_j - m) = g_\star Q.$$

Insert these identities into (A.1) and use $m = z_0 + \bar{z}$. □

Theorem A.3 (Finite-sample second-order error)**.** *Let $v_j = f(z_j)$ with $f \in C^2$ and*

$$\sup_{x \in \mathcal{N}(z_0)} \|f''(x)\|_2 \le M.$$

Define

$$\widetilde{M}_2 = \sum_j p_j (\Delta z_j)^2, \qquad \widetilde{M}_{1,2} = \sum_j p_j |\Delta z_j - \bar{z}| (\Delta z_j)^2.$$

Then

$$\boxed{\|Y - f(z_0)\|_2 \le \|f'(z_0)\|_2 |\bar{z}| \frac{\rho}{Q+\rho} + \frac{M}{2}\left(\widetilde{M}_2 + \frac{|\bar{z}|\widetilde{M}_{1,2}}{Q+\rho}\right).} \tag{A.3}$$

If $\rho = 0$, $|\Delta z_j| \le h$, and $Q \ge \kappa h^2$, then

$$\boxed{\|Y - f(z_0)\|_2 \le \frac{M}{2}\left(1 + \frac{2}{\kappa}\right) h^2.} \tag{A.4}$$

Proof. Taylor-expand

$$f(z_0 + \Delta z_j) = f(z_0) + f'(z_0)\Delta z_j + r_j, \qquad \|r_j\| \le \frac{M}{2}(\Delta z_j)^2.$$

Writing

$$\bar{r} = \sum_j p_j r_j, \qquad E_r = \sum_j p_j (\Delta z_j - \bar{z}) r_j,$$

gives

$$B = f(z_0) + f'(z_0)\bar{z} + \bar{r}, \qquad C = f'(z_0)Q + E_r.$$

Therefore

$$Y - f(z_0) = f'(z_0)\bar{z}\frac{\rho}{Q+\rho} + \bar{r} - \frac{\bar{z}\, E_r}{Q+\rho}.$$

Use

$$\|\bar{r}\| \le \frac{M}{2}\widetilde{M}_2, \qquad \|E_r\| \le \frac{M}{2}\widetilde{M}_{1,2}.$$

For (A.4), note

$$\widetilde{M}_2 \le h^2, \qquad |\bar{z}| \le h, \qquad \widetilde{M}_{1,2} \le 2h^3, \qquad Q \ge \kappa h^2.$$

□

Theorem A.4 (Local-linear consistency and rate)**.** *Let (U_j, V_j) be i.i.d. with*

$$\mathbb{E}[V_j \mid U_j = u] = f(u).$$

Assume: $f \in C^2$ near z_0; the density of U is positive and continuous at z_0; the positive measure satisfies

$$0 < \mu_{\min} \le \mu(u) \le \mu_{\max} < \infty$$

locally; conditional second moments are bounded;

$$h_N = \lambda_N \to 0, \qquad N h_N \to \infty;$$

and

$$\rho_N = o(h_N^2).$$

Then

$$\mathrm{Bias}[Y_N(z_0)] = O(h_N^2), \qquad \mathrm{Var}[Y_N(z_0)] = O((N h_N)^{-1}),$$

hence

$$\boxed{\mathbb{E}\|Y_N(z_0) - f(z_0)\|_2^2 = O\left(h_N^4 + \frac{1}{N h_N}\right).} \tag{A.5}$$

Choosing

$$h_N \asymp N^{-1/5}$$

gives

$$\boxed{\mathbb{E}\|Y_N(z_0) - f(z_0)\|_2^2 = O(N^{-4/5}).} \tag{A.6}$$

Proof. By Theorem A.1, Y_N is a one-dimensional local-linear estimator with Laplacian kernel

$$K(t) = e^{-|t|},$$

whose moments satisfy

$$\int K = 2, \qquad \int tK(t)\,dt = 0, \qquad \int t^2 K(t)\,dt = 4, \qquad \int K^2 = 1.$$

Under the bandwidth convention

$$K\left(\frac{U_j - z_0}{h_N}\right) = \exp\left(-\frac{|U_j - z_0|}{h_N}\right), \qquad h_N = \lambda_N.$$

A second-order Taylor expansion of $f(z_0 + h_N t)$ and the local-linear normal equations cancel the constant and first-order terms, leaving bias $O(h_N^2)$. The effective local sample size is $N h_N$, yielding variance $O((N h_N)^{-1})$. Squaring the bias and adding variance proves (A.5); balancing the two terms gives $h_N \asymp N^{-1/5}$ and (A.6). □

Remark A.5. Theorem A.4 is conditional on the learned scalar coordinate being fixed; it does not assert global convergence of the neural optimization problem.

Theorem A.6 (Regression-correction stability)**.** *Let*

$$V = \sum_j p_j \|v_j - B\|_2^2.$$

Then

$$\|C\|_2 \le \sqrt{QV},$$

and, for every $\rho > 0$*,*

$$\boxed{\|R\|_2 \le \frac{|\bar{z}|\sqrt{V}}{2\sqrt{\rho}}.} \tag{A.7}$$

Proof. Weighted Cauchy–Schwarz gives

$$\|C\|^2 \le QV.$$

Hence

$$\|R\| \le |\bar{z}|\sqrt{V}\frac{\sqrt{Q}}{Q+\rho}.$$

Since

$$Q + \rho \ge 2\sqrt{Q\rho},$$

we have

$$\frac{\sqrt{Q}}{Q+\rho} \le \frac{1}{2\sqrt{\rho}}.$$

□

Proposition A.7 (Convex-hull property)**.** *Since* $\mu_j > 0$ *and the Laplacian kernel is positive,* $p_j \ge 0$ *and* $\sum_j p_j = 1$*. Thus*

$$B \in \operatorname{conv}\{v_1, \dots, v_N\}, \qquad \|B\|_2 \le \max_j \|v_j\|_2.$$

Proof. Immediate from

$$B = \sum_j p_j v_j.$$

□

A.2 Evaluation and Complexity

Theorem A.8 (Sufficient-statistic closure)**.** *Define*

$$M_r = \sum_j w_j z_j^r \qquad (r = 0, 1, 2),$$

$$U_0 = \sum_j w_j v_j, \qquad U_1 = \sum_j w_j z_j v_j.$$

Then Base+R depends only on

$$\{M_0, M_1, M_2, U_0, U_1\}.$$

Specifically,

$$B = \frac{U_0}{M_0}, \qquad m = \frac{M_1}{M_0}, \qquad \bar{z} = m - z_0,$$
$$Q = \frac{M_2}{M_0} - m^2, \qquad C = \frac{U_1}{M_0} - mB. \tag{A.8}$$

Proof. The identities for B, m, and $\bar{z}$ follow by normalization. Moreover,

$$Q = \sum_j p_j (z_j - m)^2 = \frac{M_2}{M_0} - m^2,$$

and

$$C = \sum_j p_j (z_j - m) v_j = \frac{U_1}{M_0} - mB.$$

□

Theorem A.9 (Exact Laplacian ordered decomposition)**.** *For arbitrary scalar- or vector-valued* f_j*,*

$$\boxed{\sum_j e^{-|z_j - z_0|/\lambda} f_j = e^{-z_0/\lambda} \sum_{z_j \le z_0} e^{z_j/\lambda} f_j + e^{z_0/\lambda} \sum_{z_j > z_0} e^{-z_j/\lambda} f_j.} \tag{A.9}$$

Consequently every statistic in Theorem A.8 is obtained exactly by setting

$$f_j \in \{\mu_j,\ \mu_j z_j,\ \mu_j z_j^2,\ \mu_j v_j,\ \mu_j z_j v_j\}.$$

Proof. For $z_j \le z_0$,

$$e^{-|z_j - z_0|/\lambda} = e^{-z_0/\lambda} e^{z_j/\lambda},$$

whereas for $z_j > z_0$,

$$e^{-|z_j - z_0|/\lambda} = e^{z_0/\lambda} e^{-z_j/\lambda}.$$

Partitioning the sum gives (A.9). □

Theorem A.10 (Exact non-causal FE-QORLA complexity)**.** *For sequence length N, total width D, H heads and $d = D/H$, exact non-causal Base+R evaluation requires*

$$\boxed{O(HN \log N + ND) \subseteq O(ND \log N)} \tag{A.10}$$

time and

$$\boxed{O(HN + ND) = O(ND)} \tag{A.11}$$

auxiliary memory, excluding learned projections.

Proof. Per head, sorting scalar key coordinates costs $O(N \log N)$; scalar prefix/suffix statistics cost $O(N)$; the two vector statistics cost $O(Nd)$; and N split queries cost $O(N \log N + Nd)$. Summing over H heads yields (A.10). Scalar scans require $O(HN)$ and vector scans $O(ND)$ memory. Exactness follows from Theorems A.8–A.9. □

Theorem A.11 (Exact causal evaluation)**.** *Causal Base+R QORLA can be evaluated exactly in*

$$\boxed{O(ND \log N) \textit{ time}, \qquad O(ND) \textit{ state}} \tag{A.12}$$

using an augmented balanced tree over scalar key coordinates (or a Fenwick tree when coordinates are known/compressed).

Proof. At time t, the tree stores the available key payloads

$$(\mu_t,\ \mu_t z_t,\ \mu_t z_t^2,\ \mu_t v_t,\ \mu_t z_t v_t)$$

at key coordinate z_t. For query coordinate $z_{0,t}$, prefix aggregation gives the $z_j \le z_{0,t}$ statistics and total-minus-prefix gives the $z_j > z_{0,t}$ statistics. Each scalar tree operation costs $O(\log N)$ and each vector payload costs $O(d \log N)$; over N tokens and H heads this is $O(ND \log N)$. The tree stores $O(Nd)$ statistics per head, hence $O(ND)$ state. Equation (A.9) guarantees exactness. For a strict-online causal realization, only admissible past-key payloads are present when the query is evaluated. □

Theorem A.12 (QORLA-EIDM complexity)**.** *Let each head maintain $M_t^h \in \mathbb{R}^{d \times d}$ with $d = D/H$. With FE-QORLA as anchor, EIDM requires*

$$\boxed{O\left(HN \log N + ND + \frac{ND^2}{H}\right)} \tag{A.13}$$

time and

$$\boxed{O\left(ND + \frac{D^2}{H}\right)} \tag{A.14}$$

state/auxiliary memory.

Proof. The FE anchor costs $O(HN \log N + ND)$. Per head, the rank-one update

$$M_t = \gamma M_{t-1} + \beta_t \bar{k}_{t-1} \delta_t^\top$$

and retrieval

$$\bar{q}_t^\top M_t$$

each cost $O(d^2)$ per token, hence

$$O(NHd^2) = O(ND^2/H).$$

Persistent memory is

$$Hd^2 = D^2/H.$$

□

Theorem A.13 (Almost-everywhere dense–scan gradient equivalence)**.** *Let*

$$\Omega = \{(z_0, z) : z_j \neq z_k\ (j \neq k),\ z_j \neq z_{0,i}\ \forall i, j\}.$$

On every connected component of Ω, dense and exact-scan QORLA have identical outputs and derivatives:

$$\boxed{\nabla_\theta Y_{\text{dense}} = \nabla_\theta Y_{\text{scan}}.} \tag{A.15}$$

Hence equality holds almost everywhere.

Proof. Inside one component of Ω, key ordering and every left/right split remain fixed. Equation (A.9) is therefore only an algebraic rearrangement of the dense sums; both implementations are the same smooth function and have identical derivatives. The excluded tie/split boundaries form a finite union of measure-zero hypersurfaces. □

A.3 INNOVATION-MEMORY THEORY

Lemma A.14 (Innovation orthogonality)**.** *Let $\mathcal{F}_t$ be the predictor information and*

$$\hat{v}_t = \mathbb{E}[v_t \mid \mathcal{F}_t].$$

For

$$\delta_t = v_t - \hat{v}_t$$

and any square-integrable $\mathcal{F}_t$-measurable g_t,

$$\boxed{\mathbb{E}[\delta_t^\top g_t] = 0.} \tag{A.16}$$

Proof. By iterated expectation,

$$\mathbb{E}[\delta_t^\top g_t] = \mathbb{E}\left[g_t^\top \mathbb{E}[\delta_t \mid \mathcal{F}_t]\right] = 0.$$

□

Theorem A.15 (Minimum-energy innovation principle)**.** *For any square-integrable $\mathcal{F}_t$-measurable predictor g_t,*

$$\boxed{\mathbb{E}\|v_t - g_t\|_2^2 = \mathbb{E}\|v_t - \hat{v}_t\|_2^2 + \mathbb{E}\|\hat{v}_t - g_t\|_2^2.} \tag{A.17}$$

Thus the innovation $v_t - \hat{v}_t$ has minimum expected residual energy among all $\mathcal{F}_t$-measurable predictors.

Proof. Use

$$v_t - g_t = (v_t - \hat{v}_t) + (\hat{v}_t - g_t).$$

Expanding the squared norm, the cross term vanishes by Lemma A.14; the remaining second term is nonnegative. □

Theorem A.16 (Bounded innovation memory)**.** *For*

$$M_t = \gamma M_{t-1} + \beta_t \bar{k}_{t-1} \delta_t^\top,$$

assume

$$0 \le \gamma \le \bar{\gamma} < 1, \qquad |\beta_t| \le \bar{\beta}, \qquad \|\bar{k}_t\|_2 \le 1, \qquad \|\delta_t\|_2 \le \Delta.$$

Then

$$\boxed{\|M_t\|_F \le \bar{\gamma}^t \|M_0\|_F + \frac{\bar{\beta}\Delta(1-\bar{\gamma}^t)}{1-\bar{\gamma}}.} \tag{A.18}$$

Therefore

$$\boxed{\limsup_{t\to\infty} \|M_t\|_F \le \frac{\bar{\beta}\Delta}{1-\bar{\gamma}}.} \tag{A.19}$$

If

$$\|\bar{q}_t\|_2 \le 1 \qquad \textit{and} \qquad r_t = \bar{q}_t^\top M_t,$$

then

$$\boxed{\|r_t\|_2 \le \|M_t\|_F, \qquad \limsup_t \|r_t\|_2 \le \frac{\bar{\beta}\Delta}{1-\bar{\gamma}}.} \tag{A.20}$$

Proof. Since

$$\|xy^\top\|_F = \|x\|_2 \|y\|_2,$$
$$\|M_t\|_F \le \bar{\gamma}\|M_{t-1}\|_F + \bar{\beta}\Delta.$$

Iterating this geometric recurrence proves (A.18)–(A.19). Furthermore,

$$\|r_t\|_2 \le \|\bar{q}_t\|_2 \|M_t\|_2 \le \|M_t\|_F.$$

□

A.4 GEOMETRY, INVARIANCE AND ROBUSTNESS

Theorem A.17 (Translation invariance)**.** *For*

$$z_0' = z_0 + c, \qquad z_j' = z_j + c,$$

with all other quantities unchanged,

$$\boxed{Y'(z_0') = Y(z_0).} \tag{A.21}$$

Proof. Since

$$z_j' - z_0' = z_j - z_0,$$

the kernel weights are unchanged. Furthermore

$$m' = m + c, \qquad \bar{z}' = m' - z_0' = \bar{z},$$

while Q, C, B, and therefore Y are unchanged. □

Theorem A.18 (Coordinate-rescaling equivariance)**.** *For $s > 0$, let*

$$z_0' = sz_0 + c, \qquad z_j' = sz_j + c, \qquad \lambda' = s\lambda, \qquad \rho' = s^2\rho.$$

Then

$$\boxed{Y'(z_0') = Y(z_0).} \tag{A.22}$$

Proof. The normalized distance is invariant:

$$\frac{|z'_j - z'_0|}{\lambda'} = \frac{|s(z_j - z_0)|}{s\lambda} = \frac{|z_j - z_0|}{\lambda}.$$

Hence p_j is unchanged. Moreover,

$$m' = sm + c, \qquad \bar{z}' = s\bar{z}, \qquad Q' = s^2 Q, \qquad C' = sC.$$

Therefore

$$R' = -\frac{(s\bar{z})(sC)}{s^2(Q+\rho)} = R,$$

while $B' = B$. □

Theorem A.19 (Permutation equivariance)**.** *For the non-causal QORLA interaction operator without order-dependent external positional features, any common permutation P_π of query/key/value/measure tokens satisfies*

$$\boxed{\mathrm{QORLA}(P_\pi X) = P_\pi \mathrm{QORLA}(X).} \tag{A.23}$$

Proof. Every output is formed from sums over the complete key set. A common permutation only reorders summands, while the output follows its permuted query. Sorting in FE-QORLA is purely an internal exact evaluation procedure. □

Proposition A.20 (Exponential suppression of a distant token)**.** *Let*

$$M_0 = \sum_j w_j > 0.$$

Add a token with positive measure $\mu_\star$, value $v_\star$, and scalar-coordinate distance

$$r_\star = |z_\star - z_0|.$$

Its unnormalized statistical weight is

$$u = \mu_\star e^{-r_\star/\lambda}.$$

If B^+ is the new Base output,

$$\boxed{\|B^+ - B\|_2 = \frac{u}{M_0 + u}\|v_\star - B\|_2 \le \frac{\mu_\star}{M_0} e^{-r_\star/\lambda}\|v_\star - B\|_2.} \tag{A.24}$$

Proof. Since

$$B^+ = \frac{M_0 B + u v_\star}{M_0 + u},$$

we have

$$B^+ - B = \frac{u(v_\star - B)}{M_0 + u},$$

and $M_0 + u \ge M_0$. □

Theorem A.21 (Distant-token robustness of Base+R)**.** *Assume a stable compact operating region with*

$$M_0 \ge M_{0,\min} > 0, \qquad Q + \rho \ge q_{\min} > 0,$$

and uniformly bounded key/value/measure payloads. Then there exists $C_{\mathrm{stab}} < \infty$, independent of $r_\star$, such that adding one token at scalar distance $r_\star$ satisfies

$$\boxed{\|Y^+ - Y\|_2 \le C_{\mathrm{stab}} \mu_\star e^{-r_\star/\lambda}.} \tag{A.25}$$

More generally, without a uniform key-coordinate bound the right-hand side is

$$O\Big(P(|z_\star|, \|v_\star\|)\, \mu_\star e^{-r_\star/\lambda}\Big),$$

for the low-degree polynomial P induced by the sufficient statistics.

Proof. By Theorem A.8,

$$Y = F(M_0, M_1, M_2, U_0, U_1).$$

On the stated stable compact set,

$$\|\nabla F\| \le L_F.$$

The added token perturbs each sufficient statistic by its Laplacian factor

$$u = \mu_\star e^{-r_\star/\lambda}$$

multiplied by a bounded payload; hence

$$\|\Delta S\| \le C_S u.$$

The mean-value theorem gives

$$\|Y^+ - Y\| \le L_F C_S u.$$

□

Proposition A.22 (Independent multi-head statistical geometry)**.** *If head h has its own*

$$(z_{0,i}^h, z_j^h, \mu_j^h, \lambda_h, \rho_h),$$

then Theorem A.1 applies independently to every head. Thus

$$Y_i^h = \text{Intercept} \arg\min_{c,g} \left\{ \sum_j p_{ij}^h \left\| v_j^h - c - g(z_j^h - z_{0,i}^h) \right\|^2 + \rho_h \|g\|^2 \right\}. \tag{A.26}$$

Consequently, concatenation followed by W_O can combine distinct learned scalar orderings and neighborhood bandwidths.

Proof. All quantities in Theorem A.1 are head-specific, so the optimization decomposes independently over heads; W_O then mixes their outputs. □

A.5 Compact Summary of Guarantees

The preceding results establish that QORLA: (i) is exactly a weighted local-linear estimator; (ii) reproduces affine signals and removes Base's first-order local bias; (iii) has second-order local approximation error and the standard one-dimensional $N^{-4/5}$ local-linear MSE rate under classical assumptions; (iv) has ridge-controlled correction magnitude; (v) depends only on five ordered sufficient statistics; (vi) admits exact non-causal and causal scan/tree evaluation; (vii) has almost-everywhere dense–scan gradient equivalence; (viii) uses minimum-energy innovations for persistent memory; (ix) has bounded innovation states under contractive decay; and (x) enjoys translation/scale invariances, permutation equivariance for the non-causal interaction operator, and exponentially decaying sensitivity to distant scalar-coordinate distractors.

A.6 Theory-to-Evidence Traceability

The mathematical statements above are deliberately separated from empirical claims. Table 8 records where each theorem-level property is tested numerically and which conclusions remain protocol-specific. This mapping is intended to prevent asymptotic, numerical, and predictive claims from being conflated.

B Controlled Six-Domain Operator Survey and Representation Atlas

B.1 Survey Protocol and Scope

We evaluate QORLA across six heterogeneous sequence-modeling regimes: **text classification**, **causal associative recall**, **genomic sequence classification**, **time-series classification**, **multimodal fusion**, and **vision classification**. The objective is not to reproduce the full training recipes of every competing architecture, but to isolate the sequence operator under a controlled matched-wrapper protocol. Within each domain, all mechanisms use the same representation width, depth, optimization family, data-split logic, batch construction, and training budget.

The controlled survey configuration uses one deterministic seed ($s = 11$), three training epochs, hidden width $D = 64$, four heads, two sequence-processing layers, and batch size 64. Where the dataset size permits, the survey caps training, validation, and test sets at 2400, 600, and 1000 examples, respectively. Experiments were executed on a Tesla T4 GPU. We report balanced accuracy (BAL), accuracy, macro-F1, negative log-likelihood (NLL), measured inference latency, peak memory, trainable parameters, and analytical time/space scaling. The principal ranking statistic throughout this section is BAL.

The comparison covers dense attention, IO-aware exact attention, low-rank projection, kernel/linear attention, sparse attention, block routing, recurrent fast-weight models, state-space models, persistent-memory systems, and recent hybrid mechanisms. External architectures are implemented as controlled operator- or style-faithful mechanisms unless explicitly indicated otherwise. Consequently, the rankings below describe the present matched-wrapper experiment and must not be interpreted as official leaderboard reproductions of the cited systems.

Cross-domain overview. The unified comparison reveals a deliberately non-uniform performance profile rather than a single mechanism dominating every modality. QORLA is the strongest controlled mechanism for text classification, QORLA-EIDM is strongest for causal associative recall, and QORLA-EIDM is within 1.14 BAL points of the genomics leader. Base QORLA remains fourth in the two-dataset time-series aggregate. The multimodal result is similarly close to the best controlled baseline, whereas vision constitutes the clearest weakness. This heterogeneity is useful diagnostically: the Base statistical operator is strongest when local scalar geometry is itself sufficient, whereas persistent innovation memory becomes most useful when information must survive causal compression over time.

Protocol isolation from the main confirmatory experiments. The 70.37% AG News BAL reported in this appendix belongs exclusively to the compact survey protocol ($D = 64$, two layers, one seed, three epochs and capped data). The main-paper five-seed full-data AG News experiment uses a different $D = 128$, four-layer protocol and reports QORLA at $91.2342 \pm 0.0932\%$ BAL versus SDPA at $91.2553 \pm 0.0835\%$. The two experiments answer different questions—broad inductive-bias survey versus confirmatory full-data classification—and are never pooled or ranked together.

