EXACT ATTENTION MERGING BEYOND SOFTMAX MIXTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient attention relies on compact summaries that merge exactly across blocks. Does retaining more states simply amount to combining more softmax components? We show that coupled moment states provide a different design space. For positive real-rate exponential-polynomial generators, we characterize the minimum number of retained value vectors, give an attaining merge rule, and prove a matching lower bound even with nonlinear decoding. A finite Hankel criterion separates nonnegative softmax mixtures from positive, monotone maps with coupled states. The simplest instance uses two value vectors and cannot be represented by any nonnegative softmax mixture on the same scores. Controlled learning experiments connect this structural separation to estimation risk: learned two- and three-state maps approach Bayes performance, and the three-state map transfers to unseen scores. In language modeling, the two-state map improves held-out loss over softmax at three context lengths while maintaining comparable full-model latency to FlashAttention-2. Learned-shape diagnostics in the trained decoder, together with fused execution on Qwen and SmolLM2, further connect the construction to Transformer computation. These results establish state coupling as a design axis for expressive, exactly mergeable attention.

1 INTRODUCTION

Exact blockwise attention summarizes a sequence in pieces and combines the partial states. Softmax retains a maximum, a normalizer and a weighted value sum. A later maximum rescales those quantities, allowing the next block to merge without rereading old values. FlashAttention-2 exploits this operation in tiled computation, while FlashInfer makes attention-state composition an explicit system primitive (Dao, 2024; Ye et al., 2025b). Compact merging therefore ties the choice of attention map to its execution. What attention shapes can a fixed state budget support?

One natural extension stores several independently rescaled softmax components. We investigate a different use of the same memory: store moments that transform together. When a block is recentered, its zeroth moment contributes to the update of its first moment. This coupling preserves a polynomial factor alongside the exponential. The resulting attention merges exactly, yet its weights lie outside every nonnegative softmax mixture on the original scores. Figure 1 follows the resulting weight update. More retained states can change the structure of attention, not just its number of components.

We make this distinction precise for positive real-rate exponential polynomials. First, the minimum number of retained value vectors equals the dimension of the generator's translation span. Maximum-aligned moments achieve this count. A matching lower bound follows by appending zero-valued tokens at larger scores: their new maxima expose independent combinations of old values, all of which an exact summary must preserve. Second, a finite Hankel signature characterizes nonnegative mixture membership. Nonconstant polynomial blocks introduce negative directions in that signature and can do so while keeping the attention weights positive and monotone. The two results identify both the state cost and the structural gain of coupling.

Our experiments follow the construction from learnable shape to prediction and execution. In Gaussian aggregation tasks, coupled maps learn from finite data, cross a certified risk bound for positive mixtures, and transfer to unseen score values. From-scratch language modeling then compares

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

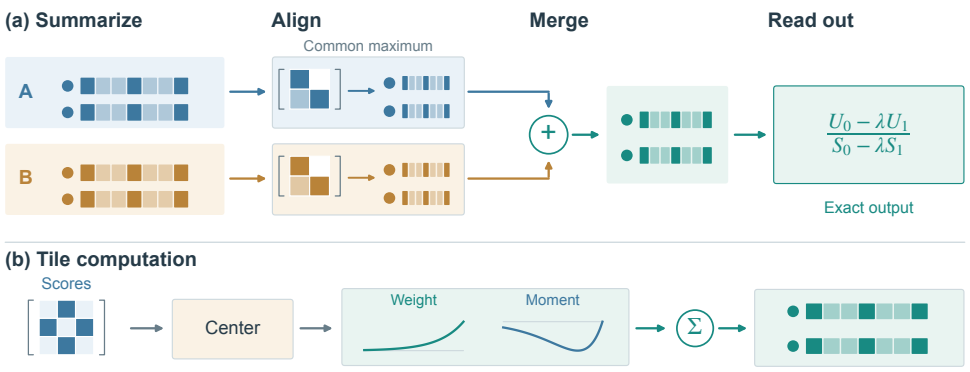


Figure 1: Blocks align to a common maximum, add and decode. Dots and strips denote scalar and value moments; the inset shows their tiled computation.

softmax, learned temperature, independent mixtures and coupled maps. The two-state instance improves mean held-out loss at all three tested context lengths. Head-wise geometry and output-vector interventions show how the learned shape survives value aggregation. Finally, full-model measurements compare its execution with FlashAttention-2 and trace longer-context scaling on two existing backbones.

2 Problem Setup

Fix a query, scores $s_i \in \mathbb{R}$ and arbitrary values $v_i \in \mathbb{R}^D$. Let $m = \max_i s_i$ and $t_i = s_i - m \le 0$. A positive generator h defines

$$w_i(s) = h(t_i)/Z_h(s), \quad Z_h(s) = \sum_j h(t_j), \quad A_h(s,V) = \sum_i w_i(s) v_i. \tag{1}$$

The weights are permutation equivariant and invariant to a common score shift. They are *own-score monotone* when increasing one score cannot decrease its own weight.

Lemma 1 (Monotonicity on the centered domain)**.** *For positive C^1 generators, own-score monotonicity on every finite list is equivalent to $h'(t) \ge 0$ for all $t \le 0$.*

For a nonmaximum score the derivative is $w_i(1 - w_i)h'(t_i)/h(t_i)$; for the unique maximum it is $w_i \sum_{j \ne i} w_j h'(t_j)/h(t_j)$. Limits cover ties, and two-token lists prove necessity (Appendix G).

Value-linear associative summaries. A block summary stores fixed-dimensional score-only state $\tau(s)$ and r value vectors,

$$\sigma(s,V) = (\tau(s), U_1, \dots, U_r), \qquad U_k = \sum_i \phi_{k,i}(s) v_i. \tag{2}$$

Coefficients can depend on the complete block score list. An associative merge operation supports exact decoding for every finite input and every value assignment; the decoder can be nonlinear. The minimum r is the *merge-rank*, which counts retained value vectors.

We consider the minimal real-rate representation

$$h(t) = \sum_{k=1}^{m_0} e^{a_k t} q_k(t), \quad q_k(t) = \sum_{p=0}^{d_k} c_{k,p} t^p, \quad R = \sum_{k=1}^{m_0} (d_k + 1). \tag{3}$$

Rates are distinct and real, $c_{k,d_k} \ne 0$, and the parameters are fixed across inputs. Repeated rates are combined before counting R. We require $h > 0$ on the entire centered half-line; coefficients within an otherwise positive generator may have either sign.

3 THEORY

3.1 EXACT STATE REQUIREMENTS

Theorem 1 (Exact merge-rank). *A strictly positive minimal generator in Equation (3) has merge-rank exactly R in the value-linear model of Equation (2). The count is unchanged when $h' \geq 0$, i.e., when the map is own-score monotone.*

In particular, rank one forces a single exponential; among monotone maps this is softmax with nonnegative inverse temperature, including uniform weights at zero. At larger rank, channels can retain coupled polynomial moments rather than separate mixture components.

3.1.1 MOMENT-BASED CONSTRUCTION

For each rate and degree, retain the block maximum and paired moments

$$S_{k,p} = \sum_i t_i^p e^{a_k t_i}, \qquad U_{k,p} = \sum_i t_i^p e^{a_k t_i} v_i, \quad 0 \leq p \leq d_k. \tag{4}$$

To align an old block maximum m to a merged maximum M, set $\delta = m - M$. The binomial expansion gives

$$S'_{k,p} = e^{a_k \delta} \sum_{j=0}^{p} \binom{p}{j} \delta^{p-j} S_{k,j}, \quad U'_{k,p} = e^{a_k \delta} \sum_{j=0}^{p} \binom{p}{j} \delta^{p-j} U_{k,j}. \tag{5}$$

These are lower-triangular Pascal actions. Two shifts compose by adding their offsets; after alignment, corresponding block states add. Every merge tree therefore aligns each original token to the same final maximum. The readout

$$A_h = \frac{\sum_{k,p} c_{k,p} U_{k,p}}{\sum_{k,p} c_{k,p} S_{k,p}} \tag{6}$$

has denominator exactly $Z_h > 0$, proving sufficiency and exact-arithmetic associativity. The state comprises R value vectors, R scalars and one maximum; Algorithm 1 gives explicit pseudocode. Direct triangular merging costs $O((D+1)\sum_k (d_k+1)^2)$ arithmetic operations.