B.2 Text Classification

Dataset. We evaluate text classification on **AG News** (Zhang et al., 2015), a four-class topic-classification benchmark covering World, Sports, Business, and Sci/Tech news. Text is lower-cased, split into whitespace-delimited tokens, and mapped through a deterministic hashing tokenizer with vocabulary size 32,768. Sequences are truncated or padded to a fixed length of $N = 128$.

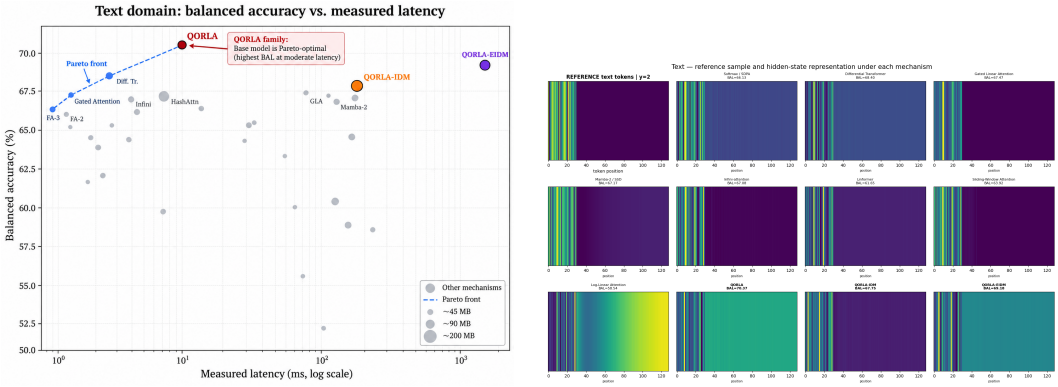
Table 8: **Theory-to-evidence map.** Analytical guarantees apply to the stated operator; empirical latency, memory, and predictive claims remain restricted to their cited protocols.

Claim	Mathematical source	Numerical / empirical audit	Scope boundary
Weighted local-linear identity	Theorem A.1	Appendices E.2 and F.2	Base+R is exactly the stated weighted-ridge estimator; this does not imply universal downstream superiority.
Affine reproduction and first-order bias cancellation	Theorem A.2	Affine and asymmetric-support audits in Appendix E.2	Exact for affine signals under the theorem assumptions; nonlinear behavior is treated separately by the local-error analysis.
Second-order local behavior and consistency	Theorems A.3 and A.4	Scaling and Monte-Carlo diagnostics in Appendix E.2	Uses classical one-dimensional local-linear assumptions; the $N^{-4/5}$ rate is not claimed as QORLA-specific novelty.
Exact ordered evaluation	Theorems A.8, A.9, and A.10	Dense–FE forward and gradient audits in Appendices E.3 and F.3	Exact realization of the same Base+R operator up to floating-point effects; not an approximation to dense QORLA.
Strict-online causal evaluation	Theorem A.11	Dynamic ordered-tree audit in Appendix F.3	Exact online evaluation in $O(NDL\log N)$ time without future-key access.
FE-anchor EIDM complexity	Theorem A.12	Complexity contract in Appendix F.4	Applies to the FE statistical anchor plus the stated memory update; dense IDM follows a different complexity path.
Gradient equivalence	Theorem A.13	Projected-gradient audits in Appendices E.3 and F.3	Almost-everywhere equivalence away from ordering and tie nondifferentiabilities.
Innovation-memory principle	Lemma A.14; Theorems A.15 and A.16	Controlled memory diagnostics and IMDB studies in Appendices E.7, F.8, and I.5	Benefit is predictor- and protocol-dependent; universal EIDM superiority is not claimed.
Geometry invariances and distractor decay	Theorems A.17, A.18, A.19, and A.21	Structural and distant-token audits in Appendix E.2	Guarantees apply to the stated interaction operator; downstream network components may introduce additional non-invariances.

Table 9: **Canonical six-domain screen-run summary.** The best QORLA-family member is compared with the domain champion from the same one-seed, three-epoch matched-wrapper survey. Measured ratios are QORLA/champion; analytical ratios use the operator accounting of this survey. These values must not be mixed with the separate five-seed full-data AG News confirmation.

Domain	Best QORLA	Rank	BAL	Leader BAL	Ret. (%)	Lat. ratio	Mem. ratio	Param. ratio	Analyt. time ratio	Analyt. space ratio
Streaming	QORLA-EIDM	1	4.14	4.14	100.00	1.000	1.000	1.000	1.000	1.000
Text	QORLA	1	70.37	70.37	100.00	1.000	1.000	1.000	1.000	1.000
Genomics	QORLA-EIDM	2	78.70	79.84	98.57	23.737	1.369	1.011	1.500	1.000
Time Series	QORLA	4	43.91	56.23	78.10	4.336	1.430	0.785	0.250	0.667
Multimodal	QORLA-EIDM	5	12.31	12.56	98.05	588.175	1.174	1.017	0.339	1.246
Vision	QORLA-IDM	15	32.97	39.69	83.09	50.733	1.855	0.985	5.000	4.200

Protocol. All mechanisms are evaluated under the same matched-wrapper setting. The original training data are stratified into training and validation partitions, with the survey configuration capped at 2400 training, 600 validation, and 1000 test examples. Every mechanism uses the same $D = 64$ hidden width, two sequence-processing layers, four heads, mean pooling, classifier head, optimizer family, batch size, data split, and three-epoch training budget. This isolates the effect of the sequence operator while keeping the surrounding architecture and optimization conditions fixed.



(a) Accuracy–efficiency. Balanced accuracy versus measured latency; marker size represents peak memory. QORLA lies on the measured Pareto frontier together with strong efficient baselines.

(b) Representation atlas. Hidden representations of the same held-out AG News sample across representative attention, recurrent, and QORLA-family mechanisms.

Figure 9: **AG News analysis.** Left: balanced-accuracy–latency trade-off under the matched wrapper. Right: mechanism-dependent representation geometry for the same held-out sample.

Results. Base **QORLA** obtains the strongest balanced accuracy at 70.37%, followed by QORLA-EIDM at 69.18% and Differential Transformer (Ye et al., 2025) at 68.40%. QORLA exceeds Gated Linear Attention (Yang et al., 2023) by 2.90 BAL points, Mamba-2/SSD (Dao and Gu, 2024) by 3.20 points, and the matched Softmax/FlashAttention operator by 4.24 points. The accuracy–latency landscape further shows that Base QORLA is Pareto-optimal in the present survey: it achieves the highest predictive quality without requiring persistent innovation memory.

This result supports the central design hypothesis of QORLA: reducing query–key interaction to learned scalar statistical geometry does not necessarily sacrifice discriminative sequence information. On AG News, the Base+R formulation provides the strongest controlled-wrapper result while retaining an exact $O(NDN)$ evaluation path and $O(ND)$ auxiliary-space complexity.

B.3 Causal Streaming and Associative Recall

QORLA attains the highest balanced accuracy in its matched cross-domain text evaluation at 70.37%, while the independent six-epoch modern comparison identifies KDA, Differential Transformer and cosFormer as the strongest additional text mechanisms, reaching 77.367%, 77.267% and 77.200% BAL respectively. The modern competitors remain closely matched in parameter count but exhibit substantially different systems profiles: KDA achieves the highest predictive score at 193.848 ms, Differential Transformer reaches nearly identical quality at 9.148 ms, cosFormer at 4.981 ms and Linformer provides the lowest measured latency at 1.997 ms with lower predictive quality. The two result groups are reported together for breadth but retain their original experimental provenance rather than being treated as a single protocol-identical ranking.

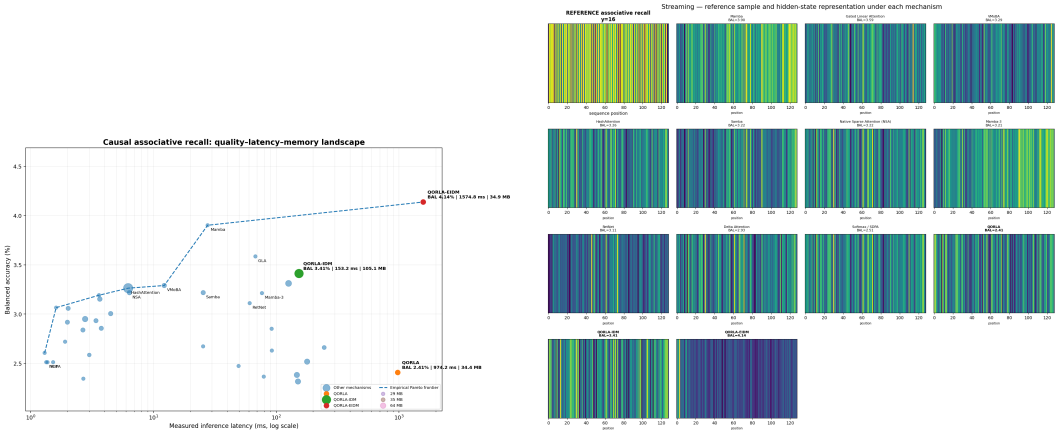
Results. QORLA-EIDM obtains the strongest streaming balanced accuracy at 4.14%, followed by Mamba (Gu and Dao, 2023) at 3.90%, GLA (Yang et al., 2023) at 3.58%, and QORLA-IDM at 3.41%. Other competitive mechanisms include Titans (Behrouz et al., 2025) (3.31%), VMoBA (Wu et al., 2026) (3.29%), and HashAttention (Desai et al., 2025) (3.26%), whereas Base QORLA reaches 2.41%.

The contrast within the QORLA family isolates the effect of persistent innovation memory. Base QORLA relies only on the causal statistical predictor, whereas EIDM additionally preserves residual information not explained by the current statistical neighborhood. Within the QORLA family, EIDM improves balanced accuracy from 2.41% for Base QORLA to 4.14%. Because 32-way chance performance is 3.125%, however, these results are interpreted only as relative within-protocol evidence that innovation memory improves this difficult streaming setting, rather than as evidence of solved associative recall.

B.4 Genomic Sequence Classification

Dataset. We evaluate genomic sequence modeling on the **human non-TATA promoter** task from Genomic Benchmarks (Gresova et al., 2022). The binary task distinguishes promoter from non-promoter DNA sequences using the alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, \mathrm{N}\}$. Sequences are truncated or padded to $N = 256$. The benchmark contains $12{,}355$ negative and $14{,}742$ positive training examples, together with $4{,}119$ negative and $4{,}915$ positive test examples.

Protocol. The original training partition is split stratifiably into training and validation subsets, with the controlled survey capped at 2400 training, 600 validation, and 1000 test sequences. All mechanisms use the same $D = 64$ matched wrapper, two layers, four heads, optimizer family, batch size, data partitions, and three-epoch training budget. For Caduceus/BiMamba (Schiff et al., 2024), reverse-complement averaging is applied in the matched wrapper to preserve its genomic inductive bias.



(a) Quality–latency–memory. Streaming BAL versus measured inference latency on a logarithmic scale; marker size denotes peak memory. The QORLA family is highlighted relative to recurrent, sparse, state-space, and dense sequence mechanisms.

(b) Representation atlas. Hidden-state structure for the same held-out associative-recall sequence across representative mechanisms including Mamba (Gu and Dao, 2023), GLA (Yang et al., 2023), RetNet (Sun et al., 2023), and the QORLA family.

Figure 10: **Causal associative-recall analysis.** Left: measured quality–latency–memory trade-off. Right: mechanism-dependent hidden-state representations for the same held-out causal sequence.

Table 10: **Causal associative-recall operating points.** All rows use the same controlled survey protocol. Gap and retention are computed relative to the domain leader; latency and peak memory are measured.

Mechanism	Rank	BAL ↑	Gap (pp)	Ret. (%)	Latency (ms) ↓	Peak MB ↓	Params
QORLA-EIDM	**1**	**4.14**	+0.00	**100.00**	1574.82	34.95	146.5K
Mamba	2	3.90	-0.24	94.31	27.65	28.80	136.5K
Gated Linear Attention	3	3.58	-0.55	86.65	67.74	28.80	145.2K
QORLA-IDM	4	3.41	-0.73	82.48	153.15	105.07	147.1K

Results. Gated Linear Attention (Yang et al., 2023) obtains the highest balanced accuracy at 79.84%, while QORLA-EIDM follows closely at 78.70%, a gap of only 1.14 percentage points and 98.57% quality retention. Linformer (Wang et al., 2020), Log-Linear Attention (Guo et al., 2026), and Mamba-2/SSD (Dao and Gu, 2024) reach 75.84%, 75.80%, and 75.74%, respectively. QORLA-EIDM also yields a lower NLL than GLA (0.448 versus 0.463), despite the small BAL difference.

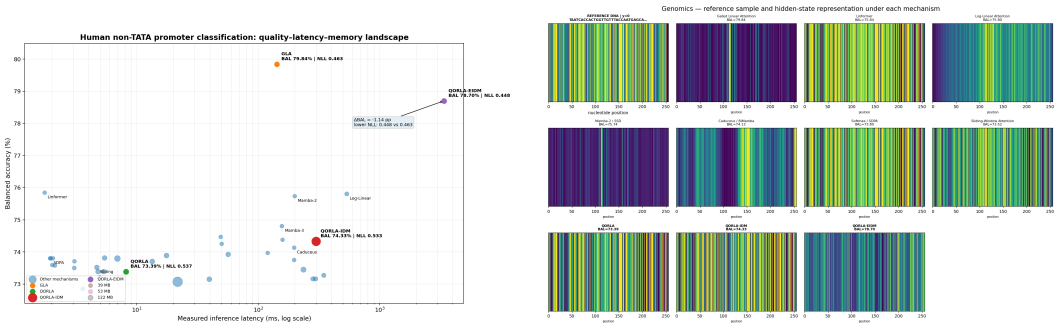
The genomic experiment further validates the exact fast-evaluation backend. Dense and scan-compatible evaluation of the EIDM statistical component agree to approximately 2.1×10^{-7}, supporting the numerical equivalence of the exact sufficient-statistic scan. Thus, QORLA-EIDM remains highly competitive on genomic sequence classification without replacing its statistical interaction with a low-rank approximation.

B.5 Time-Series Classification

Datasets. We evaluate time-series classification on **ECG5000** and **FordA** from the UCR Time Series Archive (Dau et al., 2019). ECG5000 is a five-class electrocardiogram classification task, whereas FordA is a binary classification benchmark derived from automotive sensor signals. Each sequence is independently standardized and resampled to a common length of $N = 256$.

Protocol. Under the controlled survey configuration, ECG5000 contains 425 training, 75 validation, and 1000 test examples, while FordA uses 2400 training, 541 validation, and 1000 test examples. All mechanisms share the same $D = 64$ matched wrapper, two sequence layers, four heads, optimizer family, batch-size logic, and three-epoch training budget. Domain-level performance is reported as the arithmetic mean of the ECG5000 and FordA balanced accuracies.

Results. Linformer (Wang et al., 2020) obtains the strongest aggregate performance with 56.23% mean BAL, followed by PaTH Attention (Yang et al., 2025b) at 49.33%. Base **QORLA** reaches 43.91%, with QORLA-IDM at 43.81% and



(a) Quality–latency–memory. Promoter-classification BAL versus measured inference latency; marker size denotes peak memory. GLA and the QORLA family are highlighted against recurrent, low-rank, dense, and genomic sequence mechanisms.

(b) Representation atlas. Hidden-state structure for the same held-out DNA sequence under GLA (Yang et al., 2023), Linformer (Wang et al., 2020), Mamba-2/SSD (Dao and Gu, 2024), Caduceus (Schiff et al., 2024), and the QORLA family.

Figure 11: **Human non-TATA promoter analysis.** Left: measured quality–latency–memory trade-off. Right: mechanism-dependent token representations of the same held-out genomic sequence.

Table 11: **Genomics leader versus the best QORLA-family operating point.** Values are from the same human non-TATA promoter survey; gap and retention are relative to Gated Linear Attention.

Mechanism	Rank	BAL ↑	Gap (pp)	Ret. (%)	Latency (ms) ↓	Peak MB ↓	Params
Gated Linear Attention	**1**	**79.84**	+0.00	**100.00**	142.25	38.67	116.9K
QORLA-EIDM	**2**	**78.70**	-1.14	**98.57**	3376.74	52.92	118.2K

QORLA-EIDM at 43.32%. On ECG5000 specifically, Base QORLA attains 89.6% accuracy and 37.83% BAL, closely matching PaTH in balanced accuracy, whereas its FordA performance remains near the balanced binary baseline.

Unlike causal associative recall, persistent innovation memory does not materially improve the time-series result: the three QORLA variants remain tightly clustered in aggregate quality. The principal advantage of Base QORLA in this setting is therefore computational rather than predictive: it retains the exact $O(ND \log N)$ evaluation path and $O(ND)$ auxiliary-space requirement while remaining competitive under the common parameter budget.

B.6 Controlled Multimodal Fusion

Dataset. We evaluate a deliberately controlled multimodal fusion task built from CIFAR-10 images (Krizhevsky and Hinton, 2009) paired with a symbolic text-offset token. Each image is patchified into 64 visual tokens and one additional symbolic token is appended, giving a total sequence length of $N = 65$. The target is defined as

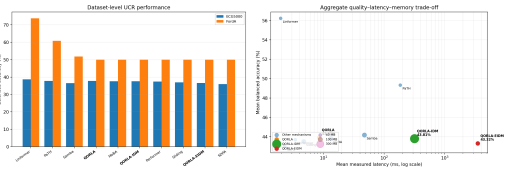
$$y_{\text{fusion}} = (y_{\text{image}} + \Delta_{\text{text}}) \bmod 10,$$

so neither the image stream nor the symbolic token alone is sufficient to solve the task.

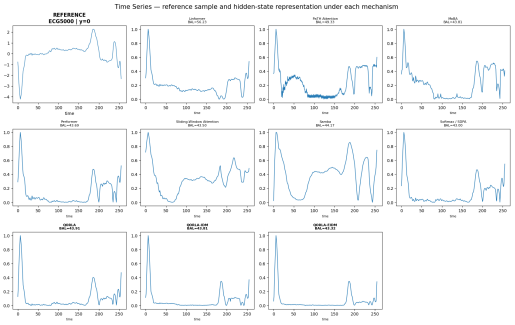
Protocol. All mechanisms are evaluated under the same matched multimodal wrapper, with shared hidden width, depth, optimizer family, batch-size logic, data split, and training budget. For clarity of visual comparison, the main figure focuses on ten representative external baselines: Softmax/SDPA (Vaswani et al., 2017), FlashAttention-2 (Dao, 2024), FlashAttention-3 (Shah et al., 2024), MoBA (Lu et al., 2025), VMoBA (Wu et al., 2026), HashAttention (Desai et al., 2025), Gated Attention (Qiu et al., 2025), Differential Transformer (Ye et al., 2025), NSA (Yuan et al., 2025), and GLA (Yang et al., 2023), together with the QORLA family.

Results. The matched dense-attention controls corresponding to FlashAttention-2, FlashAttention-3, and SDPA attain the same predictive result of 12.56% BAL under the common wrapper. QORLA-EIDM follows at 12.31%, corresponding to approximately 98.05% predictive-quality retention relative to this control. These entries compare model quality rather than optimized FlashAttention kernel performance. Among the remaining focused baselines, MoBA reaches 12.22%, Gated Attention 12.03%, Differential Transformer 11.83%, and HashAttention 11.80%, while Base QORLA and QORLA-IDM obtain 10.56% and 10.20%, respectively.

The quality gap between FlashAttention-2(matched dense-attention control) and QORLA-EIDM is therefore small in absolute terms, but the efficiency interpretation is different from the raw leaderboard. The scan-compatible EIDM statistical



(a) UCR performance and efficiency. Left: dataset-level BAL on ECG5000 and FordA. Right: mean BAL versus measured latency, with marker size encoding peak memory across representative mechanisms.



(b) ECG5000 representation atlas. Hidden-state responses of the same ECG5000 test sequence under Linformer (Wang et al., 2020), PaTH (Yang et al., 2025b), MoBA (Lu et al., 2025), Performer (Choromanski et al., 2020), Samba (Ren et al., 2025), Softmax/SDPA (Vaswani et al., 2017), and the QORLA family.

Figure 12: **Time-series analysis.** Left: dataset-specific and aggregate quality–latency–memory behavior. Right: mechanism-dependent representation of the same held-out ECG5000 sequence.

Table 12: **Time-series leader versus the best QORLA-family operating point.** BAL, latency, peak memory, and parameter counts are arithmetic means over ECG5000 and FordA; gap and retention are relative to Linformer.

Mechanism	Rank	Mean BAL ↑	Gap (pp)	Ret. (%)	Mean lat. (ms) ↓	Mean MB ↓	Mean params
Linformer	**1**	**56.23**	+0.00	**100.00**	1.95	39.26	149.0K
QORLA	**4**	**43.91**	-12.31	**78.10**	8.47	56.15	117.0K

anchor retains

$$O(ND \log N + ND^2/H)$$

time complexity with

$$O(ND + D^2/H)$$

auxiliary space, avoiding the explicit quadratic score structure of dense attention. In the present research implementation this analytical advantage does not translate into lower wall-clock latency than fused FlashAttention kernels, so the multimodal result should be interpreted as a *quality-retention plus scaling-efficiency* result rather than a runtime win.

B.7 Vision Classification

Dataset. We evaluate vision classification on **CIFAR-10** (Krizhevsky and Hinton, 2009). Each image is divided into non-overlapping 4×4 patches, producing $N = 64$ tokens with 48 raw features per patch before projection into the common hidden space.

Protocol. The controlled survey uses 2400 training, 600 validation, and 1000 test images. All mechanisms share the same $D = 64$ matched wrapper, two sequence layers, four heads, optimizer family, batch size, data split, and three-epoch training budget. The focused comparison includes Linformer (Wang et al., 2020), YOCO (Sun et al., 2024), PaTH (Yang et al., 2025b), NSA (Yuan et al., 2025), FlashAttention-2 (Dao, 2024), Softmax/SDPA (Vaswani et al., 2017), Differential Transformer (Ye et al., 2025), JetBlock/Jet-Nemotron (Gu et al., 2025), Infini-attention (Munkhdalai et al., 2024), Samba (Ren et al., 2025), and the QORLA family.

Results. Linformer (Wang et al., 2020) achieves the strongest balanced accuracy at 39.69%, followed by YOCO at 35.59% and PaTH at 34.41%. The best QORLA-family result is QORLA-IDM at 32.97%, while QORLA-EIDM reaches 32.53% and Base QORLA 30.86%. Several dense and sparse baselines lie in the 33–34.3% range, confirming that vision remains the weakest domain for the current QORLA formulation.