3.1.2 STATE LOWER BOUND

Fix R distinct old scores $x_j \leq 0$ with maximum zero, and arbitrary scalar values v_j. Any rank-r summary stores Lv for some $r \times R$ matrix L; its score-only part is now constant. Append a value-zero token at score $b > 0$. Its output is

$$\frac{\sum_j h(x_j - b) v_j}{h(0) + \sum_j h(x_j - b)}. \tag{7}$$

Each future maximum yields a different linear combination of old values.

Lemma 2 (Continuation span). *For any positive gaps $b_1, \ldots, b_N$, exactness requires*

$$r \geq \operatorname{rank} P, \qquad P_{\ell j} = h(x_j - b_\ell). \tag{8}$$

Proof. If $Lv = 0$, the retained state equals the state for zero values. Merging either with the same zero-valued singleton must decode to zero. Equation (7) gives $Pv = 0$, hence $\ker L \subseteq \ker P$ and $r \geq \operatorname{rank} P$. The argument permits a nonlinear decoder. □

Taylor expansion factors the continuation matrix as

$$P = BX^\top, \quad X_{j,(k,p)} = x_j^p e^{a_k x_j}, \quad B_{\ell,(k,p)} = e^{-a_k b_\ell} q_k^{(p)}(-b_\ell)/p!. \tag{9}$$

Taking $x_j = -j$ and $b_\ell = \ell + 1$, $j, \ell = 0, \ldots, R-1$, gives two nonsingular confluent Vandermonde factors: nodes e^{-a_k} are distinct and leading coefficients are nonzero. Thus $\operatorname{rank} P = R$. Appendix H gives determinant products and completes the necessity proof. A quadratic block needs its intermediate linear moment even when that moment has zero readout coefficient, because future translations make it observable.

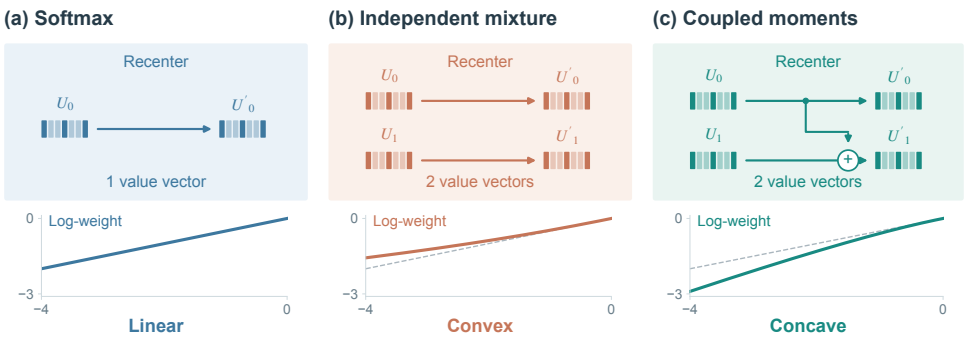


Figure 2: One state budget, distinct shapes. Independent states rescale diagonally; coupled moments mix triangularly. The curves $e^{t/2}$, $(e^{t/4} + e^{3t/4})/2$ and $e^t(1 - t/2)$ share unit value and log-slope $1/2$ at zero.

3.2 Beyond Softmax Mixtures

A nonnegative mixture *locked* to a list (gates shared by all of its tokens) has the form

$$w_i(s) = \sum_{\ell=1}^{J} \pi_\ell(s) \frac{e^{\alpha_\ell(s) s_i}}{\sum_j e^{\alpha_\ell(s) s_j}}, \qquad \pi_\ell \geq 0, \quad \sum_\ell \pi_\ell = 1. \tag{10}$$