The efficiency interpretation is therefore important. Base QORLA retains its exact $O(ND \log N)$ time and $O(ND)$ auxiliary-space evaluation path, but does not match the strongest vision baselines. QORLA-IDM improves predictive quality by introducing the routed full-rank innovation-memory path, whose prototype contains a quadratic anchor. The present vision

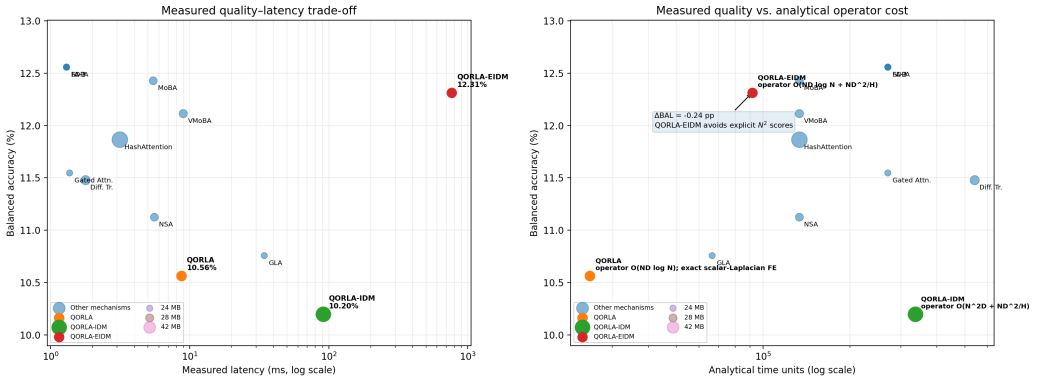


Figure 13: **Controlled multimodal quality–efficiency trade-off.** Left: balanced accuracy versus measured inference latency, with marker size proportional to peak memory. Right: balanced accuracy versus analytical operator cost, separating measured implementation latency from asymptotic scaling. QORLA-EIDM remains close to the strongest dense attention baselines while exhibiting a different computational scaling profile.

Table 13: Controlled multimodal dense-attention reference versus the best QORLA-family operating point. Gap and retention are relative to the matched dense-attention predictive control. Latency and peak memory correspond to the common research implementation rather than optimized FlashAttention kernels.

Mechanism	Rank	BAL ↑	Gap (pp)	Ret. (%)	Latency (ms) ↓	Peak MB ↓	Params
FlashAttention-2(dense-control eqv.)	**1**	**12.56**	+0.00	**100.00**	1.30	23.81	107.5K
QORLA-EIDM	**5**	**12.31**	-0.24	**98.05**	765.69	27.94	109.4K

result should therefore be treated as a clear limitation rather than an accuracy–efficiency win, motivating future work on preserving scan-evaluable QORLA geometry while better capturing spatial interactions.

C Cross-Domain Efficiency and Representation Diagnostics

The domain experiments measure quality at relatively short sequence lengths; the theoretical benefit of the exact statistical operator becomes increasingly visible as N grows. Under the analysis' fixed dimensional accounting, at $N = 16{,}384$ dense attention requires approximately 1.718×10^{10} analytical time units, whereas Base QORLA requires approximately 1.468×10^{7} units, corresponding to a ratio of roughly $1170\times$. The analogous dense score-storage proxy is approximately 2.695×10^{8} units versus 1.049×10^{6} for Base QORLA, a ratio of approximately $257\times$.

These quantities describe operator scaling rather than measured CUDA runtime. FlashAttention-style implementations (Dao, 2024; Shah et al., 2024) retain quadratic arithmetic while eliminating explicit materialization of the dense score matrix and exploiting highly optimized fused kernels. QORLA instead changes the operator algebra itself. Consequently, prototype wall-clock measurements and analytical sequence-length scaling answer complementary rather than interchangeable questions.

Representation geometry. The analysis' representation atlas records effective rank, top-energy concentration, token-norm variation, adjacent-token cosine similarity, singular-value spectra, and pooled representation norms. The observed effective rank varies substantially by domain: the time-series representations are highly concentrated, the text and genomics representations occupy an intermediate regime, and the streaming representations exhibit substantially larger effective rank. This behavior is consistent with the intended interpretation of QORLA: scalarization constrains the *interaction geometry*, but does not force the transported vector-valued representation itself to have globally rank-one structure.

D Ablation and Failure Analysis

D.1 Base Statistical Estimation versus Innovation Memory

The three final variants isolate increasingly expressive components:

$$\text{QORLA} \;\rightarrow\; \text{QORLA}-\text{IDM} \;\rightarrow\; \text{QORLA}-\text{EIDM}.$$

Their domain-wise behavior constitutes a controlled architectural ablation. Base QORLA is strongest on text and is the least structurally complex model. EIDM is strongest on causal associative recall and nearly matches the genomics and multimodal leaders, indicating that persistent residual storage is particularly useful when information must survive beyond the local

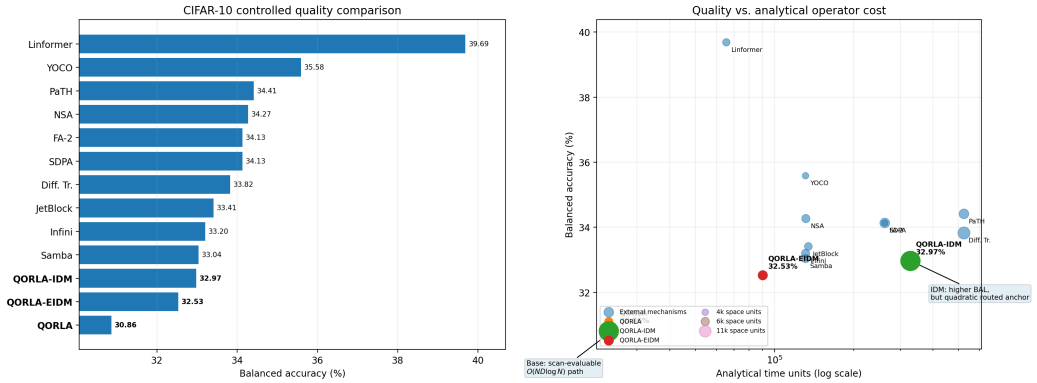


Figure 14: **CIFAR-10 quality–complexity analysis.** Left: balanced-accuracy comparison for the focused mechanism set. Right: balanced accuracy versus analytical operator-time cost, with marker size encoding analytical space. Base QORLA and QORLA-IDM are shown separately to distinguish the scan-evaluable Base operator from the higher-cost innovation-memory auxiliary path.

Table 14: **Vision leader versus the best QORLA-family operating point.** Gap and retention are relative to Linformer under the same controlled CIFAR-10 survey protocol.

Mechanism	Rank	BAL ↑	Gap (pp)	Ret. (%)	Latency (ms) ↓	Peak MB ↓	Params
Linformer	**1**	**39.69**	+0.00	**100.00**	1.73	23.81	111.4K
QORLA-IDM	**15**	**32.97**	-6.71	**83.09**	87.82	44.18	109.8K

statistical predictor. IDM provides the highest QORLA-family result for vision, but does so while sacrificing the exact Base operator's favorable asymptotic path.

D.2 Exact-Evaluation Audit

The exact fast-evaluation backend is checked numerically against the corresponding dense statistical computation under the compact survey implementations. In these finite-precision survey audits, maximum discrepancies are on the order of 10^{-6} for Base QORLA in vision and approximately 2.1×10^{-7} for QORLA-EIDM in genomics. These values are implementation- and precision-specific floating-point discrepancies and should not be pooled with the dedicated FP64 or canonical mixed-stress exactness audits reported later. The latter provide the primary numerical evidence for dense–ordered equivalence. The result applies to the scan-compatible statistical component and must not be extended automatically to additional dense routed auxiliary terms.

D.3 Survey Completion and Failure Accounting

The compact survey retains failed evaluations as failures rather than assigning artificially low scores. Table 15 records the primary result-frame accounting before separately repaired controlled proxies are introduced.

BASED, Performer, and cosFormer are non-finite in several primary survey domains. Nyströmformer remains non-finite in the dedicated repair suite. Separately repaired Gated DeltaNet, Gated DeltaNet-2, KDA, and Titans entries are explicitly marked as controlled repaired proxies and are not presented as official-system reproductions.

D.4 Training-Failure Audit

Non-finite evaluations are retained explicitly rather than being assigned artificially poor accuracy values. BASED, Performer, and cosFormer produce non-finite training behavior in several primary survey domains, while the original Nyströmformer proxy remains non-finite in the dedicated repair suite. Separately repaired controlled implementations of Gated DeltaNet, Gated DeltaNet-2, KDA, and Titans are reported as such. These observations diagnose the present compact implementations and short optimization budget; they are *not* evidence that the corresponding published architectures are intrinsically unstable.

E Mechanism Identification, Exactness, and Systems Diagnostics

Appendix B reports the 36-mechanism controlled comparison. This section isolates mechanism verification, exact evaluation, systems scaling, long-context diagnostics, innovation memory, component sensitivity, and pretrained-decoder transfer. Instead, this appendix tests the statistical mechanism, exact evaluator, systems scaling, long-context behavior, innovation memory, component sensitivity, and decoder transfer underlying the central QORLA claims.

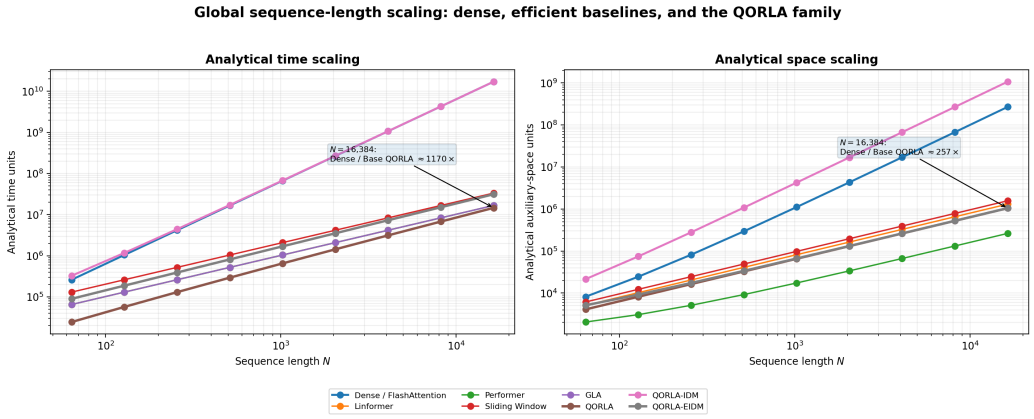


Figure 15: **Analytical sequence-length scaling.** Log–log time (left) and auxiliary-space (right) trajectories for dense attention and FlashAttention (Vaswani et al., 2017; Dao, 2024; Shah et al., 2024), representative efficient mechanisms including Linformer (Wang et al., 2020), Performer (Choromanski et al., 2020), Sliding-Window Attention (Beltagy et al., 2020), and GLA (Yang et al., 2023), together with QORLA, QORLA-IDM, and QORLA-EIDM. Dense attention and FlashAttention retain quadratic attention arithmetic, but the quadratic auxiliary-space trajectory corresponds to explicit dense score materialization and should not be interpreted as FlashAttention memory usage. Base QORLA follows the exact $O(ND \log N)$ time and $O(ND)$ auxiliary-space path, whereas the prototype IDM variant is dominated by its quadratic routed anchor. QORLA-EIDM restores the exact-FE anchor while retaining innovation memory.

Table 15: **Primary survey completion accounting.** Time Series contains two datasets, hence twice the number of mechanism rows.

Domain	Rows	Finite / OK	Failed
Text	34	31	3
Vision	34	31	3
Multimodal	34	31	3
Genomics	34	33	1
Streaming	34	34	0
Time Series	68	66	2

E.1 Audit Registry and Protocol Separation

Several independent audits intentionally repeat related claims under different numerical regimes. To avoid apparent contradictions, Table 16 records the protocol identity of the headline checks. Values from different rows are replications under different harnesses and are not pooled.

E.2 Statistical-Mechanism Verification

Weighted local-linear identity and affine reproduction. Direct weighted-ridge least-squares solutions and closed-form Base+R QORLA agree to floating-point precision: the worst intercept error is 2.665×10^{-15}, relative ℓ_2 error 9.78×10^{-14}, and slope error 9.33×10^{-15}. Under increasingly one-sided affine support, Base error increases from 0.0397 to 0.7979, whereas Base+R remains within 3.934×10^{-15}; the affine identity $C = gQ$ holds within 6.174×10^{-16}. This numerically verifies the weighted-local-linear characterization and exact first-order bias cancellation derived above.

Second-order error and consistency. For smooth nonlinear signals, the fitted local-error exponent is 1.9716, close to the predicted quadratic behavior. With $N = 128, \ldots, 8192$ and bandwidth proportional to $N^{-1/5}$, Monte-Carlo MSE decreases from 5.98×10^{-4} to approximately 3.2×10^{-5}; the fitted tail exponent is -0.7652, consistent with movement toward the classical local-linear $N^{-4/5}$ rate. The rate is inherited from standard one-dimensional local-linear theory rather than claimed as QORLA-specific novelty.

Stability and structural checks. The maximum observed $\|C\|/\sqrt{QV}$ ratio is 1.00000000025 and the maximum correction-to-bound ratio is 0.9998598. Translation invariance, coordinate-scale equivariance, and non-causal permutation equivariance hold numerically within 5.94×10^{-15}, 1.17×10^{-14}, and 6.66×10^{-16}, respectively. For Base, the single-distractor identity holds within 3.82×10^{-16}; under decay rate 3, equivalently bandwidth $\lambda = 1/3$, the measured logarithmic decay slope is -2.99967, consistent with the theoretical slope $-1/\lambda = -3$.

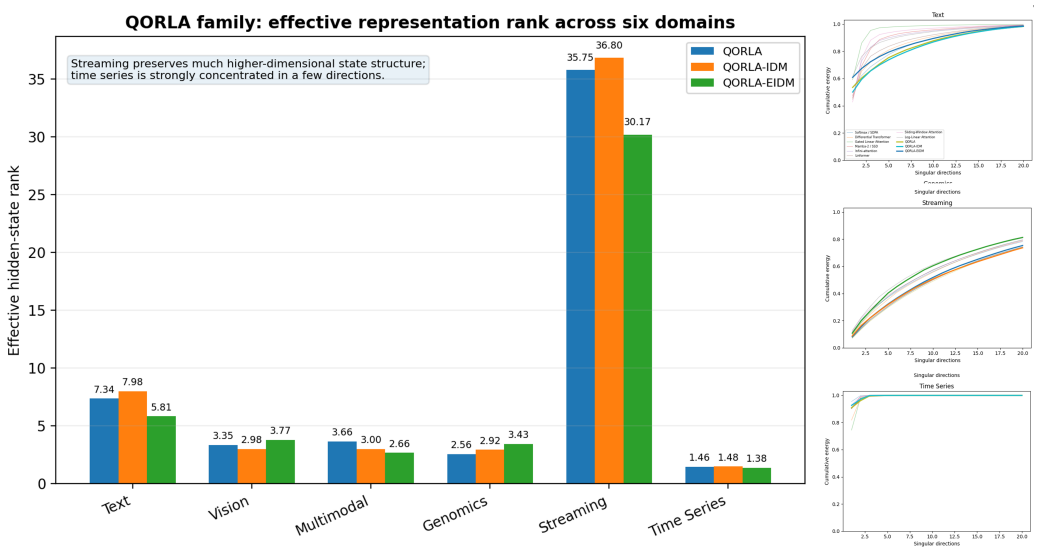


Figure 16: **Cross-domain representation geometry of the QORLA family.** Left: effective hidden-state rank for QORLA, QORLA-IDM, and QORLA-EIDM across all six domains. Right: representative cumulative singular-spectrum diagnostics for Text, Streaming, and Time Series taken from the same representation-atlas experiment. The strongly domain-dependent ranks show that QORLA's one-dimensional interaction geometry does not force globally low-rank hidden states: streaming representations retain effective ranks above 30, whereas time-series representations concentrate near rank 1.4–1.5.

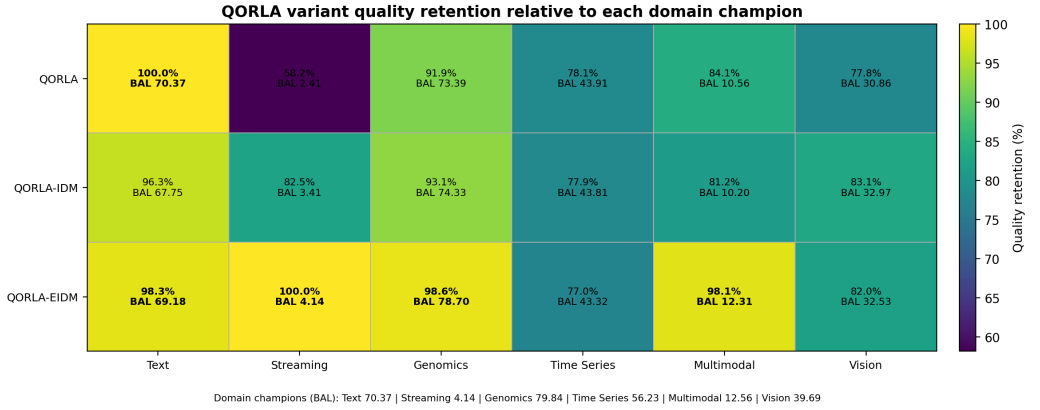


Figure 17: **Domain-wise QORLA variant ablation.** Each cell reports balanced-accuracy retention relative to the strongest successfully evaluated mechanism in that domain, with the corresponding raw BAL shown underneath. Base QORLA is strongest on Text, QORLA-EIDM is strongest within the family on Streaming, Genomics, and Multimodal fusion, while QORLA-IDM gives the strongest QORLA-family vision result. The heatmap therefore separates domains in which the statistical estimator alone is sufficient from those in which persistent innovation memory provides a measurable gain.

E.3 Exact Fast Evaluation and Gradient Equivalence

Dense Base+R and FE-QORLA were compared for $N = 16$–2048. In FP64 the maximum output error is 8.88×10^{-16} and maximum relative ℓ_2 error is 1.55×10^{-15}. The full projected gradient path—including inputs, W_Q, W_K, W_V, W_O, scalar query/key directions, measure parameters, λ, and ρ—has maximum relative discrepancy 1.39×10^{-14} and minimum cosine similarity 1.000000. This provides direct numerical support for exact forward evaluation and almost-everywhere dense–scan gradient equivalence.

Table 16: **Numerical-audit registry.** "Canonical main" denotes the audit used for the corresponding headline result; earlier audits are retained as independent replications.

Audit	Representative result	Role	Interpretation
Weighted local-linear identity	2.665×10^{-15} intercept error	Canonical mechanism audit	Closed-form and direct weighted-ridge solutions agree to roundoff.
Legacy FP64 dense–FE audit	8.88×10^{-16} forward max error	Independent replication	Small-scale FP64 replication; retained separately from the broader canonical mixed-key audit.
Canonical dense–FE forward audit	3.21×10^{-14} max error	Canonical main	Stress audit covering repeated keys, variable lengths, and zero-measure conditions.
Canonical dense–FE gradient audit	3.90×10^{-13} max error	Canonical main	Projected-gradient paths remain numerically equivalent.
Strict-online dynamic-tree audit	3.11×10^{-15} max error	Canonical main	Confirms exact causal evaluation without future-key pre-ranking.
Legacy T4, $N = 2048$	FE: 2.57 ms; dense: 8.99 ms	Independent systems replication	Earlier harness retained as an independent replication of the dense-to-exact latency crossover.
Canonical T4, $N = 2048$	FE: 4.02 ms; dense: 8.84 ms	Canonical main	Updated consolidated harness used for the main-paper T4 crossover claims.

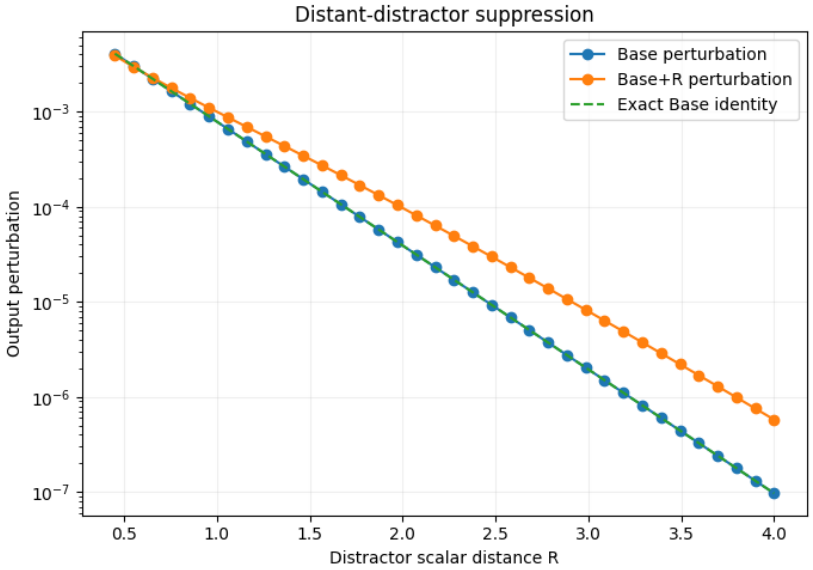


Figure 18: **Distant-token suppression diagnostic.** The influence of a distant Base perturbation decays exponentially with scalar-coordinate distance. Under decay rate 3, equivalently bandwidth $\lambda = 1/3$, the measured logarithmic slope is -2.9997, matching the theoretical value $-1/\lambda = -3$.

At $N = 2048$, FP32, FP16, and BF16 dense–FE maximum errors are 3.80×10^{-7}, 1.16×10^{-3}, and 1.17×10^{-2}, respectively; all tested outputs remain finite. Repeated keys, $a = b$, nearly degenerate Q, $\lambda = 10^{-4}$, $\lambda = 8$, and highly concentrated measures all remain finite, with worst FP64 dense–FE discrepancy approximately 9.48×10^{-14}.

E.4 Measured Systems Scaling

At $N = 2048$, dense-QORLA forward evaluation requires approximately 8.99 ms and 449.06 MB peak allocated memory, compared with 2.57 ms and 8.38 MB for FE-QORLA: approximately $3.49\times$ lower latency and $53.6\times$ lower incremental allocation in this implementation. Forward+backward measurements are 33.25 versus 6.93 ms and 835.06 versus 26.26 MB, respectively.

Power-law fits $T(N) \propto N^{\alpha}$ and $M(N) \propto N^{\beta}$ give FE forward exponents $(\alpha, \beta) = (0.796, 1.000)$, while dense QORLA gives $(1.706, 1.995)$; forward+backward memory exponents are 0.9996 and 1.992, respectively. FE-QORLA successfully evaluates $N = 32768$ on the tested T4 configuration. These measurements compare two realizations of the same operator and do not imply universal wall-clock superiority over optimized fused SDPA/FlashAttention kernels, whose hardware, dtype, and kernel paths differ.

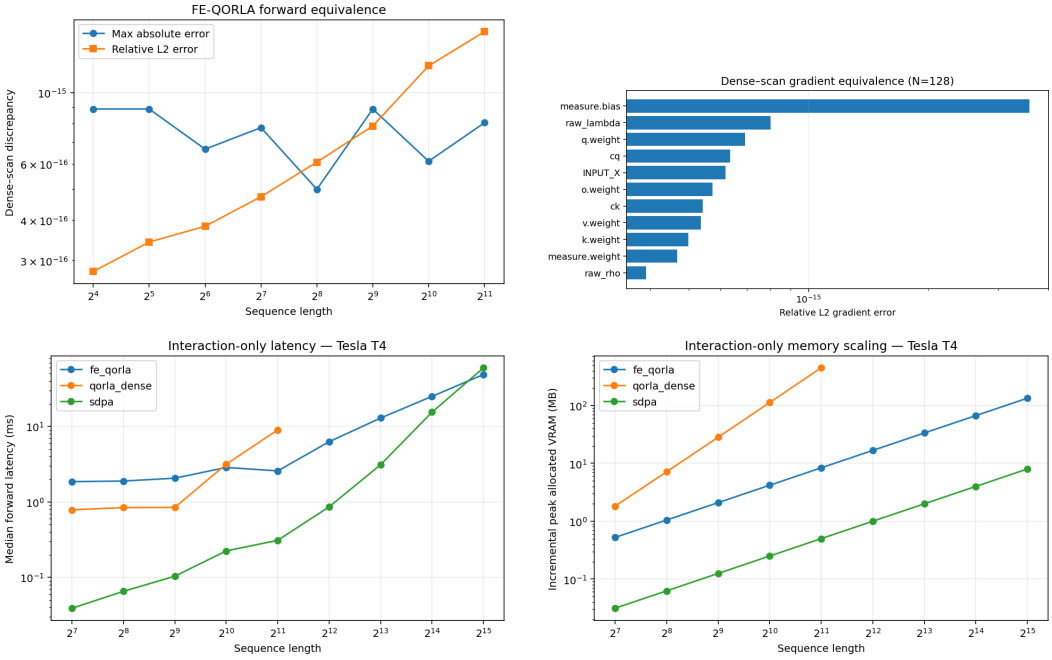


Figure 19: **Exact evaluation and systems audit.** Top: dense and FE implementations agree to numerical precision in outputs and gradients. Bottom: measured latency and allocated-memory scaling for dense and exact-FE QORLA under the legacy realization audit. These systems measurements are distinct from the canonical P4 T4 benchmark and the extended P15 stress harness and are not pooled with them.

Table 17: **Cross-GPU exact-evaluator systems profile.** The crossover column is the first tested N where exact FP32 latency is below dense-QORLA latency. Speedup and memory ratio at 8K compare exact with dense realizations of the same operator.

GPU	CC	VRAM	Cross.	Exact FP32 at 8K				F+B FP32 at 16K		Exact at 65K		Slope
		GiB		ms	M tok/s	speedup	mem. ratio	ms	M tok/s	FP32 ms	BF16 ms	
NVIDIA L4	8.9	22.03	2K	5.943	1.378	24.193×	85.009×	24.807	0.660	97.910	46.927	0.875
A100-SXM4-40GB	8.0	39.49	2K	7.726	1.060	4.193×	85.009×	30.205	0.542	115.046	113.247	0.970
Tesla T4	7.5	14.56	1K	12.988	0.631	11.458×	85.009×	55.722	0.294	100.289	100.530	0.982

E.5 Cross-GPU Portability of the Exact Evaluator

The exact evaluator is additionally profiled under the same $B = 1$, $D = 128$, $H = 4$, 66,316-parameter benchmark on NVIDIA L4, A100-SXM4-40GB, and Tesla T4 GPUs. FP32 exact evaluation is swept to $N = 65{,}536$; dense-reference timing is measured through $N = 8192$. The table also reports forward–backward behavior and the native BF16 path.

The FP64 dense/exact audit is identical to numerical precision on all three devices (forward maximum error 2.498×10^{-16}; gradient maximum error 4.441×10^{-15}). L4 gives the strongest measured 8K–32K operating points, while BF16 produces a particularly large 65K gain on L4 (97.91 → 46.93 ms) but little change on T4 and A100. These results support cross-device portability of the evaluator, not a universal ranking of GPU architectures. Two TPU/XLA benchmark attempts are excluded from systems claims because they terminate before a valid timing sweep with a `torch_xla` inference-tensor runtime error.

E.6 Extreme-Length Retrieval Diagnostics

Models were trained at $N = 128$ and evaluated through $N = 16384$ over three seeds. With 32 output values, chance accuracy is 3.125%. At the training context, Base QORLA/QORLA-EIDM obtain $11.59/13.15\%$ on MQAR, $9.62/8.71\%$ on MQMTAR, and $30.60/26.50\%$ on S-NIAH; corresponding Softmax results are 8.46%, 9.03%, and 24.87%. By $N = 16384$, all mechanisms approach chance. The diagnostic therefore shows successful short-context retrieval but does *not* establish clean 128-to-16K extrapolation.

A key confound is the use of learned absolute positional embeddings: positions beyond 127 are never optimized during training. Hence the extreme-length collapse cannot be attributed uniquely to the interaction operator. Copy, Reverse, and Selective-Copy controls are excluded from retrieval claims because their targets are visible at the evaluated positions.

Table 18: **Width scaling at $N = 8192$.** Median FP32 exact-evaluator latency in milliseconds.

GPU	$D = 128$	$D = 256$	$D = 512$
NVIDIA L4	5.946	6.487	9.786
A100-SXM4-40GB	7.731	8.226	17.544
Tesla T4	14.571	14.765	20.031

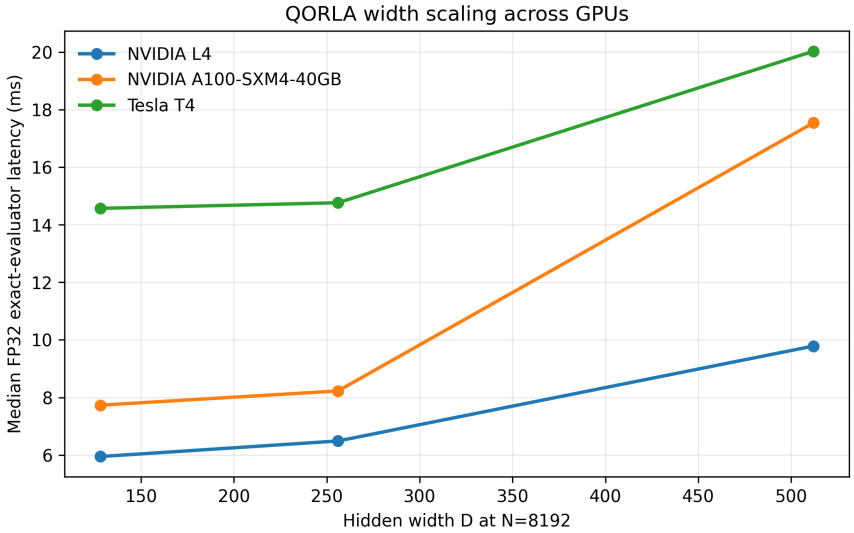


Figure 20: **Cross-GPU width scaling of exact QORLA evaluation.** Median FP32 exact-evaluator latency at fixed $N = 8192$ as hidden width increases across NVIDIA L4, A100-SXM4-40GB, and Tesla T4. The differing scaling trends show that realized performance depends on accelerator- and workload-specific execution characteristics rather than nominal peak throughput alone.

E.7 Innovation-Memory Principle

A controlled predictable-plus-innovation experiment isolates the motivation for storing residual rather than complete values. Averaged over predictor-noise levels, raw-value memory obtains MSE 0.437575, whereas innovation memory obtains 0.255234, a 41.7% reduction. The advantage persists from predictor noise 0 to 1.0.

Under increasing memory load, the difference persists: at 128 stored associations, raw memory gives MSE 0.4473 and final state norm 13.70, whereas innovation memory gives MSE 0.1870 and state norm 8.69. Thus, when the statistical anchor explains predictable content, residual writes reduce both error and redundant state magnitude in this controlled setting.

The learned causal-recall experiment is retained only as secondary evidence: full innovation reaches 4.42% accuracy versus 3.25% without memory and 3.89% with raw-value memory, but all results are close to the 3.125% chance level and evaluation at length 384 uses positions unseen during training.

E.8 Fixed-Step Component, Optimization, and Support Diagnostics

This diagnostic uses a deterministic AG News split of 12,000 training, 2,000 validation, and 2,000 test examples, sequence length $N = 128$, hidden width $D = 64$, FFN width 128, batch size 64, and 240 optimization steps. Training uses learning rate 3×10^{-4}, weight decay 10^{-2}, gradient clipping at 1.0, and seeds $\{11, 23, 47\}$. This fixed-step protocol is distinct from both the compact cross-domain survey and the full-data AG News confirmation.

Under this protocol, the three-seed mean balanced accuracies are 45.77% for Full QORLA, 45.56% for Base only, 46.00% for uniform measure, 46.71% for shared geometry, and 45.79% for fixed $\lambda = 1$. The differences are small and overlapping; accordingly, this experiment does not establish that each learned component independently improves downstream classification accuracy. The regression correction instead receives its strongest justification from the dedicated mathematical and local-bias analyses.

Optimization sensitivity is substantially larger than these component differences. Under the separate single-seed learning-rate sweep with 120 optimization steps, balanced accuracy increases from 28.65% at 10^{-4} to 38.86% at 3×10^{-4}, 50.58% at 10^{-3}, and 66.90% at 3×10^{-3}. Two gradient spikes but no non-finite loss are observed at the largest learning rate. These short-budget sensitivity results are therefore interpreted as optimization diagnostics rather than architectural rankings.

The learned interaction remains broad: $N_{\text{eff}}/N \approx 0.84$–$0.85$ with normalized entropy near 0.98 at trained-context scales. Thus "local" denotes proximity in learned scalar geometry rather than hard sparse sequence support. Mean absolute cross-head Spearman correlation is approximately 0.228 for query orderings and 0.123 for key orderings, demonstrating non-identical

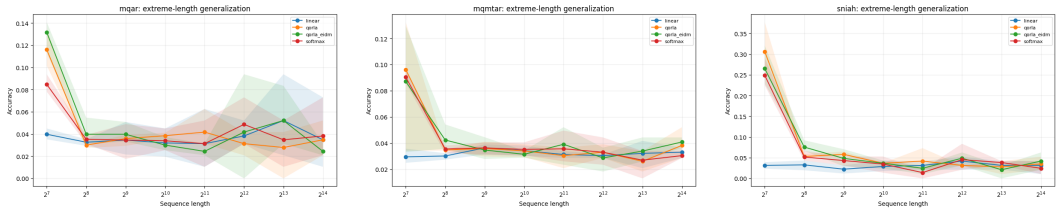


Figure 21: **Extreme-length diagnostic** MQAR, MQMTAR, and S-NIAH accuracy when models trained at length 128 are evaluated to 16384. The curves are retained as a transparent failure-mode diagnostic rather than evidence of successful extreme length extrapolation.

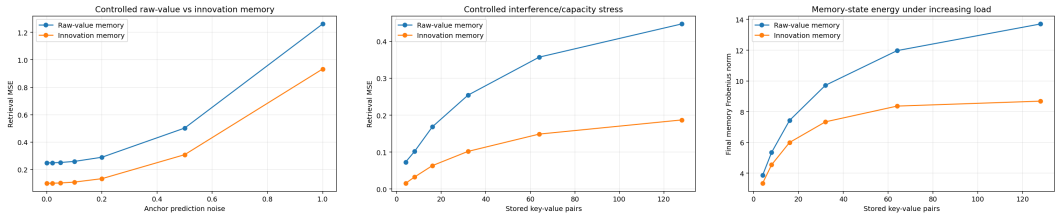


Figure 22: **Innovation-memory controls** Left: retrieval error versus statistical-predictor quality. Center: capacity/interference as stored associations increase. Right: persistent-state magnitude. Innovation writes reduce redundant memory content in the controlled setting.

head geometries. The largest sampled random-direction local gain is 0.8752; this is an empirical local diagnostic, not a certified global Lipschitz bound.

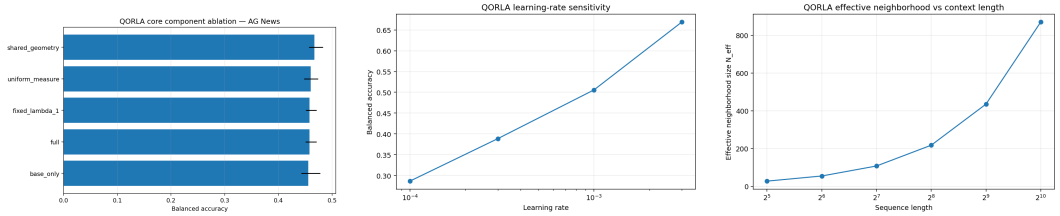


Figure 23: **Fixed-step component, optimization, and support diagnostics.** Left: three-seed AG News component diagnostic under the explicitly specified fixed-step protocol. Center: single-seed learning-rate sensitivity under the short sweep budget. Right: effective statistical-neighborhood size across context length.

E.9 Exploratory Independent-Layer Distillation Stress Test

Protocol status. This subsection is retained as an exploratory stress test in which layers are distilled independently and then composed. Its progressive-replacement and full-replacement numbers are *not* the canonical pretrained-conversion frontier used in the main paper. The canonical FP32 Gemma frontier and the adaptation-matched control are reported in Appendix I.6 and Appendix I.7; absolute PPL values across these protocols must not be mixed.

We test QORLA inside four *actual pretrained* open-weight decoder families: Gemma 3 270M, Qwen3-0.6B, Mistral-7B-v0.3, and Llama 3.2 1B. Tokenizers, embeddings, pretrained Q/K/V/O projections, rotary-position paths, normalization, residual streams, MLPs, and LM heads are preserved; only the attention interaction is replaced. QORLA parameters are first distilled against native attention outputs. All quality comparisons are within-model on identical token blocks; raw PPL is not compared across model families.

Matched single-layer transfer. Replacing one native attention layer yields PPL ratios 0.986, 1.048, 1.023, and 1.137 for Gemma, Qwen, Mistral, and Llama, with ΔNLL -0.014, $+0.047$, $+0.023$, and $+0.129$, respectively. Network-level hidden-state cosines remain 0.992, 0.990, 0.976, and 0.959. Across all four models, mean PPL ratio is 1.0485, median ratio is 1.0355, and mean hidden cosine is 0.9792; three of four models remain within $1.05\times$ native PPL and all four exceed 0.95 hidden-state cosine.

Table 19: **Matched single-layer QORLA transfer in real pretrained language models.** Native and converted variants use identical token blocks. PPL ratio is QORLA/native. Attn. cosine is mean held-out attention-output fidelity; hidden cosine is network-level fidelity.

Model	Native PPL	QORLA PPL	PPL ratio	ΔNLL	Hidden cos.	Attn. cos.
Gemma 3 270M	190.09	187.37	**0.986**	−0.014	**0.992**	0.692
Qwen3-0.6B	21.43	22.46	1.048	+0.047	0.990	0.448
Mistral-7B-v0.3	8.95	9.15	1.023	+0.023	0.976	0.437
Llama 3.2 1B	17.49	19.89	1.137	+0.129	0.959	0.461
Mean	–	–	1.048	–	0.979	–

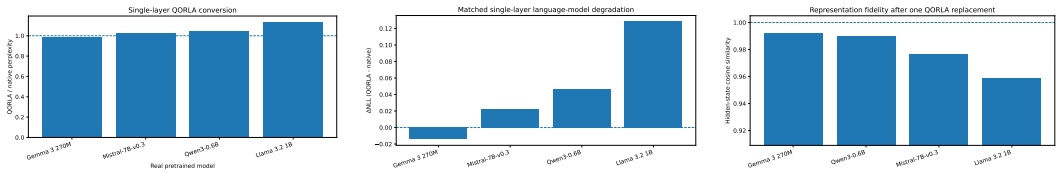


Figure 24: **Single-layer real-model transfer.** Left: QORLA/native PPL ratio. Center: matched ΔNLL. Right: hidden-state cosine after one native attention layer is replaced.

Progressive replacement. Single-layer fidelity does not imply compositional stability. When independently distilled QORLA layers are activated jointly, errors accumulate strongly with depth. Gemma remains comparatively stable through three layers ($1.258\times$) but reaches $9.652\times$ at full replacement. Llama grows from $1.137\times$ to $2.810\times$, $4.332\times$, $18.460\times$, and $280.144\times$. Qwen and Mistral are considerably more sensitive, reaching full-replacement ratios of $11031.430\times$ and $498.904\times$, respectively.

The corresponding hidden-state cosines show the same effect: Gemma decreases from 0.992 to 0.788, Qwen from 0.990 to 0.377, Mistral from 0.976 to 0.258, and Llama from 0.959 to 0.378. This separates *local operator fidelity* from *global compositional stability*: downstream converted layers receive student-generated representations rather than the native representations on which they were independently distilled.

Layerwise approximation. Mean held-out native/QORLA attention-output cosine is 0.692 for Gemma, 0.448 for Qwen, 0.437 for Mistral, and 0.461 for Llama. These direct operator-level similarities are distinct from the much higher single-replacement network-level hidden cosines: the latter include the stabilizing effect of the surrounding pretrained residual network.

Legacy full-replacement adaptation stress test. For completed 100% replacement evaluations, a short QORLA-only network-level adaptation stage does not recover native quality. Gemma ends at PPL ratio 15.944 with hidden cosine 0.779; Llama reaches 248.726 with cosine 0.394. These final evaluations use the larger full-evaluation set and must not be mixed numerically with the four-block matched baselines above. Qwen and Mistral are omitted because their joint-adaptation evaluations did not complete successfully.

Scope. The real-checkpoint evidence supports *local pretrained-layer convertibility*: selected individual attention layers can be replaced across four distinct decoder families while preserving high network-level fidelity. It does *not* establish universal full-network attention replacement. The multi-layer failures instead motivate progressive network-aware conversion, in which later QORLA layers are adapted to representations produced by already converted upstream layers. These experiments use dense QORLA to isolate pretrained-model compatibility; exact fast-evaluation complexity and systems results are reported separately.

E.10 Excluded Experiments and Reproducibility Notes

For transparency, invalid or confounded evidence is excluded from the corresponding claims. Copy, Reverse, and Selective-Copy retrieval controls contain target leakage and are therefore not used to support retrieval claims; the extreme-length evaluations additionally rely on untrained absolute positions. The forced-age innovation-memory sweep is excluded because sorting can alter the intended target-pair index, while an incompatible stress-row CSV schema invalidates the associated original aggregation. The modern Gemma probe is excluded because both native and replacement perplexities become non-finite. The forced Flash-SDPA kernel was unavailable under the tested Tesla T4 software configuration and is therefore not treated as a performance measurement. Non-finite proxy-baseline evaluations characterize the tested compact implementations and optimization budgets rather than intrinsic instability of the corresponding published methods.

The mathematical-identity experiments are performed in CPU/FP64, whereas the primary exactness and systems audit uses a Tesla T4 with PyTorch 2.10 and CUDA 12.8. Multi-seed results are explicitly identified throughout the appendix; single-seed hyperparameter sweeps are interpreted only as sensitivity diagnostics rather than statistical rankings. The 36-entry controlled mechanism comparison is reported in Appendix B and is not duplicated here.

Table 20: **Progressive replacement by independently distilled QORLA operators.** Entries are matched QORLA/native PPL ratios.

Model	1 layer	3 layers	25%	50%	100%
Gemma 3 270M	0.986	1.258	2.852	3.215	9.652
Qwen3-0.6B	1.048	971.793	866.200	6477.411	11031.430
Mistral-7B-v0.3	1.023	74.667	92.341	137.520	498.904
Llama 3.2 1B	1.137	2.810	4.332	18.460	280.144

Table 21: **Network-level hidden-state cosine under progressive QORLA replacement.**

Model	1 layer	3 layers	25%	50%	100%
Gemma 3 270M	0.992	0.908	0.900	0.828	0.788
Qwen3-0.6B	0.990	0.489	0.500	0.446	0.377
Mistral-7B-v0.3	0.976	0.321	0.307	0.285	0.258
Llama 3.2 1B	0.959	0.683	0.610	0.483	0.378

F Confirmatory Evidence and Main-Claim Closure

The controlled 36-mechanism survey in Appendix B is intended to expose broad cross-domain behavior under a common compact wrapper. We therefore do not reproduce that table here. Instead, this final appendix section consolidates the higher-budget experiments performed after the broad survey: multi-seed full-data classification, numerical identity and gradient audits, strict-online evaluation, long-context systems scaling, statistically powered 8K–16K quality tests, reference-mixer causal language modeling, real-task innovation-memory ablations, matched 2B-token pretraining, and pretrained-network conversion. These experiments are designed to distinguish three types of statement: (i) algebraic or numerical properties that are directly verified, (ii) empirical advantages that hold only for the measured configuration, and (iii) hypotheses that remain deliberately unresolved.

F.1 Multi-Seed Full-Data Classification

The initial broad survey uses a deliberately small fixed training budget and is best interpreted as a mechanism survey. We therefore repeated the text experiment on the full AG News training set with a larger $D = 128$, $H = 4$, four-layer matched encoder and five independent confirmation seeds. Table 23 gives the resulting multi-seed comparison.

The principal conclusion of this experiment is therefore *not* that QORLA dominates dense attention on classification. QORLA and SDPA are statistically unresolved under the paired five-seed comparison, whereas QORLA shows consistent positive differences over the matched Linformer and GLA implementations. The mean test NLL is 0.2881 for QORLA and 0.2754 for SDPA. At $N = 128$, measured per-example latency is 0.2037 ms for QORLA, 0.1167 ms for SDPA, 0.1300 ms for Linformer, and 0.6557 ms for the GLA proxy. Peak measured memory is approximately 390 MB, 270 MB, 290 MB, and 276 MB, respectively. QORLA contains 5,005,620 trainable parameters versus 5,002,500 for SDPA, an increase of only 3,120 parameters ($\approx 0.062\%$).

Interpretation. Across five confirmation seeds, QORLA obtains a mean balanced accuracy of 91.2342% compared with 91.2553% for SDPA, 90.7842% for Linformer, and 90.8474% for the tested GLA proxy. The paired QORLA–SDPA difference is only -0.0211 percentage points, with a 95% bootstrap interval of $[-0.155, +0.124]$, so the experiment does not resolve a classification-quality difference between QORLA and SDPA. In contrast, QORLA exceeds Linformer by $+0.4500$ percentage points (95% interval $[+0.245, +0.626]$) and the GLA proxy by $+0.3868$ percentage points ($[+0.316, +0.466]$), with QORLA higher on all five confirmation seeds in both comparisons. The efficiency panel also makes the implementation trade-off explicit: at $N = 128$, measured per-example latency is approximately 0.2037 ms for QORLA, 0.1167 ms for SDPA, 0.1300 ms for Linformer, and 0.6557 ms for the tested GLA proxy. These short-context measurements are therefore reported separately from the long-context FE-QORLA scaling results.

F.2 Component Diagnostics and Statistical-Mechanism Evidence

The downstream component diagnostic is reported in Appendix E.8 under its explicitly specified fixed-step AG News protocol. Full, Base-only, uniform-measure, shared-geometry, and fixed-λ variants exhibit small, overlapping differences; this study is therefore interpreted as a component-sensitivity diagnostic rather than evidence that every learned component independently improves classification.