Rates, gates and component count may adapt to the complete list; the gates are shared across its tokens. On scores $0, -x, \ldots, -(2R-2)x$, each component is a positive geometric sequence, so its Hankel matrix is a positive outer product. Every such mixture must therefore yield a positive semidefinite (PSD) matrix. The common positive normalizer preserves the signature of

$$H_R(x) = [h(-(i+j)x)]_{i,j=0}^{R-1}, \qquad x > 0. \tag{11}$$

Theorem 2 (Block inertia and mixture criterion)**.** *For every positive spacing, the inertia (n_+, n_-, n_0) of $H_R(x)$ is the sum over rate blocks of*

$$\begin{cases} ((d_k+1)/2, (d_k+1)/2, 0), & d_k \text{ odd}, \\ (d_k/2+1, d_k/2, 0), & d_k \text{ even}, \ c_{k,d_k} > 0, \\ (d_k/2, d_k/2+1, 0), & d_k \text{ even}, \ c_{k,d_k} < 0. \end{cases} \tag{12}$$

For $h > 0$, its map admits a locked nonnegative mixture on every finite list if and only if all $d_k = 0$ and all $c_{k,0} > 0$.

Proof mechanism. Congruence preserves inertia (Horn & Johnson, 2012). The confluent Vandermonde congruence in Equation (22) reduces the weight matrix to coefficient blocks. Their anti-diagonals are nonzero, and their determinants are unchanged as lower coefficients shrink to zero. Off-center anti-diagonal pairs contribute one positive and one negative direction; an even-degree central entry has the leading coefficient's sign. Hence every nonconstant block contributes a negative direction. Conversely, positive constant blocks reproduce a softmax mixture using gates proportional to their coefficient-weighted block normalizers. Appendix I proves the complete signature, including irregular offsets.

The minimal representation determines the value-channel count, while the coefficient-block signature determines mixture membership. Appendix E develops *legal* (positive and own-score monotone on $t \leq 0$) families, including generators whose first three weights miss the obstruction. We next instantiate the coupled construction with two and three value states.

4 Method

Two-state attention. We use the degree-one generator, denoted *Ray*,

$$h(t) = e^{at}(1 - \lambda t), \qquad a \geq \lambda > 0, \quad t \leq 0. \tag{13}$$

Its positive weights are own-score monotone, and its merge-rank is two. The normalized readout is $(U_0 - \lambda U_1)/(S_0 - \lambda S_1)$. Figure 2 gives the recentering action: the exponential of a size-two Jordan block, whose zeroth moment feeds the first. On three equally spaced scores, the mixture obstruction is

$$h(-x)^2 - h(0)h(-2x) = e^{-2ax}\lambda^2 x^2 > 0. \tag{14}$$

Thus two retained vectors suffice for a map outside all nonnegative mixtures in Equation (10). In learned heads, we set $\lambda = \mathrm{softplus}(\ell)$ and $a = \lambda + \mathrm{softplus}(b)$.

Ray has negative log-curvature, $(\log h)''(t) = -\lambda^2/(1-\lambda t)^2$, whereas positive exponential sums have nonnegative log-curvature. Figure 2(c) illustrates this distinction; Section 5.4 measures the learned head geometry.

At the maximum, Ray's log-slope is $a - \lambda$. Holding this slope fixed while changing λ changes how the weighting rule bends away from the maximum. The second retained moment therefore carries a shape parameter distinct from temperature. During recentering, its update combines the old first moment with the zeroth moment, preserving the information needed to apply this shape at the new origin. The same two states are used for every block size and merge order.

Three-state attention. The quadratic family $h(t) = e^{at}(1 + \beta t^2)$, with $\beta > 0$ and $a \geq \sqrt{\beta}$, retains moments of orders zero, one and two. Its logarithmic curvature changes sign across the score domain, providing a richer shape than the degree-one family. We parameterize $a > 0$ and $\beta = \mathrm{sigmoid}(u)a^2$. The same moment merge applies, and the readout uses $U_0 + \beta U_2$ and $S_0 + \beta S_2$.

Blockwise execution. Algorithm 1 implements the state-optimal construction by aligning and adding block summaries. Our Transformer kernels use tiled score computation, causal masking and fused value accumulation. The longer-context FP32 implementation uses two passes, first finding row maxima and then recomputing tiles for accumulation. The compact-model FP16 implementation evaluates the full model in a CUDA graph. Section 5 reports the corresponding learning and execution measurements.