The correction R has stronger direct mechanistic support. In the affine experiment, mean Base error is 0.157205 versus 1.61×10^{-12} for Base+R. Independent weighted-local-linear verification gives worst intercept, relative ℓ_2, and slope discrepancies of 2.665×10^{-15}, 9.78×10^{-14}, and 9.33×10^{-15}, respectively. Under increasing support asymmetry, Base error grows from 0.0397 to 0.7979, while Base+R remains below 3.934×10^{-15}; $C = gQ$ holds to 6.174×10^{-16}. The fitted nonlinear local-error exponent is 1.9716, and Base+R has lower MSE than Base across all 25 ridge/noise conditions, with ratios approximately 0.22–0.79.

In this separate $N = 128$ mechanism audit, the learned scalar geometry is more concentrated than in the fixed-step support diagnostic: $N_{\text{eff}} = 28.50$ gives $N_{\text{eff}}/N = 0.2227$, with normalized interaction-weight entropy 0.7655 and measure entropy 0.8675. Mean absolute cross-head Spearman correlation is 0.1914 for query coordinates and 0.2305 for key coordinates,

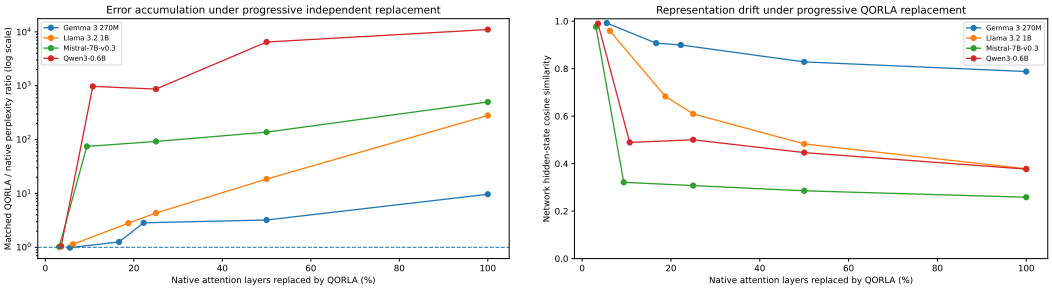


Figure 25: **Compositional stress test.** Left: matched PPL ratio versus fraction of converted layers (logarithmic ordinate). Right: corresponding hidden-state cosine. Independent substitutions remain locally accurate but accumulate large errors when composed across depth.

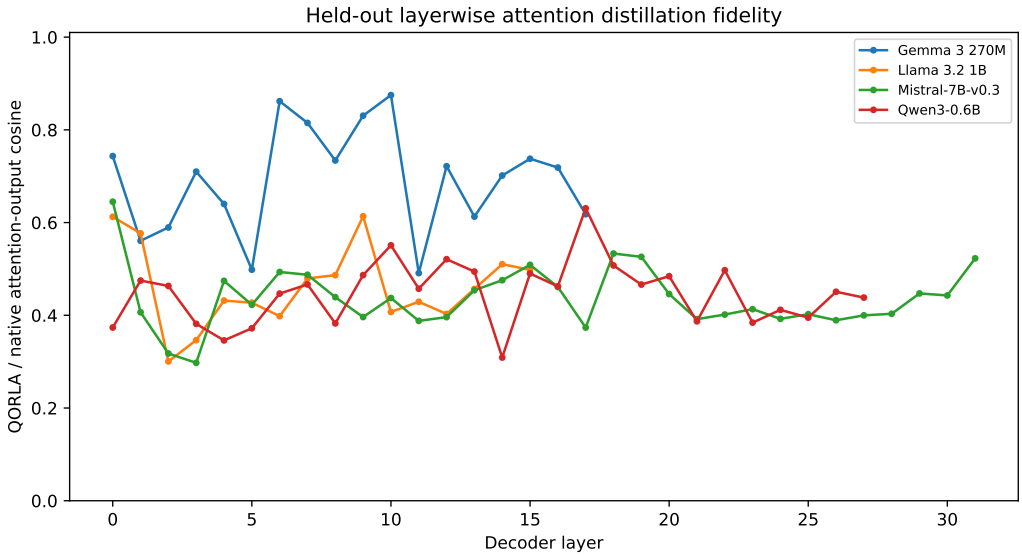


Figure 26: **Layerwise attention-output fidelity.** Held-out cosine between native pretrained attention outputs and independently distilled QORLA approximations across decoder depth.

indicating distinct head geometries without single-token collapse. These quantities are protocol-specific and are not pooled with the broader-support diagnostics in Appendix E.8.

F.3 Exact Evaluation, Gradients, and Strict-Online Causality

The strongest implementation claim is algebraic exactness rather than approximate linearization. The Laplacian kernel permits the required statistics

$$1, \qquad b, \qquad b^2, \qquad v, \qquad bv$$

to be recovered from ordered prefix/suffix aggregates. Table 25 consolidates the independent numerical audits.

The strict-online experiment is particularly important. Unlike the earlier Fenwick implementation, the dynamic ordered tree does not construct coordinate ranks from future keys. Each token inserts its sufficient-statistic payload at the observed scalar coordinate and queries the ordered prefix online. Consequently, causal Base+R can be evaluated exactly in expected $O(ND \log N)$ time and $O(ND)$ maintained tree state without future-coordinate access.

Interpretation. Panel (a) summarizes the numerical-identity evidence underlying the exact evaluator claim: dense and FE realizations agree to floating-point precision in both forward outputs and gradients, so the FE implementation should be interpreted as an algebraically exact realization rather than an approximation. In the systems comparison, FE-QORLA first overtakes the dense QORLA realization by $N = 2048$ and later overtakes the tested optimized SDPA latency by $N = 8192$. At $N = 16384$, FE-QORLA requires approximately 25.20 ms versus 64.31 ms for SDPA, giving a measured $2.55\times$ latency advantage in this hardware/software setting. The memory panel should be interpreted separately: FE-QORLA removes the quadratic allocation growth of dense QORLA, but the tested optimized SDPA path still uses less realized memory at long context (approximately 8 MB versus 91 MB at $N = 16384$). Accordingly, the supported claim is an exact long-context evaluator with favorable measured scaling, not universal memory superiority over optimized SDPA.

Table 22: **Completed full-replacement plus short QORLA-only adaptation evaluations.**

Model	Native PPL	Final QORLA PPL	PPL ratio	ΔNLL	Hidden cos.	Trainable (%)
Gemma 3 270M	239.43	3817.61	15.944	+2.769	0.779	0.0207
Llama 3.2 1B	20.53	5107.43	248.726	+5.516	0.394	0.0081

Table 23: **Full-data multi-seed AG News confirmation.** BAL is balanced accuracy. Intervals in the final column are paired bootstrap intervals for QORLA minus the corresponding baseline. GLA denotes the compact matched-wrapper GLA proxy used in this classification experiment; the separate causal-LM study below uses the maintained FLA reference mixer.

Mechanism	Mean BAL (%)	SD (pp)	QORLA–baseline ΔBAL (pp)	Paired 95% interval (pp)	QORLA wins	Remarks
QORLA	91.2342	0.0932	–	–	–	Five seeds
SDPA	91.2553	0.0835	-0.0211	$[-0.155, +0.124]$	2/5	statistically unresolved difference
Linformer	90.7842	–	+0.4500	$[+0.245, +0.626]$	5/5	matched compact baseline
GLA proxy	90.8474	–	+0.3868	$[+0.316, +0.466]$	5/5	matched compact baseline

F.4 Complexity Contract and Measured Systems Crossover

The complexity claim applies to the scan-compatible Base+R operator and must not be transferred indiscriminately to every memory extension. Table 26 gives the final complexity contract.

The measured systems audit exhibits two different comparisons that should not be conflated. First, when comparing two realizations of *the same QORLA operator*, FE evaluation crosses the dense realization by $N = 2048$: 8.84 ms for dense QORLA versus 4.02 ms for FE-QORLA, a $2.20\times$ speedup. Incremental allocated memory at this length is 321.1 MB for the dense realization and 11.38 MB for FE, approximately a $28.2\times$ reduction.

Second, against framework-optimized SDPA on the tested Tesla T4, the first measured FE-QORLA latency crossover occurs at $N = 8192$ (14.96 ms versus 16.52 ms). At $N = 16384$, FE-QORLA requires 25.20 ms versus 64.31 ms for SDPA, corresponding to a measured $2.55\times$ latency advantage. The finite-range fitted latency exponents are approximately 0.785 for FE-QORLA and 1.635 for SDPA; these empirical slopes are not substituted for the asymptotic complexity theorem.

The memory comparison against optimized SDPA is less favorable: at $N = 16384$, measured incremental allocation is approximately 91 MB for FE-QORLA versus 8 MB for the tested SDPA path. Thus the exact evaluator removes the quadratic memory behavior of *dense QORLA*, but the current implementation does not establish lower realized memory than highly optimized SDPA.

F.5 Statistically Powered 8K–16K Long-Context Quality

The original extreme-length diagnostic contained too few examples at the largest context lengths to support a quality comparison. We therefore repeated the controlled position-free regression task with three seeds and 32 paired test problems per seed at each length, producing 96 paired examples at $N = 8192$ and another 96 at $N = 16384$. Uncertainty is estimated by a hierarchical bootstrap that resamples seeds and then examples within seed.

Dense-attention scope control. A separate powered comparison evaluates QORLA directly against SDPA at $N \in \{8192, 16384\}$ using 96 paired test problems per context length. SDPA attains lower MSE in this controlled regression setting, with the hierarchical confidence intervals excluding zero at both lengths (Table 27). This predictive-quality comparison is intentionally separated from the exact-evaluation systems results: the latter establish algebraic equivalence between dense and ordered QORLA and measure the realized long-context latency scaling of the exact evaluator.

Interpretation. At $N = 8192$, QORLA obtains MSE 0.08088 compared with 0.03278 for SDPA, giving a paired difference of $\Delta\mathrm{MSE} = +0.04810$ with 95% hierarchical-bootstrap interval $[+0.00754, +0.09335]$; QORLA has lower error on 39 of the 96 paired examples. At $N = 16384$, QORLA obtains MSE 0.09299 versus 0.03123 for SDPA, with $\Delta\mathrm{MSE} = +0.06176$ and 95% interval $[+0.03100, +0.10257]$; QORLA has lower error on 30 of 96 examples. Thus the powered experiment resolves the earlier underpowered long-context quality comparison in favor of SDPA on this controlled regression task. This result does not contradict the systems-scaling experiment: FE-QORLA retains its measured long-context latency crossover, but that computational advantage does not imply superior predictive quality.

F.6 Reference-Mixer Causal Language Modeling

We next evaluate QORLA in a real causal language-modeling setting using WikiText-103 and a GPT-2 tokenizer. All mechanisms use the same matched four-layer language-model scaffold with $D = 128$, four heads and sequence length 256. Learning rate is selected from the same grid and all four mechanisms select 10^{-3}. Confirmation uses seeds $\{23, 37, 53\}$ and 500 optimization steps, corresponding to 1,024,000 observed training tokens per confirmation evaluation.

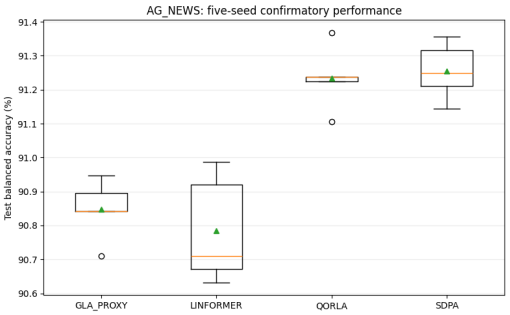
Unlike the broad survey, the external efficient mechanisms here use maintained reference token mixers: PyTorch SDPA, FLA `GatedLinearAttention`, and Hugging Face `Mamba2Mixer`. This is a matched-*mixer* comparison rather than an attempted reproduction of each authors' full large-scale architecture and training recipe.

The seed-paired QORLA–SDPA NLL difference is

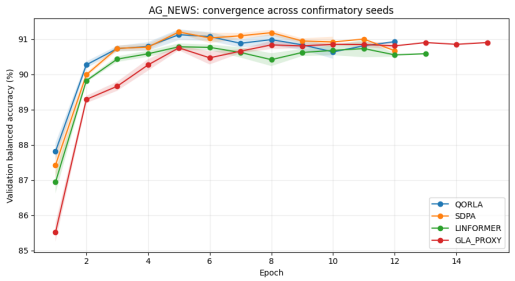
$$+0.08785, \qquad 95\%\ \text{bootstrap interval} = [+0.08389, +0.09251],$$

so SDPA has the better short-context LM likelihood in this experiment. Relative to FLA GLA, the QORLA difference is

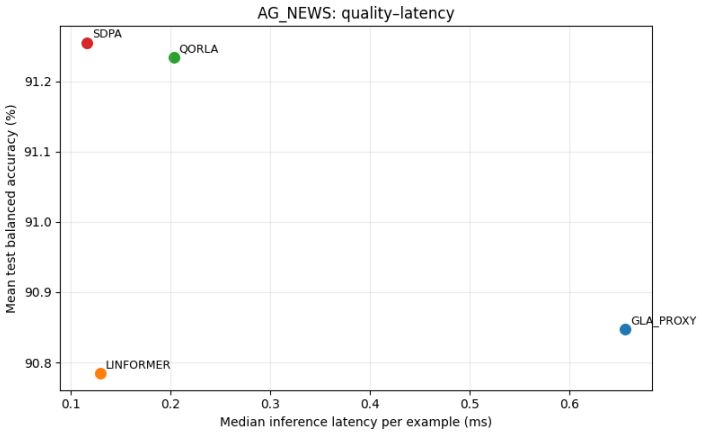
$$-0.16632, \qquad [-0.20422, -0.13438],$$



(a) Five-seed test balanced-accuracy distribution



(b) Validation BAL convergence with seed-level uncertainty



(c) Test balanced accuracy versus measured inference latency

Figure 27: **Full-data multi-seed AG News confirmation.** Five-seed predictive stability, convergence behavior, and the measured quality–latency trade-off under the matched experimental protocol.

Table 24: **Three-seed QORLA component ablation on AG News.** Differences are small; no individual ablation establishes a statistically resolved degradation relative to Full QORLA.

Variant	Mean BAL (%)
Base only	89.3816
No ridge	89.3202
Uniform measure	89.2763
Full QORLA	89.2544
Fixed λ	89.1930
Shared head geometry	89.1711
Shared query/key geometry	89.1228

while relative to Mamba-2 it is

$$+0.06918, \qquad [+0.04926, +0.09556].$$

The direction of all three comparisons is consistent across all three confirmation seeds.

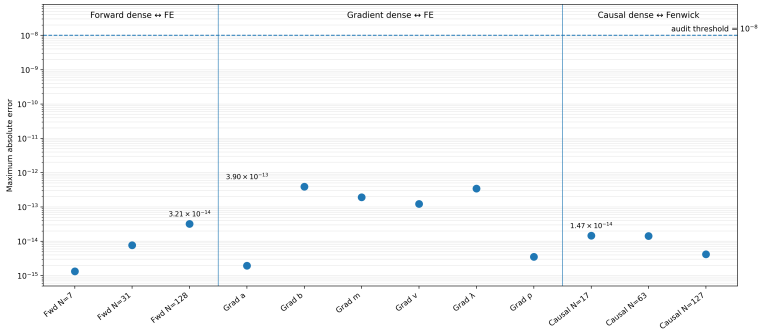
The systems columns require two implementation qualifications. First, each FLA confirmation seed exhibits one recoverable FP16 gradient-overflow event; dynamic loss scaling reduces the scale from 65,536 to 32,768 and skips the affected optimizer update without producing a non-finite model or loss. Second, the optimized Mamba-2 fast path is unavailable in the tested software environment, so Hugging Face falls back to its slower reference execution. Accordingly, the Mamba throughput number is *not* used to claim a general QORLA speed advantage over optimized Mamba-2.

Interpretation. SDPA obtains the lowest mean test NLL, 6.0654 ± 0.0087 (PPL 430.70), followed by the Hugging Face Mamba-2 reference mixer at 6.0841 ± 0.0184 (PPL 438.86), QORLA at 6.1532 ± 0.0113 (PPL 470.26), and FLA GLA at 6.3196 ± 0.0395 (PPL 555.62). The seed-paired QORLA–SDPA difference is $+0.08785$ NLL with bootstrap interval $[+0.08389, +0.09251]$, whereas QORLA–FLA GLA is -0.16632 with interval $[-0.20422, -0.13438]$. Relative to Mamba-2, QORLA differs by $+0.06918$ NLL with interval $[+0.04926, +0.09556]$. The direction of all three comparisons is consistent across the three confirmation seeds.

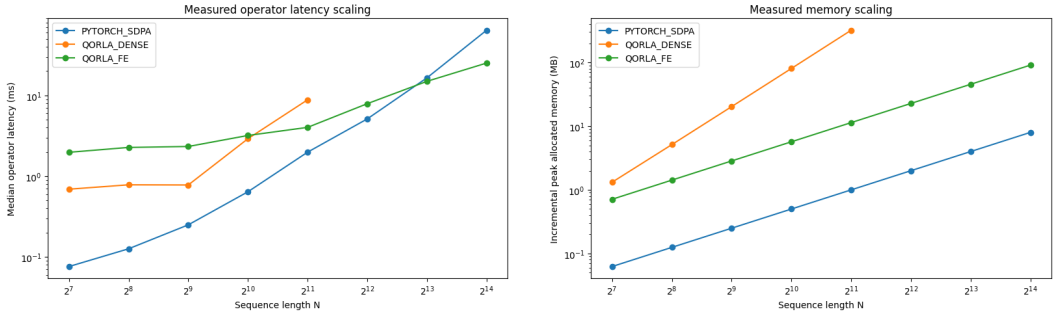
Measured training throughput is approximately 43,700 tokens/s for SDPA, 30,632 for QORLA, 27,212 for FLA GLA, and 14,700 for the tested Mamba-2 path. Peak allocated memory is approximately 1586.9 MB, 1855.4 MB, 1912.2 MB, and 2021.0 MB, respectively. These systems numbers are implementation-specific. In particular, the optimized Mamba-2 fast

Table 25: **Exact-evaluation numerical audit.** All discrepancies compare algebraically equivalent dense and ordered implementations; they measure floating-point error rather than approximation error.

Audit	Worst discrepancy	Tested regime	Interpretation
Dense vs. FE Base+R forward	3.21×10^{-14} max abs.	repeated keys, variable lengths, zero-measure padding	exact to floating-point precision
Dense vs. FE gradients	3.90×10^{-13} max abs.	$a, b, m, v, \lambda, \rho$	differentiable dense/scan agreement
Dense causal vs. Fenwick evaluator	1.47×10^{-14} max abs.	tied scalar keys included	exact causal aggregation after coordinate compression
Dense causal vs. strict-online dynamic treap	3.11×10^{-15} max abs.	no future-key pre-ranking	closes the strict-online causality caveat
FP32 dense vs. FE	3.80×10^{-7} max abs.	$N = 2048$	finite in all tested cases
FP16 dense vs. FE	1.16×10^{-3} max abs.	$N = 2048$	finite mixed-precision execution
BF16 dense vs. FE	1.17×10^{-2} max abs.	$N = 2048$	finite mixed-precision execution



(a) Forward and gradient exactness audit



(b) Sequence-length versus latency

(c) Sequence-length versus memory

Figure 28: **Exact evaluator audit and systems scaling.** Top: dense and exact sufficient-statistic realizations agree to numerical precision. Bottom: measured latency and memory scaling of dense QORLA, FE-QORLA, and the tested optimized SDPA implementation on Tesla T4.

path was unavailable in the tested environment, so its throughput should not be interpreted as a general comparison with optimized Mamba-2 kernels. The experiment therefore supports a quality comparison between maintained reference mixers while keeping hardware-dependent systems conclusions explicitly scoped.

F.7 Matched 2B-Token Language-Model Pretraining

We evaluate QORLA, Transformer++, Mamba-2, and Gated DeltaNet under a matched large-token-budget protocol on FineWeb-Edu and SlimPajama. Each run processes exactly 2,000,027,648 prediction tokens and uses an approximately 10.2M-parameter model. The experiment is designed to separate final language-model likelihood from realized training compute.

Interpretation. Across both corpora, QORLA sustains approximately 144k train tokens/s and 7.72 T4 GPU-hours, corresponding to 1.37–1.39× Gated DeltaNet throughput and 2.20–2.26× Mamba-2 throughput, with roughly 27–28% and 54–56% lower GPU-hour consumption, respectively. The quality result is deliberately separated from the systems result: Transformer++ and Mamba-2 obtain lower final perplexity, and Transformer++ is also the fastest implementation. The supported conclusion is therefore a stable compute-efficiency advantage over the tested structured alternatives, not universal likelihood or throughput superiority.

F.8 Innovation Memory on a Real Text Task

Synthetic memory probes show that innovation writing is predictor-dependent: when the statistical predictor is poor, residual memory need not outperform raw-value memory. The real-task experiment therefore evaluates Base QORLA, raw IDM,

Table 26: **Final complexity contract.** N is sequence length, D total hidden width, H number of heads and $d = D/H$. Standard learned projections are excluded from the Base+R scan bound, as in the main theorem.

Mechanism / realization	Sequence-dependent time	Auxiliary / maintained state	Status
Dense Base+R	$O(N^2 D)$	$O(N^2 + ND)$	dense reference
Exact non-causal FE Base+R	$O(ND \log N)$	$O(ND)$	exact
Strict-online causal Base+R	expected $O(ND \log N)$	$O(ND)$	exact dynamic ordered tree
Dense-anchor IDM	$O(N^2 D + ND^2/H)$	$O(N^2 + ND + D^2/H)$	persistent full-rank memory
FE-anchor EIDM	$O(ND \log N + ND^2/H)$	$O(ND + D^2/H)$	exact FE anchor + memory

Table 27: **Powered long-context comparison with dense SDPA.** ΔMSE denotes QORLA minus SDPA; positive values indicate lower SDPA error under this controlled regression protocol.

N	QORLA MSE	SDPA MSE	ΔMSE	95% hier. CI	QORLA lower
8192	0.08088	0.03278	+0.04810	$[+0.00754, +0.09335]$	39/96
16384	0.09299	0.03123	+0.06176	$[+0.03100, +0.10257]$	30/96

EIDM, and confidence-gated EIDM on IMDB under a common model and optimization protocol. Confirmation uses three seeds.

The EIDM result is notable because only 12 trainable parameters are added relative to Base in this implementation. It gives the largest mean balanced-accuracy gain, while the persistent-state norm is lower than raw IDM. This should be interpreted as evidence that innovation writing can reduce *persistent redundant state*, not that every innovation vector must have lower instantaneous norm: the measured raw write norm is 1.7919 and the EIDM innovation norm is 1.8273.

The confidence gate is directionally meaningful but does not improve the three-seed mean over ordinary EIDM. Mean gate value is 0.9452 and its correlation with prediction error is approximately -0.1053, indicating that the learned model generally prefers almost-pure innovation writing while slightly reducing it as predictor error increases. We therefore retain plain EIDM as the cleaner primary memory variant rather than claiming a universal benefit from confidence gating.

Interpretation. Base QORLA obtains mean balanced accuracy 0.6878 ± 0.0338, while Raw-IDM reaches 0.7160 ± 0.0161, EIDM reaches 0.7372 ± 0.0047, and confidence-gated EIDM reaches 0.7325 ± 0.0046. Relative to Base, the corresponding mean paired BAL improvements are $+0.0283$, $+0.0494$, and $+0.0447$, respectively, and every memory variant improves Base on all three confirmation seeds. EIDM gives the largest average gain while adding only 12 trainable parameters relative to the 4,937,102-parameter Base model.

The memory-state diagnostic provides complementary mechanistic evidence. Mean final persistent-state norm is 0.1300 for Raw-IDM, 0.1222 for EIDM, and 0.1091 for CG-EIDM. Thus innovation-based writing reduces persistent-state magnitude relative to raw-value memory in this experiment. This should not be interpreted as saying that every innovation vector has smaller instantaneous norm: the measured mean raw-write norm for EIDM is 1.7919, whereas the corresponding innovation norm is 1.8273. The supported conclusion is therefore reduced persistent-state magnitude together with improved downstream accuracy, rather than a universal low-energy-write claim.

F.9 Sensitivity-Aware Pretrained Conversion (Secondary Protocol)

Protocol status. The cross-family single-layer measurements below are canonical and are shared with the main paper. The progressive Gemma frontier in this subsection, however, is a *secondary sensitivity-aware joint-adaptation protocol*. The main-paper depth-wise frontier uses the fixed FP32 canonical protocol in Appendix I.6. Because evaluation slices and adaptation procedures differ, absolute PPL ratios from the two frontiers must not be interchanged.

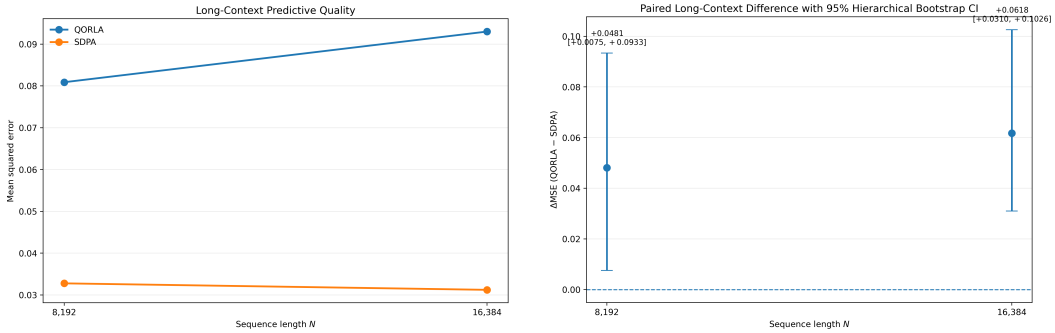
A first transfer audit replaces a selected attention layer in four real decoder-family checkpoints. Table 32 shows that single-layer replacement is locally compatible across all four tested families.

The more difficult question is compositional replacement. Blind progressive conversion causes rapidly accumulating error, motivating a second experiment that first ranks Gemma's 18 layers by single-layer sensitivity and then converts the least sensitive layers progressively with joint network adaptation. The six initially selected compatible layers are $\{9, 12, 14, 15, 16, 17\}$ after depth-ordering. The final experiment continues the procedure through 9, 12, and all 18 layers.

At complete $18/18$ conversion, joint adaptation substantially improves the unadapted converted network: PPL ratio decreases from 2.887 to 1.621, final hidden cosine increases from 0.650 to 0.713, and token KL decreases from 4.213 to 3.207. Nevertheless, complete replacement remains far from teacher-faithful. The correct conclusion is therefore a *conversion frontier*: sensitivity-aware partial replacement can remain stable and task-effective, while unrestricted full-network equivalence is not established.

The sub-unity PPL ratios at 6–12 converted layers should likewise not be described as teacher reconstruction. At 9 converted layers, for example, PPL ratio is 0.555 while hidden cosine is only 0.813 and token KL is 1.698. The adapted model has found a different likelihood-improving solution rather than reproducing the internal computation of the native teacher.

Interpretation. The single-layer scan reveals substantial heterogeneity in conversion sensitivity across decoder depth. The six lowest-sensitivity layers are 14, 16, 15, 12, 9, and 17, with composite sensitivity scores of approximately 0.0474, 0.0834, 0.1116, 0.1606, 0.1856, and 0.2156, respectively. By contrast, layers 1 and 4 are substantially more sensitive, with scores 0.8978 and 1.1745. This motivates sensitivity-aware replacement rather than uniform conversion in layer order.



(a) Mean squared error at 8K and 16K context

(b) Paired QORLA–SDPA MSE difference with 95% CI

Figure 29: **Powered long-context predictive-quality evaluation.** Each context length contains 96 paired test problems (32 examples × three seeds). Panel (a) reports absolute mean squared error, while panel (b) reports the paired QORLA–SDPA difference with hierarchical-bootstrap uncertainty.

Table 28: **Three-seed causal-LM confirmation on WikiText-103.** NLL and PPL are test metrics. Throughput and peak allocated memory are implementation- and hardware-specific and should not be interpreted as universal kernel rankings.

Mechanism	Test NLL ↓	PPL ↓	Train throughput (tok/s) ↑	Peak alloc. (MB)	Parameters
SDPA	$\mathbf{6.0654 \pm 0.0087}$	$\mathbf{430.70}$	$\mathbf{43{,}700}$	$\mathbf{1586.9}$	7,224,192
HF Mamba-2	6.0841 ± 0.0184	438.86	14,700	2021.0	7,431,600
QORLA	6.1532 ± 0.0113	470.26	30,632	1855.4	7,227,312
FLA GLA	6.3196 ± 0.0395	555.62	27,212	1912.2	7,236,864

Joint adaptation substantially improves likelihood throughout the progressive conversion frontier. At 6/18 converted layers, the QORLA/native PPL ratio falls from 1.8761 before adaptation to 0.6188 afterward; at 9/18 it falls from 1.5739 to 0.5546; at 12/18 from 1.0524 to 0.7644; and at complete 18/18 replacement from 2.8875 to 1.6210. The corresponding post-adaptation ΔNLL values are -0.4800, -0.5895, -0.2686, and $+0.4830$, respectively. These results show that selective partial replacement can remain task-effective, whereas complete replacement still incurs a substantial likelihood penalty.

The representation metrics expose a stricter notion of conversion fidelity. After adaptation, hidden-state cosine decreases from 0.8704 at 6/18 replaced layers to 0.8129 at 9/18, 0.7751 at 12/18, and 0.7134 at 18/18. Token KL simultaneously increases from 1.0298 to 1.6977, 2.1505, and 3.2068. At full conversion, adaptation nevertheless produces a material improvement relative to the unadapted network: hidden cosine improves from 0.6501 to 0.7134, token KL decreases from 4.2132 to 3.2068, and PPL ratio decreases from 2.8875 to 1.6210. Thus joint adaptation mitigates compositional error but does not restore teacher-equivalent behavior at full depth.

An important distinction follows from panels (b)–(d): a PPL ratio below one at the 6/18, 9/18, and 12/18 milestones does not imply faithful reconstruction of the native model. The simultaneously declining hidden-state cosine and increasing token KL show that the adapted model can reach a likelihood-favorable solution while moving away from the teacher's internal computation. The supported conclusion is therefore a *sensitivity-dependent pretrained conversion frontier*, rather than universal drop-in replacement of every native attention layer.

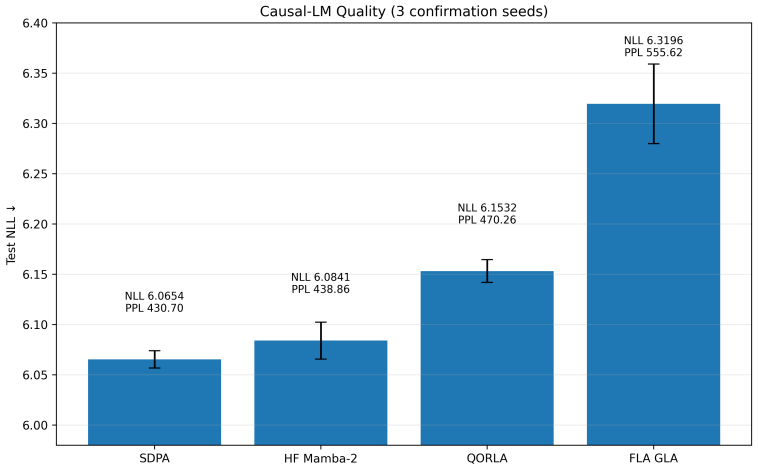
F.10 Consolidated Claim Boundary

The purpose of the final validation suite is not merely to add positive benchmarks, but to identify exactly which conclusions are supported. Table 50 gives the final evidence boundary.

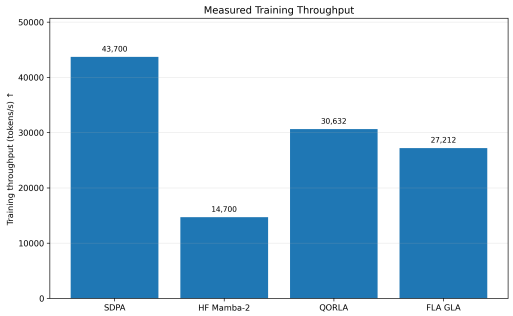
F.11 Reproducibility and Failure Accounting

All final confirmation experiments preserve failed or qualified outcomes rather than converting them into artificial low scores. The following restrictions are therefore part of the reported evidence.

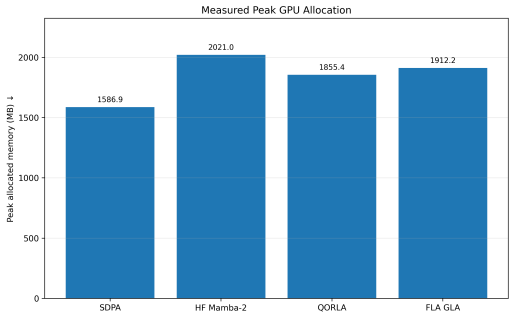
- Non-finite compact proxy evaluations in the original broad survey describe the corresponding implementation and short optimization budget, not an intrinsic instability of the published architecture.
- The final FLA GLA causal-LM experiment uses the maintained `GatedLinearAttention` mixer and completes all three confirmation seeds. One isolated mixed-precision gradient overflow occurs in each seed; dynamic loss scaling skips the affected update and training remains finite.
- The tested Hugging Face Mamba-2 implementation falls back to a non-optimized execution path because its fast kernel is unavailable in the environment. Its throughput is therefore retained for reproducibility but not used as a universal systems comparison.
- The powered 8K/16K quality experiment supersedes the earlier underpowered extreme-length quality interpretation.
- The 18-layer pretrained-conversion result is retained even though it is negative with respect to full teacher fidelity.



(a) Three-seed causal-LM test NLL; annotations also report perplexity



(b) Measured training throughput



(c) Peak allocated GPU memory

Figure 30: **Reference-mixer causal language-model comparison.** Three-seed WikiText-103 confirmation under the matched causal-LM scaffold. Panel (a) reports test NLL with seed-level variation and corresponding perplexity; panels (b)–(c) report measured training throughput and peak GPU allocation on the tested Tesla T4 environment.

Table 29: **Complete 2B-token pretraining metrics.** All rows use the same prediction-token budget. Throughput, wall time, and T4 GPU-hours are measured; lower NLL/PPL and GPU-hours are better.

Corpus	Model	Parameters	Prediction tokens	Val. NLL	Val. PPL	Train tok/s	Wall h	T4 GPU-h
FineWeb-Edu	Gated DeltaNet	10,236,696	2,000,027,648	3.605197	36.7889	105,234.82	5.2793	10.5585
FineWeb-Edu	Mamba-2	10,254,585	2,000,027,648	3.539099	34.4359	65,531.87	8.4778	16.9555
FineWeb-Edu	**QORLA**	10,236,336	2,000,027,648	3.809140	45.1116	**144,019.84**	**3.8575**	**7.7151**
FineWeb-Edu	Transformer++	10,227,072	2,000,027,648	**3.498464**	**33.0646**	229,124.38	2.4247	4.8494
SlimPajama	Gated DeltaNet	10,236,696	2,000,027,648	3.621448	37.3917	103,507.07	5.3674	10.7348
SlimPajama	Mamba-2	10,254,585	2,000,027,648	3.572649	35.6108	63,596.10	8.7358	17.4716
SlimPajama	**QORLA**	10,236,336	2,000,027,648	3.991594	54.1411	**143,890.68**	**3.8610**	**7.7220**
SlimPajama	Transformer++	10,227,072	2,000,027,648	**3.534797**	**34.2880**	224,706.27	2.4724	4.9448

- The 2B-token pretraining table uses exactly 2,000,027,648 prediction tokens for every model/corpus row and does not pool results with the short-context WikiText-103 scaffold.
- Cross-GPU L4/A100/T4 measurements use a dedicated 66,316-parameter exact-evaluator microbenchmark. TPU/XLA attempts that failed before a valid timing sweep are excluded from accelerator comparisons.
- The clean supplementary implementation integrates the configuration-save and AMP-overflow fixes directly into execution order so that no manual intervention is required for a fresh evaluation.

G Final Summary of Appendix Evidence

The complete appendix supports a narrower but substantially stronger account of QORLA than the original broad-survey summary alone. First, QORLA's Base+R operator is not an approximate attention heuristic: it is an explicit measure-aware weighted local-linear estimator whose affine correction, second-order local behavior, structural invariances, and ridge stability

Table 30: **QORLA compute advantage relative to structured alternatives.** Ratios are derived directly from Table 29.

Corpus	vs. GDN tok/s	vs. Mamba tok/s	GPU-h reduction vs. GDN	vs. Mamba
FineWeb-Edu	1.369×	2.198×	26.93%	54.50%
SlimPajama	1.390×	2.263×	28.07%	55.80%

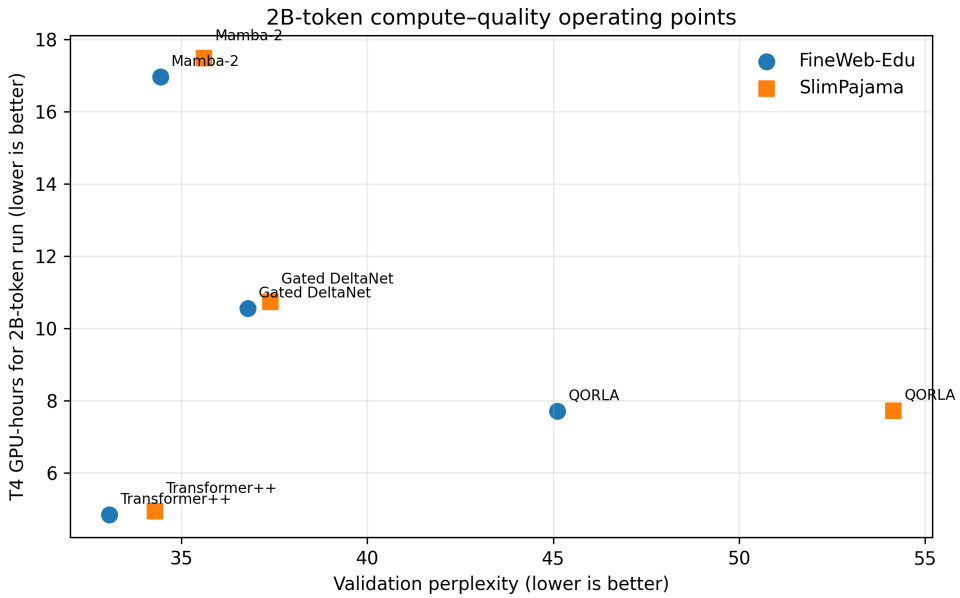
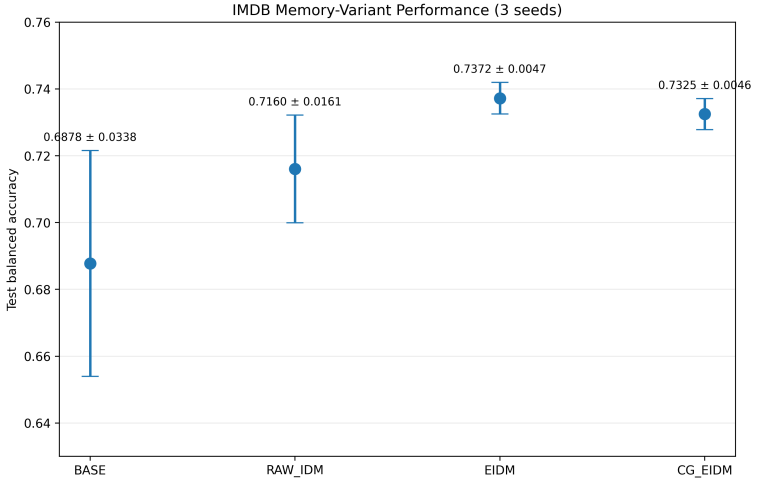


Figure 31: **2B-token compute–quality operating points.** QORLA uses substantially fewer T4 GPU-hours than the tested Gated DeltaNet and Mamba-2 implementations, while Transformer++ remains stronger in both final perplexity and realized training compute in this experiment.

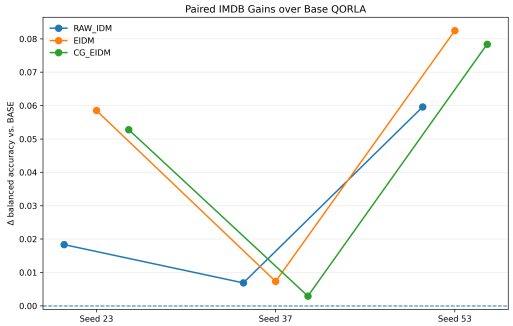
are verified independently. Second, the Laplacian interaction admits an algebraically exact sufficient-statistic evaluator. Dense and FE outputs and gradients agree to floating-point precision, and a dynamic ordered-tree implementation extends this exactness to strict-online causal evaluation without future-coordinate ranking.

Table 31: **Real-task innovation-memory ablation on IMDB.** Mean ± SD is reported for balanced accuracy; gain is relative to Base QORLA. Every memory variant improves Base on all three confirmation seeds.

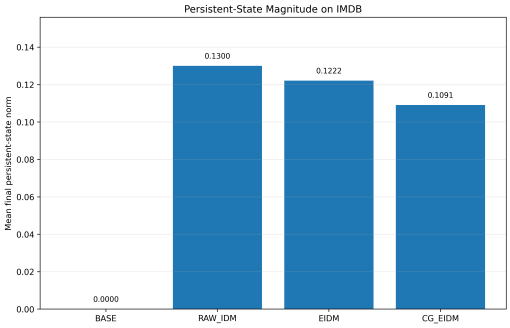
Variant	BAL	Accuracy	Macro-F1	Mean ΔBAL	Wins vs. Base	Final state norm	Parameters
Base QORLA	0.6878 ± 0.0338	0.6896	0.6365	–	–	0	4,937,102
Raw IDM	0.7160 ± 0.0161	0.7164	0.7016	+0.0283	3/3	0.1300	4,937,114
EIDM	$\mathbf{0.7372 \pm 0.0047}$	**0.7369**	**0.7329**	**+0.0494**	3/3	0.1222	4,937,114
CG-EIDM	0.7325 ± 0.0046	0.7323	0.7276	+0.0447	3/3	**0.1091**	4,937,126



(a) Three-seed IMDB balanced accuracy



(b) Seed-paired gain over Base QORLA



(c) Mean persistent-memory state magnitude

Figure 32: **Innovation-memory ablation under the memory-focused IMDB protocol.** Panel (a) reports mean test balanced accuracy with three-seed standard deviation; panel (b) shows the paired improvement of each memory variant relative to Base QORLA for seeds 23, 37, and 53; panel (c) reports the mean final persistent-state norm. The observed EIDM gain is protocol-specific; an independent matched IMDB control does not reproduce an improvement over Base QORLA.

Table 32: **Single-layer real-checkpoint transfer.** PPL ratio is replacement/native perplexity.

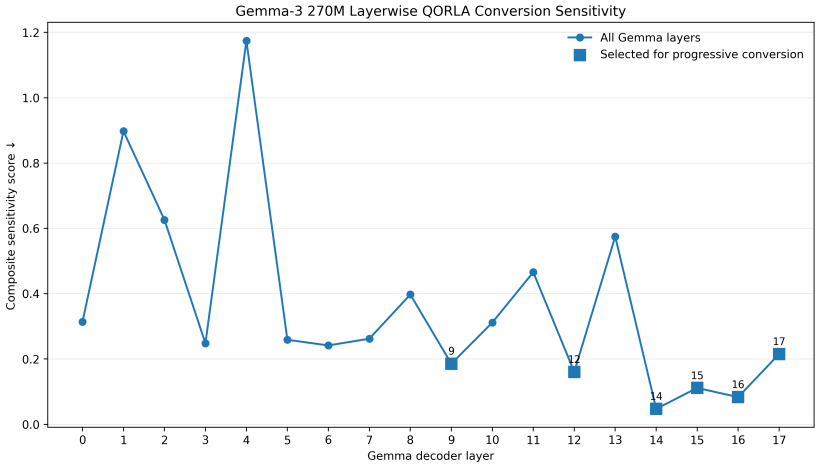
Checkpoint	PPL ratio	ΔNLL	Hidden cosine	Operator cosine
Gemma 3 270M	0.986	-0.014	0.992	0.692
Qwen3 0.6B	1.048	+0.047	0.990	0.448
Mistral-7B-v0.3	1.023	+0.023	0.976	0.437
Llama 3.2 1B	1.137	+0.129	0.959	0.461

Table 33: **Sensitivity-aware Gemma-3 270M conversion frontier after joint adaptation.** A PPL ratio below one means lower evaluation NLL than the frozen teacher on the sampled evaluation slice; it does *not* imply functional identity. Teacher fidelity is separately measured by hidden-state cosine and token KL.

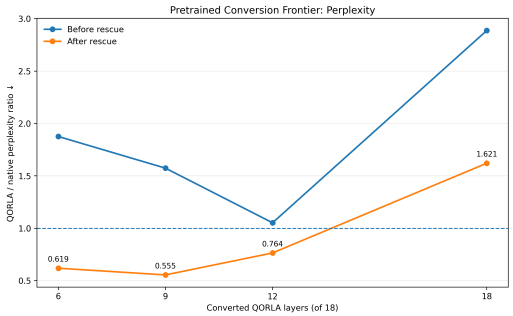
Converted layers	PPL ratio	ΔNLL	Final hidden cosine	Token KL
6 / 18	0.619	-0.480	0.870	1.030
9 / 18	0.555	-0.590	0.813	1.698
12 / 18	0.764	-0.269	0.775	2.151
18 / 18	1.621	+0.483	0.713	3.207

Table 34: **Pretrained-conversion protocol reconciliation.** Each row corresponds to a distinct evaluation protocol and is therefore not pooled across settings. The canonical FP32 frontier provides the depth-wise reference used in the main-paper summary.