5 Experiments

We evaluate three consequences of the coupled-state construction: learning attention shapes beyond positive mixtures, improving prediction with learned scores and values, and executing the resulting maps efficiently.

5.1 Experimental Setup

Controlled learning. We use Gaussian aggregation tasks with tokens $x_i = y + \epsilon_i$, target $y \sim \mathcal{N}(0, 1)$, and score-dependent independent noise. Exact group means supply the training inputs. Each map learns its parameters and a scalar readout from 32, 128, 512, 2048 or 8192 labeled examples, with 20 paired training seeds. The two-state task uses scores $(-2, -1, 0)$ and noise variance proportional to $\exp(-4t + 0.6t^2)$. The three-state task uses $(-8, -4, -2, -1, 0)$ and variance proportional to $[e^t(1 + 0.64t^2)]^{-1}$. Counts and the common variance scale set Bayes MSE to 0.5. We evaluate population risk analytically after fitting. Appendix A specifies the tasks, optimization and transfer grids.

Language modeling. We train a 5.510M-parameter decoder with four layers, four heads per layer, hidden size 256, and an 8192-token byte-level BPE vocabulary on an article-disjoint WikiText-derived corpus, optimizing with AdamW (Loshchilov & Hutter, 2019). Softmax, learned temperature, a positive two-exponential mixture (Mix-2), Ray and a two-state Coxian variant each receive 2^{27} (134M) target tokens under three paired seeds. Evaluation covers 262097 targets from 48 held-out articles at context lengths 128, 256 and 512. The final checkpoint of each run is evaluated through its FP16 deployment route. Training hyperparameters and the Coxian construction appear in Appendix B.

Execution. All GPU measurements use one RTX 4090 24 GB. Compact-model costs include the batch-one CUDA-graph forward and one-position vocabulary readout, using three independent-process rounds per seed. Softmax is evaluated with official FlashAttention-2, PyTorch Flash and

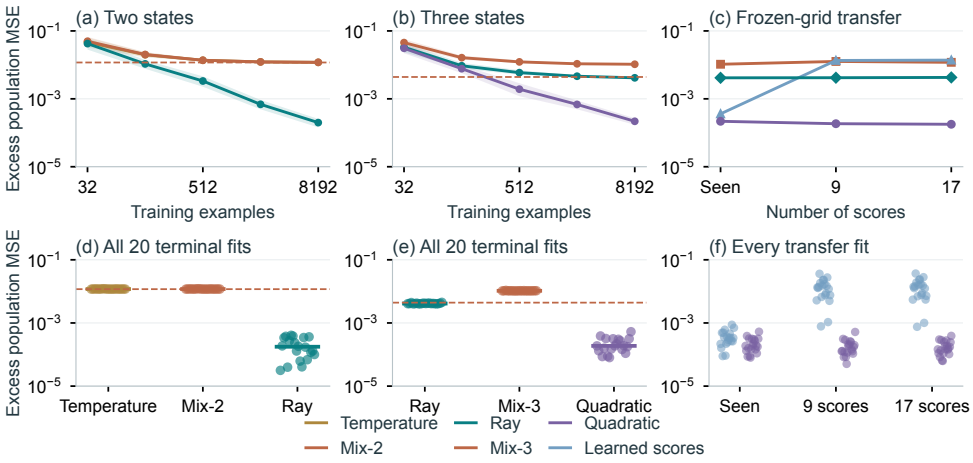


Figure 3: Coupled shapes learn and transfer. Top: excess population MSE with pointwise 95% bootstrap bands over 20 seeds, and frozen-grid transfer. Bottom: every terminal fit. Dashed lines mark positive-mixture risk floors.

the shared fused implementation. A separate longer-context study measures FP32 forward scoring on Qwen2.5-0.5B and SmolLM2-135M (Yang et al., 2024a; Allal et al., 2025), comparing dense, query-blocked and fused evaluation of the Ray map.

5.2 Learning and Unseen-Score Transfer

The finite-score certificate has a predictive consequence. In the Gaussian tasks, a positive-mixture weight vector satisfies linear inequalities inherited from exponential squares. Minimizing aggregate noise variance subject to one such inequality gives a population-risk lower bound, including an optimal scalar readout. Appendix A.2 derives this connection. For the two-state task we use the exact positive-mixture optimum.