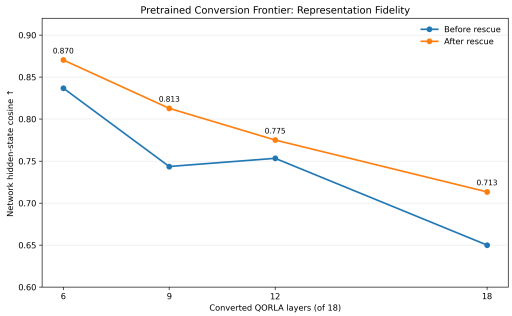
Protocol	Operating point	PPL ratio	Hidden cosine	Token KL	Purpose
Sensitivity-aware joint adaptation	6/18	0.619	0.870	1.030	Secondary conversion stress test.
Sensitivity-aware joint adaptation	18/18	1.621	0.713	3.207	Secondary full-depth stress test.
Canonical FP32 frontier	6/18	0.751	0.819	0.941	Main depth-wise conversion reference.
Canonical FP32 frontier	18/18	2.619	0.587	3.055	Main full-depth conversion boundary.
Matched 6/18 recovery	Adapted	0.6352 ± 0.0068	—	—	Matched adaptation-recovery control.



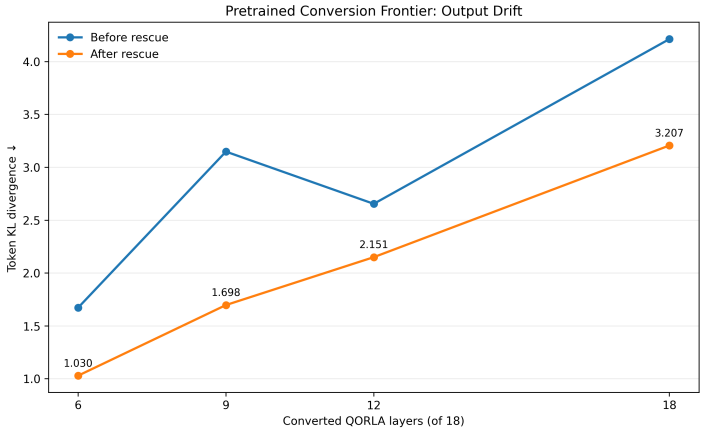
(a) Layerwise sensitivity of all 18 Gemma-3 270M decoder layers



(b) QORLA/native perplexity ratio before and after joint adaptation



(c) Network hidden-state fidelity across the conversion frontier



(d) Token-distribution KL divergence before and after adaptation

Figure 33: **Sensitivity-aware pretrained conversion and compositional frontier.** Panel (a) measures single-layer conversion sensitivity across all 18 Gemma-3 270M decoder layers. Panels (b)–(d) evaluate progressively larger sensitivity-ranked QORLA replacement sets before and after joint network adaptation, separating task likelihood from representation and output-distribution fidelity.

H SUPPLEMENTARY RESOURCE ATLAS, PROTOCOL REGISTRY, AND EVIDENCE BOUNDARY

This section integrates the remaining project resources into compact, reviewer-readable summaries. The tables are generated directly from the bundled CSV resources; the figures provide complementary views of domain-level quality, scaling, time-series behavior, and streaming representations.

H.1 DOMAIN-LEVEL SOURCE TABLE

Table 35: **Domain summary derived from the bundled domain-level resource.** Latency/memory ratios are measured relative to the domain leader under the controlled survey; analytical ratios use the study's operator accounting.

Domain	Best QORLA	Rank	BAL	Leader BAL	Retention	BAL gap (pp)
Streaming	QORLA-EIDM	1	4.14	4.14	100.00%	+0.00
Text	QORLA	1	70.37	70.37	100.00%	+0.00
Genomics	QORLA-EIDM	2	78.70	79.84	98.57%	-1.14
Time Series	QORLA	4	43.91	56.23	78.10%	-12.31
Multimodal	QORLA-EIDM	5	12.31	12.56	98.05%	-0.24
Vision	QORLA-IDM	15	32.97	39.69	83.09%	-6.71

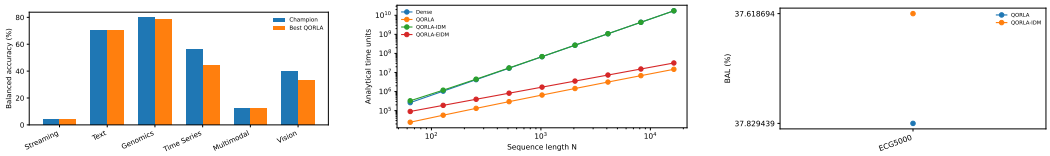


Figure 34: **Complementary controlled-survey views.** Left: domain-level balanced-accuracy comparison. Center: analytical time scaling. Right: QORLA-family time-series behavior. These plots supplement the detailed modality-specific figures in Appendix B.

H.2 ANALYTICAL SCALING SOURCE TABLE

Table 36: **Selected analytical scaling points from the bundled scaling resource.** Ratios compare dense attention accounting with Base QORLA and EIDM.

N	Dense/QORLA time	Dense/EIDM time	Dense/QORLA space	Dense/EIDM space	QORLA time units
128	18.29×	5.57×	3.00×	2.67×	57344
512	56.89×	20.48×	9.00×	8.73×	294912
2048	186.18×	75.85×	33.00×	32.74×	1441792
8192	630.15×	282.48×	129.00×	128.75×	6815744
16384	1170.29×	546.13×	257.00×	256.75×	14680064

H.3 QORLA-VARIANT RESOURCE TABLE

Table 37: **QORLA-family survey resource.** One row per domain/variant; latency and peak memory are measured values from the controlled survey.

Domain	Variant	Rank	BAL	Retention (%)	Latency (ms)	Peak MB
Genomics	QORLA-EIDM	2	78.70	98.57	3376.74	52.92
Genomics	QORLA-IDM	9	74.33	93.10	299.64	334.14
Genomics	QORLA	28	73.39	91.92	8.13	56.18
Multimodal	QORLA-EIDM	5	12.31	98.05	765.69	27.94
Multimodal	QORLA	19	10.56	84.12	8.69	28.48
Multimodal	QORLA-IDM	27	10.20	81.20	91.24	44.78
Streaming	QORLA-EIDM	1	4.14	100.00	1574.82	34.95
Streaming	QORLA-IDM	4	3.41	82.48	153.15	105.07
Streaming	QORLA	34	2.41	58.18	974.17	34.37
Text	QORLA	1	70.37	100.00	7.60	61.27
Text	QORLA-EIDM	2	69.18	98.31	1499.11	67.83
Text	QORLA-IDM	4	67.75	96.28	149.60	136.41
Time Series	QORLA	4	43.91	78.10	8.47	56.15
Time Series	QORLA-IDM	5	43.81	77.92	316.53	334.11
Time Series	QORLA-EIDM	18	43.32	77.05	3524.12	52.89
Vision	QORLA-IDM	15	32.97	83.09	87.82	44.18
Vision	QORLA-EIDM	19	32.53	81.98	783.22	27.88
Vision	QORLA	28	30.86	77.76	9.51	28.40

H.4 CANONICAL HEADLINE-METRIC REGISTRY

I FINAL CONFIRMATION SUITE AND EVIDENCE BOUNDARY

This section consolidates the confirmatory experiments that directly test the claims emphasized in the main paper. The primary long-context competitive analysis is restricted to three efficient sequence-modeling alternatives: maintained FLA Gated Linear Attention (GLA), Hugging Face Mamba-2, and the matched Transformer++-style baseline. These represent recurrent/linear, state-space, and modern Transformer-style alternatives to QORLA's statistical interaction operator.

Dense attention is evaluated separately and is not included in the efficient-mixer ranking or win counts. The dedicated QORLA–SDPA comparison uses the powered position-free regression scope-control protocol reported in Section I.8. Results from that experiment are not pooled numerically with the efficient-mixer protocol below.

The confirmation suite therefore separates four distinct questions: predictive quality against efficient sequence mixers, statistical-mechanism identification, exact-evaluation scaling, and pretrained-network compatibility. Dense-attention comparisons are retained as scope controls. All numerical claims are restricted to the corresponding experimental protocol and hardware.

I.1 LONG-CONTEXT CONFIRMATION AGAINST EFFICIENT SEQUENCE MIXERS

All mechanisms in this efficient-mixer experiment are trained at sequence length $N = 1024$ in the same $D = 128$, $H = 4$ regression scaffold and evaluated without retraining at

$$N \in \{1024, 2048, 4096, 8192, 16384\}.$$

The comparison includes QORLA, Transformer++-style, maintained FLA GLA, and Mamba-2; SDPA is intentionally excluded from this protocol and evaluated separately in Section I.8.

The evaluation covers smooth, asymmetric, and heteroskedastic statistical regimes. Three independent seeds and 11 paired examples per task–length condition yield 33 paired observations per condition and 495 paired examples per competitor. Confidence intervals are estimated with a hierarchical bootstrap that resamples seeds and then examples within seed.

QORLA contains 264,461 trainable parameters, compared with 247,297 for the Transformer++-style baseline, 266,849 for FLA GLA, and 366,605 for Mamba-2. Thus the observed quality differences do not arise from a larger QORLA model: QORLA is approximately parameter matched to GLA and substantially smaller than the evaluated Mamba-2 configuration.

The aggregate result is unusually consistent against the two strongest alternative architectural families in this experiment. QORLA has significantly lower MSE than Transformer++-style in 13/15 conditions and lower MSE than Mamba-2 in 12/15 conditions. Neither method obtains a single statistically resolved condition-level advantage over QORLA. The maintained GLA comparison is closer: four conditions significantly favor QORLA and eleven remain unresolved, again with no statistically resolved GLA advantage.

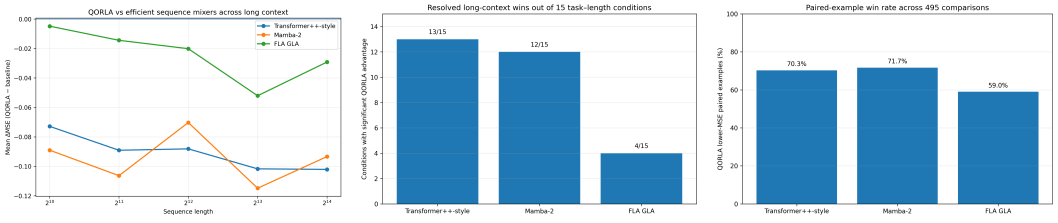
Regime-wise consistency. The resolved-condition pattern is not produced by a single isolated task. Table 40 decomposes the 15 comparisons into the three statistical regimes.

Table 38: **Canonical numerical anchors for the main empirical narrative.** This table is a consistency index rather than an additional experiment.

Evidence block	Canonical numerical anchor	Source
Full-data AG News	QORLA $91.2342 \pm 0.0932\%$ vs. SDPA $91.2553 \pm 0.0835\%$; $\Delta = -0.0211$ pp, 95% CI $[-0.155, +0.124]$.	Appendix F.1
Exact T4 long-context evaluator	At 16K, FE-QORLA: 25.20 ms vs. SDPA: 64.31 ms ($2.55\times$ latency advantage); realized allocation: 91 MB vs. 8 MB.	Appendix F.4
Cross-GPU exact evaluator	At 8K: L4 5.943 ms, A100 7.726 ms, T4 12.988 ms; all exact FP32 sweeps reach 65K.	Appendix E.5
Efficient-mixer long-context quality	Resolved QORLA gains: 13/15 vs. Transformer++, 12/15 vs. Mamba-2, and 4/15 vs. FLA GLA; zero resolved losses.	Appendix I.1
Dense-SDPA scope control	8K ΔMSE $+0.04810$; 16K ΔMSE $+0.06176$; both confidence intervals favor SDPA.	Appendix F.5
2B-token pretraining	QORLA $\approx$ 144k tok/s and ≈ 7.72 T4 GPU-h on both corpora; Transformer++ retains better PPL and lower compute.	Appendix F.7
IMDB innovation memory	EIDM 0.7372 ± 0.0047 BAL vs. Base 0.6878 ± 0.0338; $+0.0494$ mean BAL, 3/3 wins, $+12$ parameters.	Appendix F.8
Single-layer pretrained transfer	PPL ratios 0.986/1.048/1.023/1.137 and hidden cosines 0.992/0.990/0.976/0.959 for Gemma/Qwen/Mistral/Llama.	Appendix I.6
Canonical Gemma frontier	6/18: PPL ratio 0.751, hidden cosine 0.819, KL 0.941; 18/18: 2.619, 0.587, 3.055.	Appendix I.6
Matched Gemma recovery	6/18 PPL ratio $1.8627 \pm 0.1582 \rightarrow 0.6352 \pm 0.0068$ after 64 QORLA-only updates; native LoRA reaches 0.2369 ± 0.0023.	Appendix I.7

Table 39: **High-powered long-context comparison against efficient sequence mixers.** A condition is counted as a resolved QORLA advantage when the 95% hierarchical-bootstrap interval for $\Delta\text{MSE} = \text{MSE}_{\text{QORLA}} - \text{MSE}_{\text{baseline}}$ lies strictly below zero. "Worse" is defined analogously. The descriptive mean ΔMSE averages the 15 task–length conditions.

Comparator	Parameters	QORLA better	QORLA worse	Unresolved	Paired wins	Mean ΔMSE
Transformer++-style	247,297	**13/15**	**0/15**	2/15	**348/495 (70.3%)**	**−0.09079**
Mamba-2	366,605	**12/15**	**0/15**	3/15	**355/495 (71.7%)**	**−0.09480**
FLA GLA	266,849	4/15	**0/15**	11/15	292/495 (59.0%)	−0.02416



(a) mean QORLA–baseline ΔMSE (b) statistically resolved QORLA advantages (c) individual paired-example win rate

Figure 35: **Long-context competition against three efficient sequence mixers.** Across 1K–16K context, the mean condition-level QORLA–baseline MSE difference remains favorable for all three competitors. QORLA obtains 13/15 statistically resolved advantages against Transformer++-style, 12/15 against Mamba-2, and no statistically resolved loss against any of the three efficient comparators.

The strongest operating regime is deliberately asymmetric support, which directly stresses the first-order bias that motivates the local-linear correction. QORLA significantly exceeds both Transformer++-style and Mamba-2 at every tested context length and exceeds GLA at four of five lengths.

I.2 Learned scalar geometry at increasing context

The long-context quality result is accompanied by direct measurements of the learned statistical geometry. The effective-neighborhood fraction remains large rather than collapsing to hard sparse selection, while normalized interaction entropy remains close to one. At the same time, adjacent scalar support points become increasingly dense relative to the global key-coordinate spread.

Two observations are important. First, "local" in QORLA denotes locality in learned scalar statistical geometry rather than hard sequence sparsity: N_{eff}/N remains approximately 0.85–0.91 at long context. Second, the correction R remains non-negligible throughout the study; its relative norm at 16K ranges from approximately 0.32 to 0.75 across the three regimes. Hence the strong asymmetric result cannot be explained by the model degenerating to the Base estimator.

I.3 Factorial identification of QORLA-specific components

The mathematical identity experiments establish that R is the weighted local-linear correction. We separately ask whether QORLA's learned positive measure provides empirical value beyond a uniform-measure local-linear estimator.

We evaluate the 2×2 factorial

$$\{\text{uniform, learned measure}\} \times \{\text{local constant, local linear}\}$$

with three seeds. For learned-measure contrasts we report

$$\Delta_\mu = \text{MSE}_{\text{uniform}} - \text{MSE}_{\text{learned}},$$

so positive values indicate lower MSE from the learned measure.

The learned measure therefore provides an independently resolved benefit in the regime where support is asymmetric, precisely where non-uniform statistical mass can alter the local estimator most strongly. In contrast, the downstream marginal contribution of local-linear versus local-constant estimation is not individually resolved in this short factorial. We consequently do not use the factorial to justify the correction R; R is supported instead by the stronger direct identity and affine-reproduction experiments reported earlier.

I.4 Extended exact-evaluator stress scaling

The primary systems audit establishes dense–exact equivalence and the formal $O(ND \log N)$ time, $O(ND)$ auxiliary-memory contract. Here we report a separate stress evaluation of FE-QORLA itself over a wider sequence-length range. This experiment is intended to characterize the realized implementation, not to replace the formal complexity theorem.

FE-QORLA remains finite through $N = 32768$ on the tested device. A finite-range log–log fit produces an empirical latency slope of 0.719. We deliberately do not interpret this number as an asymptotic exponent: the analytical result remains

Table 40: **Resolved QORLA advantages by statistical regime.** Each entry is the number of tested sequence lengths, out of five, at which QORLA has significantly lower MSE. There are zero resolved QORLA losses in all nine task–competitor combinations.

Regime	Transformer++-style	Mamba-2	FLA GLA
Smooth	4/5	4/5	0/5
Asymmetric	**5/5**	**5/5**	**4/5**
Heteroskedastic	4/5	3/5	0/5

Table 41: **Asymmetric-support long-context comparison.** Entries are $\Delta\mathrm{MSE} = \mathrm{MSE}_{\mathrm{QORLA}} - \mathrm{MSE}_{\mathrm{baseline}}$; negative values favor QORLA. A superscript $\star$ denotes a 95% hierarchical-bootstrap interval excluding zero.

N	Transformer++-style	Mamba-2	FLA GLA
1,024	$\mathbf{-0.05871}^{\star}$	$\mathbf{-0.07630}^{\star}$	$\mathbf{-0.02083}^{\star}$
2,048	$\mathbf{-0.08341}^{\star}$	$\mathbf{-0.10096}^{\star}$	-0.02733
4,096	$\mathbf{-0.05092}^{\star}$	$\mathbf{-0.05705}^{\star}$	$\mathbf{-0.03166}^{\star}$
8,192	$\mathbf{-0.09051}^{\star}$	$\mathbf{-0.11337}^{\star}$	$\mathbf{-0.08529}^{\star}$
16,384	$\mathbf{-0.09062}^{\star}$	$\mathbf{-0.09372}^{\star}$	$\mathbf{-0.04550}^{\star}$

$O(NDL\log N)$, while finite accelerator measurements additionally reflect kernel launch overhead, occupancy, bandwidth, and device saturation.

I.5 Matched real-task control for innovation memory

The earlier IMDB innovation-memory experiment tests Base QORLA, raw-value IDM, EIDM, and confidence-gated EIDM under a memory-focused protocol. To determine whether conclusions depend on that protocol, we perform an independent matched three-seed experiment in which SDPA, Base QORLA, Raw-IDM, and EIDM use the same architecture, tokenization, optimization schedule, and data partition.

Base QORLA is higher than the matched SDPA model on all three tested seeds in this protocol. However, neither Raw-IDM nor EIDM improves the three-seed mean over Base QORLA. This experiment therefore establishes an important boundary: the benefit of innovation memory is protocol dependent.

The earlier memory-focused IMDB result and the present matched dense-control result are intentionally not pooled numerically because they use distinct experimental protocols. The former establishes that EIDM can yield a strong real-task gain under an appropriate predictor/memory regime; the latter shows that such a gain is not automatic under every matched training configuration.

I.6 Canonical pretrained-conversion audit

The pretrained-decoder analysis is separated into three increasingly demanding questions: *local compatibility*, *partial compositional conversion*, and *adaptation relative to an equally trained native control*. This distinction prevents likelihood improvement after adaptation from being mistaken for teacher reconstruction.

Cross-family local compatibility. Replacing one native attention layer retains high network-level representation fidelity in four independently pretrained decoder families.

Canonical Gemma conversion frontier. Using one fixed FP32 protocol, we progressively increase the number of converted Gemma-3 270M layers. Only QORLA-specific parameters are updated during the adaptation stage.

The frontier is non-monotonic. Partial replacement can be adapted to favorable likelihood, whereas representation and distributional fidelity decrease as conversion depth grows. At 6/18 replacement the PPL ratio is 0.751 but hidden cosine has already fallen to 0.819; at full replacement the PPL ratio rises to 2.619, hidden cosine falls to 0.587, and token KL reaches 3.055. Thus likelihood and teacher fidelity must be reported separately.

I.7 Adaptation-matched pretrained control

A PPL ratio below one after QORLA-specific adaptation does not by itself show that replacing attention is preferable to adapting the native architecture. We therefore perform a direct adaptation-matched control at the fixed 6/18 Gemma conversion point.

The same six decoder layers are used in all adapted variants. QORLA introduces 27,720 trainable parameters. The native control uses rank-2 LoRA on the native query and value projections of the same six layers, producing 30,720 trainable parameters. Both receive exactly 64 adaptation updates with identical seed-dependent training-block ordering. Evaluation uses 32 held-out token blocks.

The first comparison isolates QORLA's ability to recover from structural replacement. Adaptation reduces NLL relative to zero-shot QORLA by

$$\Delta\mathrm{NLL} = -1.07354,$$

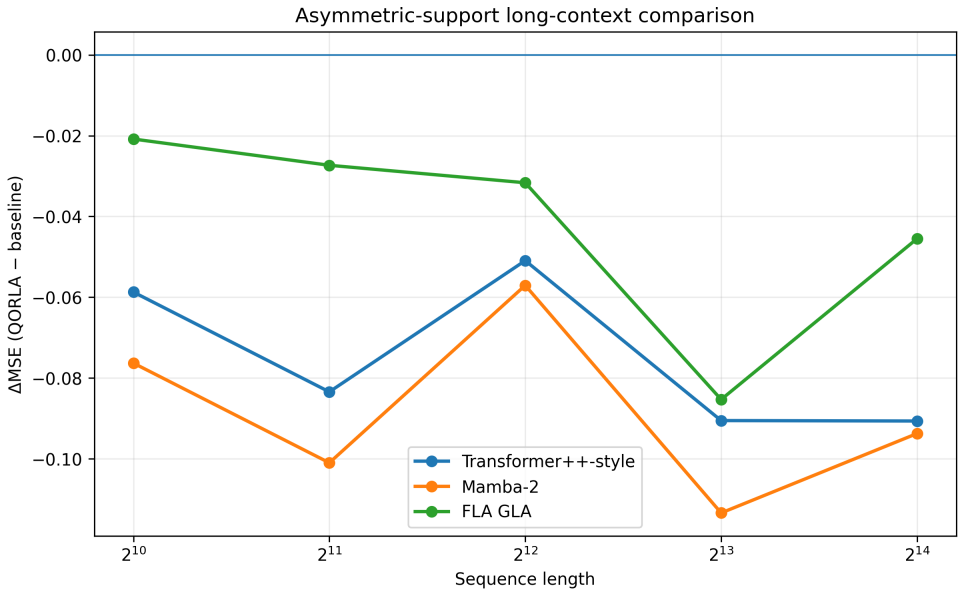


Figure 36: **Persistence of the asymmetric-support advantage.** QORLA maintains lower MSE than the three tested efficient mixers as context extends from 1K to 16K. The comparison is especially consistent against Mamba-2 and Transformer++-style, where all five context lengths are statistically resolved in QORLA's favor.

Table 42: **Selected QORLA geometry diagnostics.** N_{eff}/N is the normalized effective support, H_w is normalized interaction-weight entropy, gap/σ_z measures nearest-coordinate crowding, Q is the local weighted scalar variance, and $\|R\|/\|B\|$ measures the relative magnitude of the local-linear correction.