Figure 3 shows learned coupled maps approaching Bayes risk as data increases. At 8192 examples, Ray reaches MSE 0.5002, below the two-state task's positive-mixture optimum of 0.5118. Mix-2 reaches 0.5119. The quadratic map reaches 0.5002 on the three-state task, below the positive-mixture lower bound of 0.5044. All 20 quadratic fits cross that bound.

The quadratic map also transfers its shape. On the nine-score grid, which contains eight unseen values, its MSE is 0.5002; on a denser seventeen-score grid it is again 0.5002. A softmax control with learned score features fits the training grid to 0.5004 and attains 0.5135 and 0.5138 on these transfer grids. The learned generator thus acts as a score-to-weight rule that is reused on new grids. The same fitted parameters and readout serve every grid.

Figure 4 examines what transfers. We reconstruct each fitted generator from its saved parameters and normalize it at score zero. The quadratic fits recover the inverse-noise weight function across the interval, including between training scores. Their induced group weights also follow the inverse-variance allocation on the unseen grid. Mix-3 and the learned-score control leave systematic allocation errors at intermediate scores. This functional view explains the transfer result: fitting the coupled shape recovers a weighting rule that extends across score locations.

5.3 Language Modeling

Ray improves mean held-out NLL over ordinary softmax at all three context lengths, with lower NLL in all nine paired seed–length comparisons (Figure 5). The mean differences are -2.480×10^{-3}, -2.455×10^{-3} and -3.095×10^{-3}, corresponding to perplexity reductions of 0.2477%, 0.2452% and 0.3090%. Figure 5 shows the individual paired seeds.

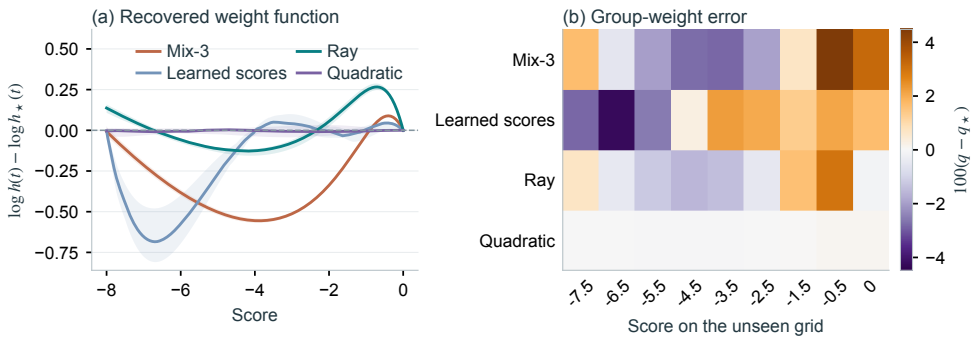


Figure 4: Frozen weighting rules between training scores. Left: normalized log-generator error, median/IQR over 20 fits. Right: mean nine-score weight error against inverse-variance weights. Both use 8,192-example fits.

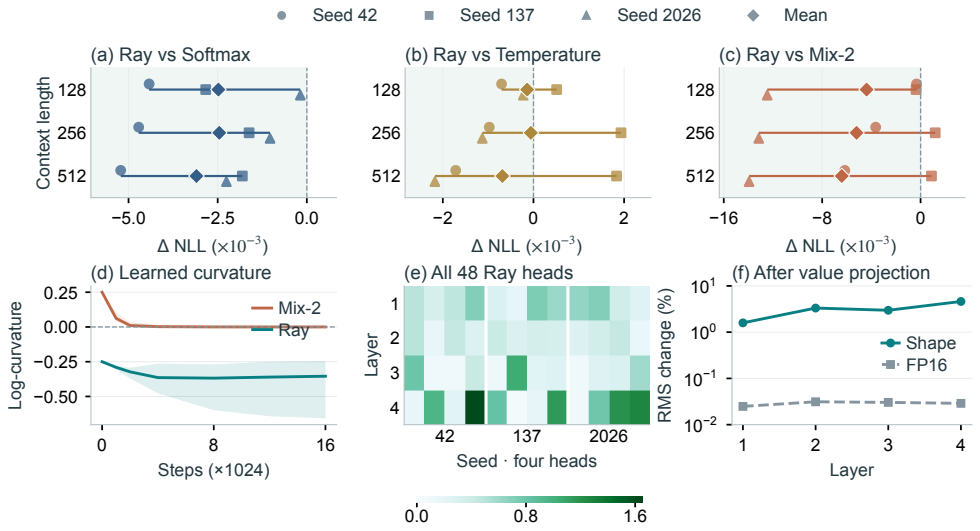


Figure 5: Prediction and learned shape. Top: paired Ray-minus-baseline NLL across three seeds. Bottom: log-curvature (median/IQR, 48 heads), final negative-curvature magnitudes, and output changes under temperature projection versus FP16 reconstruction on 16 training windows.

The controls separate state count from state organization. Temperature changes one exponential rate; Mix-2 retains two independent exponential components; Ray couples two moments. Table 1 includes the Coxian two-state variant. At length 512, Ray improves on Mix-2 by 6.406×10^{-3} NLL and on learned temperature by 6.859×10^{-4}. The 95% article-bootstrap interval for its mean difference from ordinary softmax at length 512 is $[-5.512, -0.6111] \times 10^{-3}$.

5.4 Learned Shape and Value Aggregation

We track shape in a slope-normalized coordinate, separating curvature from a change in temperature. Ray's median log-curvature moves from -0.2500 at initialization to -0.3545 after training. Mix-2 moves from 0.2500 toward 9.399×10^{-5}. Figure 5 shows the trajectory and all 48 trained Ray heads. The learned maps maintain a substantial curvature component throughout the model.

To measure the effect on actual attention outputs, we project each Ray weight row to its best positive-temperature distribution, then retain the original value vectors and output projection. Across the four layers of the seed-42 model, the relative output changes are 1.589%, 3.320%, 2.974% and 4.608%. These changes are 64.04–159.6 times the corresponding FP16 reconstruction error. The generator's shape difference persists through the operations that convert weights into model representations.

Table 1: Three-seed mean NLL changes, median forward latency and peak allocation. Bold/underlined NLL values rank first/second. Softmax and temperature use FA2; other maps use fused routes.

Map	ΔNLL $\downarrow$ ($\times 10^{-3}$)			Forward latency (ms) $\downarrow$			Memory
	128	256	512	128	256	512	512 (MiB)
Softmax	0	0	0	0.3457	0.3800	0.3862	40.53
Temperature	$\underline{-2.341}$	$\underline{-2.398}$	-2.409	0.3648	0.4007	0.4075	40.53
Mix-2	1.908	2.750	3.311	0.3340	0.3605	0.3853	40.53
Coxian	-2.206	-2.155	$\underline{-3.029}$	0.3364	0.3646	0.4017	40.53
Ray	$\mathbf{-2.480}$	$\mathbf{-2.455}$	$\mathbf{-3.095}$	0.3331	0.3587	0.3809	40.53

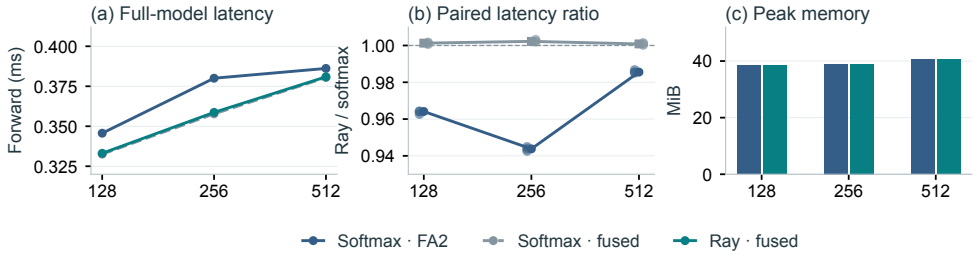


Figure 6: Batch-one FP16 execution with one-position vocabulary readout: latency, paired Ray/softmax ratios and peak allocation. Points summarize three seeds and three process rounds; memory includes model state.