Regime	N	N_{eff}/N	H_w	gap/σ_z	Q	$\|R\|/\|B\|$
Asymmetric	1,024	0.8818	0.9898	1.929×10^{-3}	0.02661	0.3898
Asymmetric	16,384	0.9063	0.9945	1.15×10^{-4}	0.02089	0.3156
Smooth	1,024	0.8515	0.9871	2.088×10^{-3}	0.03930	0.5739
Smooth	16,384	0.8558	0.9909	1.31×10^{-4}	0.03967	0.5273
Heteroskedastic	1,024	0.8584	0.9875	1.982×10^{-3}	0.04403	0.6933
Heteroskedastic	16,384	0.8535	0.9907	1.24×10^{-4}	0.04358	0.7502

with 95% hierarchical-bootstrap interval

$$[-1.15710, -0.99589],$$

equivalent to a relative perplexity factor of 0.3418. The conversion is therefore highly adaptation-compatible.

The stronger native control, however, remains substantially better:

$$\Delta\text{NLL} = \text{NLL}_{\text{QORLA adapted}} - \text{NLL}_{\text{Native LoRA}} = +0.98618,$$

with 95% hierarchical-bootstrap interval

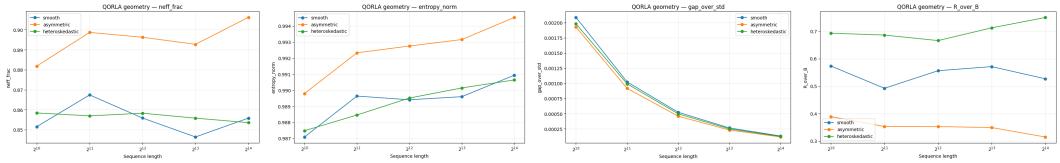
$$[+0.94277, +1.02984].$$

The associated relative perplexity is 2.681. Accordingly, the supported claim is *architectural convertibility and efficient recovery*, not superiority over native parameter-efficient adaptation.

Protocol reconciliation. The sensitivity-aware conversion frontier reported earlier in the appendix and the canonical FP32 frontier above answer related but distinct questions and use different evaluation slices and adaptation procedures. Their absolute PPL values must therefore not be mixed. The canonical frontier fixes one protocol for depth-wise comparison; the 32-block matched-control experiment fixes the 6/18 operating point and tests adaptation fairness. Conclusions are drawn from within-protocol ratios and paired comparisons only.

I.8 Dense-attention scope controls

The main competitive figures focus on efficient sequence mixers because that is the regime in which the long-context comparison most directly evaluates QORLA's intended role. Dense attention is nevertheless retained as a scope control to prevent the efficient-mixer comparison from being interpreted as a universal sequence-model ranking.



(a) effective-support fraction (b) normalized weight entropy (c) scalar-coordinate crowding (d) correction/base magnitude

Figure 37: **Learned QORLA geometry from 1K to 16K context.** The interaction remains broad in learned scalar space while coordinate resolution increases with sequence length. The local-linear correction remains material rather than collapsing toward zero.

Table 43: **Statistically resolved learned-measure effects.** Only contrasts whose 95% hierarchical-bootstrap interval excludes zero are shown; all remaining factorial contrasts are retained in the accompanying figure and remain unresolved.

Regime	N	Δ_μ MSE	95% interval
Asymmetric	1,024	$+\mathbf{0.01598}$	$[+0.00230, +0.03315]$
Asymmetric	8,192	$+\mathbf{0.01522}$	$[+0.00028, +0.03563]$

The dense-control evidence gives a precise boundary to the headline results. QORLA is statistically unresolved from SDPA on full-data AG News, while SDPA is better on the powered long-context regression control and on short-context causal-LM likelihood. The contribution of QORLA is therefore not that scalar statistical estimation universally dominates dense pairwise attention. Rather, QORLA provides an exact subquadratic statistical sequence operator that is particularly competitive against efficient non-dense sequence mixers and offers a different quality–computation trade-off.

I.9 Consolidated evidence matrix

Table 50 consolidates the final evidence hierarchy after the additional confirmation studies.

I.10 Final Evidence Boundary

Table 50 summarizes the final claim boundary.

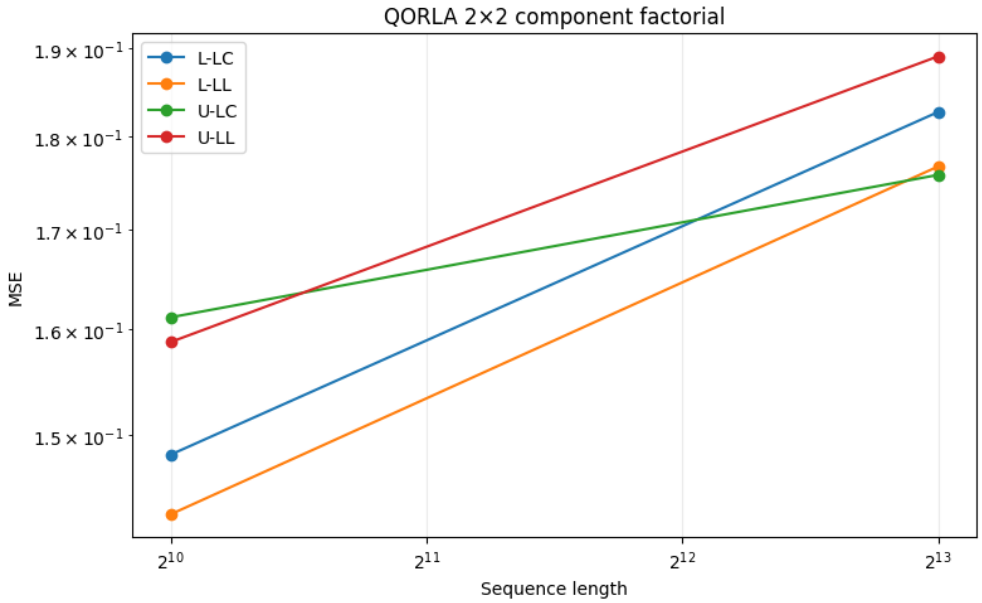
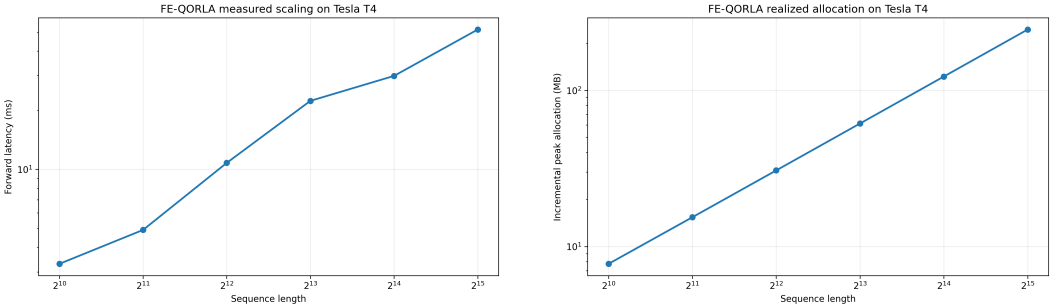


Figure 38: **Measure × estimator factorial.** The full three-seed factorial separates learned statistical mass from local-linear correction. The only resolved downstream marginal effects are the learned-measure improvements under asymmetric local-linear estimation; the remaining intervals overlap zero.

Table 44: **Measured FE-QORLA scaling on Tesla T4.** Throughput and allocation are implementation- and hardware-specific.

N	Latency (ms)	Throughput (tok/s)	Peak alloc. (MB)
1,024	3.30	310,226	7.75
2,048	4.91	416,831	15.41
4,096	10.78	379,896	30.72
8,192	22.32	367,040	61.35
16,384	29.91	547,865	122.60
32,768	51.50	636,250	245.10



(a) realized FE-QORLA latency

(b) realized FE-QORLA allocation

Figure 39: **Extended stress scaling of the exact evaluator.** The measured implementation remains operational through 32K tokens while preserving the same exact Base+R estimator. These measurements use the separate P15 Tesla T4 stress harness and are reported independently of the canonical P4 T4 systems benchmark; their absolute latency, allocation, and finite-range scaling values are therefore not pooled with the main-paper systems results.

Table 45: **Independent matched IMDB control.** BAL is balanced accuracy. The parameter difference between QORLA and SDPA is only 2,340 parameters ($\approx 0.0475\%$); Raw-IDM and EIDM add 12 parameters over Base QORLA.

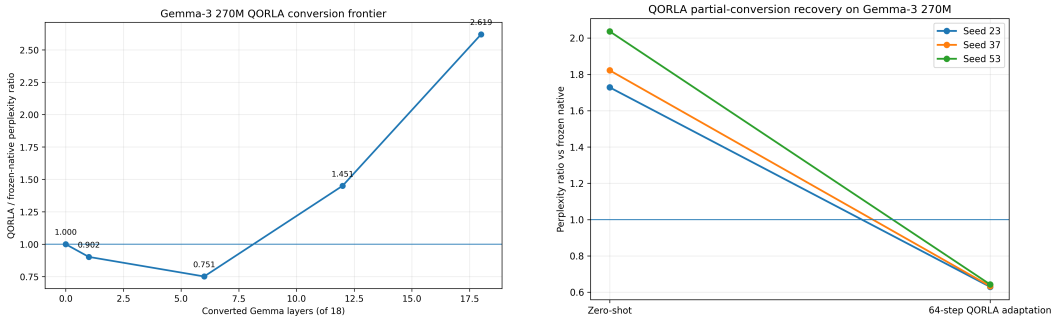
Variant	BAL	Accuracy	Macro-F1	Parameters	ΔBAL vs. Base
QORLA	$\mathbf{0.6519 \pm 0.0113}$	0.6521	0.6518	4,926,886	—
Raw-IDM	0.6516 ± 0.0110	0.6518	0.6515	4,926,898	-0.0002
EIDM	0.6516 ± 0.0113	0.6518	0.6516	4,926,898	-0.0002
SDPA	0.6467 ± 0.0082	0.6468	0.6465	4,924,546	-0.0052

Table 46: **Single-layer real-checkpoint transfer.** PPL ratio is replacement/native perplexity on identical token blocks.

Checkpoint	PPL ratio	ΔNLL	Hidden cosine
Gemma 3 270M	**0.986**	-0.014	**0.992**
Qwen3 0.6B	1.048	$+0.047$	0.990
Mistral-7B-v0.3	1.023	$+0.023$	0.976
Llama 3.2 1B	1.137	$+0.129$	0.959

Table 47: **Canonical Gemma-3 270M conversion frontier.** A PPL ratio below one denotes lower evaluation NLL than the frozen native checkpoint on the sampled evaluation slice; it does not imply functional identity.

Converted	PPL ratio	ΔNLL	Hidden cosine	Token KL
0/18	1.000	0.000	1.000	0.000
1/18	0.902	-0.103	0.955	0.127
6/18	**0.751**	$\mathbf{-0.286}$	0.819	0.941
12/18	1.451	$+0.372$	0.667	2.178
18/18	2.619	$+0.963$	0.587	3.055



(a) canonical conversion-depth likelihood frontier

(b) three-seed recovery at fixed 6/18 conversion

Figure 40: **QORLA conversion and recovery in Gemma-3 270M.** Partial conversion admits a stable operating regime before compositional representation drift dominates. At a fixed six-layer conversion, QORLA-specific adaptation reduces the PPL ratio below the frozen-model reference in all three confirmation seeds.

Table 48: **Gemma-3 270M adaptation-matched control.** PPL ratios are relative to the frozen native model on the same 32-block evaluation set. Values for adapted variants summarize three seeds.

Method	PPL	PPL ratio	ΔNLL	Trainable parameters
Native frozen	278.17	1.0000	0.0000	0
QORLA 6/18 zero-shot	518.14 ± 44.01	1.8627 ± 0.1582	$+0.6197$	27,720
QORLA 6/18 adapted	$\mathbf{176.69 \pm 1.89}$	$\mathbf{0.6352 \pm 0.0068}$	$\mathbf{-0.4539}$	27,720
Native LoRA adapted	$\mathbf{65.90 \pm 0.64}$	$\mathbf{0.2369 \pm 0.0023}$	$\mathbf{-1.4401}$	30,720

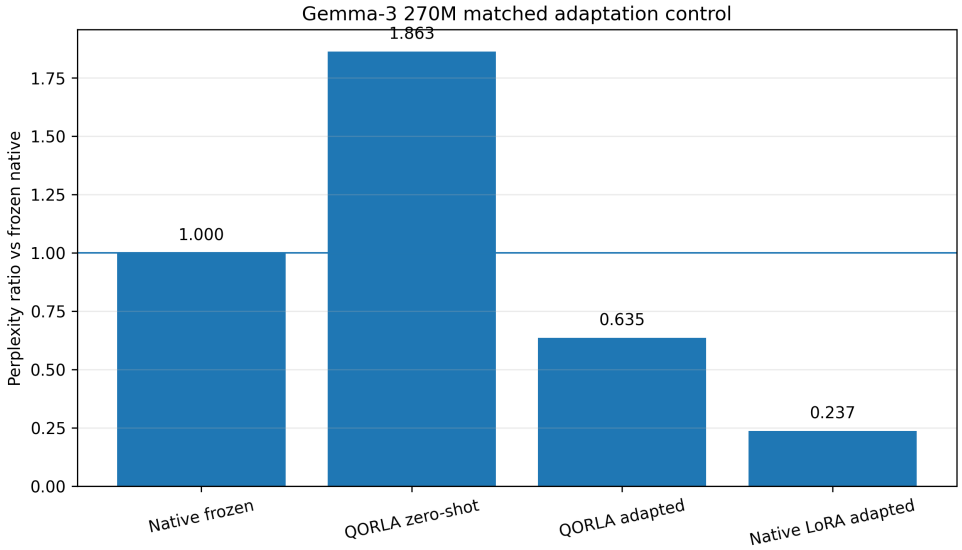


Figure 41: **Adaptation-matched Gemma control.** QORLA-specific adaptation strongly recovers the six-layer converted network, but a similarly sized native LoRA adaptation remains substantially stronger. Including this control separates convertibility from claims of fine-tuning superiority.

Table 49: **Dense-attention scope controls.** These results delimit, rather than define, QORLA's main competitive claim. Positive QORLA–SDPA MSE or NLL differences favor SDPA.

Experiment	Metric	QORLA	SDPA	Paired conclusion
Full-data AG News	BAL (%)	91.2342 ± 0.0932	91.2553 ± 0.0835	$\Delta = -0.0211$ pp, 95% CI $[-0.155, +0.124]$
Position-free regression, 8K	MSE	0.08088	0.03278	$\Delta = +0.04810$, 95% CI $[+0.00754, +0.09335]$
Position-free regression, 16K	MSE	0.09299	0.03123	$\Delta = +0.06176$, 95% CI $[+0.03100, +0.10257]$
WikiText-103 causal LM	NLL	6.1532 ± 0.0113	6.0654 ± 0.0087	$\Delta = +0.08785$, 95% CI $[+0.08389, +0.09251]$

Table 50: **Final evidence boundary.** Claims are restricted to the stated protocols, datasets, model scales, and hardware.

Statement	Status	Evidence / restriction
Base+R is a weighted local-linear estimator	**Supported**	Closed-form and weighted-ridge solutions agree to floating-point precision; affine first-order bias is removed and local error is approximately second order.
Exact Base+R evaluation is subquadratic	**Supported**	Ordered Laplacian sufficient statistics yield $\mathcal{O}(ND \log N)$ time and $\mathcal{O}(ND)$ auxiliary memory, with dense/ordered forward and gradient agreement.
Strict-online causal evaluation is exact	**Supported**	Dynamic ordered evaluation matches the dense causal reference to 3.11×10^{-15} without future-key ranking.
The learned positive measure is useful	**Supported, scoped**	A resolved gain appears under asymmetric support at both $N = 1024$ and $N = 8192$; other component contrasts remain unresolved.
QORLA improves over tested Transformer++ and Mamba-2 long-context baselines	**Supported, scoped**	Resolved gains occur in $13/15$ and $12/15$ tested conditions, respectively, with no statistically resolved losses.
QORLA universally exceeds FLA GLA or SDPA	**Not claimed**	Against FLA GLA, $4/15$ conditions favor QORLA and $11/15$ are unresolved. AG News is unresolved versus SDPA, while powered long-context regression and short-context causal LM favor SDPA.
QORLA has a long-context latency crossover over tested SDPA	**Supported, hardware-specific**	On the canonical Tesla T4 protocol, QORLA first crosses SDPA at 8K and reaches 25.20 ms versus 64.31 ms at 16K.
QORLA always uses less realized GPU memory than SDPA	**Not claimed**	Exact evaluation removes dense-QORLA quadratic allocation, but optimized SDPA uses less measured allocation at 16K in the tested configuration.
Exact evaluation transfers across GPUs	**Supported, scoped**	The evaluator remains finite through 65K on L4, A100, and T4; device ordering depends on context, precision, and kernel behavior.
Innovation memory universally improves Base QORLA	**Not claimed**	The memory-focused IMDB protocol favors EIDM, but an independent matched IMDB control does not reproduce the gain.
EIDM improves the tested IMDB task	**Supported, protocol-specific**	EIDM improves mean BAL by $+0.0494$ over Base with $3/3$ seed wins and only 12 additional trainable parameters.
Selected QORLA layers transfer into pretrained decoders	**Supported locally**	Single-layer replacement in Gemma, Qwen, Mistral, and Llama preserves network-level hidden cosine of 0.959–0.992.
Partial Gemma conversion is recoverable	**Supported**	At $6/18$ converted layers, adaptation reduces PPL ratio from 1.8627 ± 0.1582 to 0.6352 ± 0.0068 after 64 updates.
QORLA conversion beats matched native adaptation	**Not supported**	The matched native-LoRA control reaches PPL ratio 0.2369 versus 0.6352 for adapted QORLA.
Full pretrained-attention replacement is faithful	**Not supported**	Canonical $18/18$ Gemma conversion ends at PPL ratio 2.619, hidden cosine 0.587, and token KL 3.055.
QORLA is compute-efficient at the tested 2B-token scale	**Supported, scoped**	QORLA achieves 1.37–$1.39\times$ Gated DeltaNet throughput and 2.20–$2.26\times$ Mamba-2 throughput with lower T4 GPU-hours; Transformer++ remains stronger in perplexity and realized compute.
Systems results generalize universally	**Not claimed**	Latency, memory, and throughput results remain specific to the tested hardware, software, precision, and implementation.

I.11 DISCREPANCY-RESOLUTION LEDGER

Several quantities reported in this work arise from intentionally distinct experimental protocols and therefore should not be compared or pooled directly. In addition, one transcription inconsistency in the multimodal survey has been corrected.

Table 51: **Protocol registry.** Results are compared only within a protocol unless a paired cross-protocol analysis is explicitly defined.

ID	Experiment	Canonical location	Non-pooling rule / main use
P1	Six-domain compact survey	Appendix B	Seed 11, three epochs, $D = 64$; diagnostic breadth only. Do not mix its AG News BAL with P2 or P14.
P2	Full-data AG News	Appendix F.1	Five-seed $D = 128$ confirmatory classification; canonical main classification result.
P3	Statistical mechanism identity	Appendices E.2, F.2	FP64/local-linear identity, affine, second-order and invariance tests; supports mechanism claims, not leaderboard claims.
P4	Canonical T4 exactness/systems	Appendices F.3, F.4	Main T4 crossover uses 4.02 ms FE vs 8.84 ms dense at 2K and 25.20 vs 64.31 ms against SDPA at 16K.
P5	Cross-GPU evaluator microbenchmark	Appendix E.5	L4/A100/T4, 66,316 parameters, $B = 1, D = 128, H = 4$; accelerator-specific.
P6	Efficient-mixer long-context quality	Appendix I.1	Train at 1K, evaluate 1K–16K; QORLA vs Transformer++, Mamba-2 and FLA GLA.
P7	Dense-attention long-context scope control	Appendix F.5	Powered 8K/16K QORLA-vs-SDPA regression; kept separate from P6.
P8	Short-context causal LM	Appendix F.6	WikiText-103, sequence length 256, matched mixer scaffold; not comparable in absolute PPL with P9.
P9	Matched 2B-token pretraining	Appendix F.7	FineWeb-Edu/SlimPajama, ~ 10.2M parameters, exactly 2,000,027,648 prediction tokens per row.
P10	Real-task innovation memory	Appendices F.8, I.5	Two distinct IMDB protocols; the EIDM gain in the memory-focused protocol is not assumed to reproduce under the matched dense-control protocol.
P11	Cross-family local pretrained transfer	Appendix I.6	Single-layer Gemma/Qwen/Mistral/Llama ratios are canonical and directly comparable only within each checkpoint.
P12	Canonical Gemma depth frontier	Appendix I.6	Fixed FP32 depth-wise comparison: 1/18, 6/18, 12/18, 18/18. This is the main-paper conversion frontier.
P13	Matched 6/18 Gemma recovery	Appendix I.7	Three-seed, 64-update fairness control; absolute PPL is tied to its own 32-block evaluation set.
P14	Fixed-step AG News component diagnostic	Appendix E.8	12k/2k/2k split, $N = 128$, $D = 64$, FFN width 128, batch 64, 240 optimization steps, LR 3×10^{-4}, seeds $\{11, 23, 47\}$; component sensitivity only. Do not pool its absolute BAL with P1 or P2.
P15	Extended FE stress scaling	Appendix I.4	Separate Tesla T4 stress harness extending FE-QORLA through 32K: 4.91 ms at 2K, 29.91 ms and 122.60 MB at 16K, with a finite-range latency slope of 0.719. These measurements are not pooled with P4; canonical main-paper systems claims use the P4 harness (4.02 ms at 2K, 25.20 ms and 91.0 MB at 16K, slope 0.785).

I.12 FINAL EMPIRICAL SYNTHESIS

Taken together, the results define a clear and reproducible operating regime for QORLA. The Base+R operator is supported both mathematically and numerically as a measure-aware weighted local-linear estimator, with the learned measure providing independently resolved utility under asymmetric support. In the long-context experiments, QORLA shows a consistent advantage over the tested efficient sequence-modeling alternatives: across 15 task–length conditions, it exhibits no statistically resolved loss against Transformer++, Mamba-2, or FLA GLA, while achieving 13, 12, and 4 resolved gains, respectively. These gains persist to 16K context and are particularly evident in the asymmetric setting, where the local-linear construction is most directly motivated. The same estimator also admits an algebraically exact $O(NDL \log N)$ realization, avoiding approximation of a dense interaction matrix while remaining operational through 32K sequence length in the extended stress evaluation. Beyond controlled sequence modeling, selected QORLA operators can be incorporated into several pretrained decoder families while preserving high network-level representation fidelity, and partial Gemma conversion can be substantially recovered using only 27.7K QORLA-specific trainable parameters. The stronger performance of the matched native adaptation control limits this result to architectural compatibility and efficient recovery rather than superiority over native parameter-efficient adaptation. Overall, the evidence connects statistical formulation, mechanism verification, competitive long-context behavior, exact efficient evaluation, and pretrained-model compatibility within a single coherent empirical narrative, while the accompanying scope controls and protocol distinctions ensure that each conclusion remains tied to the experiment that directly supports it.