The curvature, projection and value-aggregation measurements agree across successive stages of attention, carrying the distinction between coupled and independent states from the score-to-weight rule to the residual-stream vectors of the trained decoder.

5.5 Execution Cost and Backbone Scaling

Figure 6 reports complete compact-model forward costs. Ray's median paired latency ratios to the official FlashAttention-2 softmax route are 0.9641, 0.9438 and 0.9855 at lengths 128, 256 and 512. The shared fused softmax route isolates the map-dependent cost: the corresponding ratios are 1.001, 1.002 and 1.001. Ray and this fused control have identical peak allocations. At length 512, the full-model Ray forward takes 0.3809 ms and allocates 40.53 MiB.

The longer-context study evaluates a trained Ray generator in all 336 Qwen heads and a fixed generator in 270 SmolLM2 heads. Figure 7 combines the latency and memory results. On Qwen at 8192 tokens, fusion reduces forward scoring from 1360 to 521.1 ms relative to the fastest dense route, a $2.610\times$ speedup. Peak allocation falls from 13.22 GiB for dense materialization to 2.848 GiB; blocked dense uses 2.872 GiB. The scaling curves show why tiling reduces memory. Dense evaluation materializes score-dependent intermediates whose size grows with both query and key lengths. Blocking limits their lifetime to part of the sequence, while fusion evaluates the polynomial factor within score tiles and accumulates weighted values directly. On SmolLM2 at 4096 tokens, fusion attains $1.856\times$ the speed of the fastest dense route.

The full Qwen held-out comparison has a maximum absolute mean-NLL difference of 1.179×10^{-8} between execution routes, so the measured reduction in allocation accompanies the same normalized Ray output. Appendix C provides numerical fidelity, all timing samples, head certificates and continued-training results. With the state-optimal construction and the compact-model experiment, these measurements give both an exact merging rule and a tiled implementation of the coupled attention map.

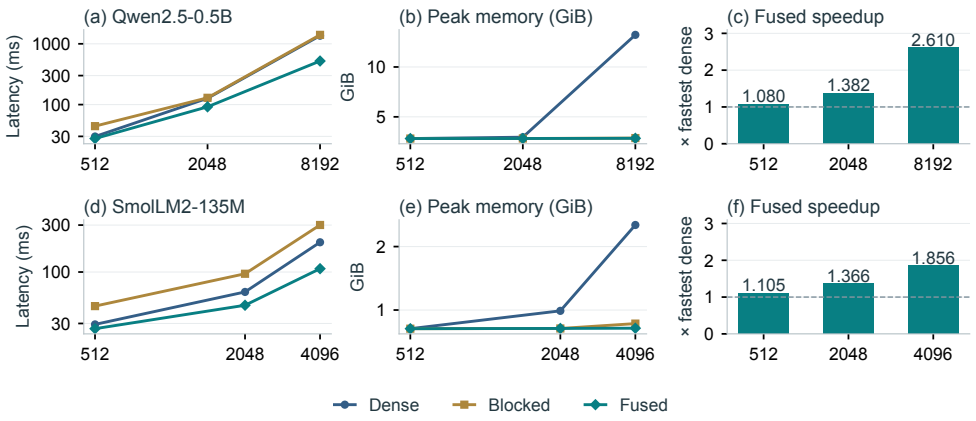


Figure 7: FP32 Ray execution on Qwen and SmolLM2: latency, peak allocation and speedup over the fastest dense route. Latencies are ten-repeat medians; memory includes the full model.

6 Related Work

Exact attention computation. In Transformers (Vaswani et al., 2017; Touvron et al., 2023), online normalization and FlashAttention organize softmax computation around compact, mergeable states (Milakov & Gimelshein, 2018; Dao et al., 2022; Dao, 2024; Shah et al., 2024); FlashInfer exposes state composition as a systems primitive (Ye et al., 2025b). Related systems organize context processing (Liu & Abbeel, 2023; Liu et al., 2023) and serving memory (Kwon et al., 2023; Zheng et al., 2024). Attention I/O bounds characterize the required cache transfers (Saha & Ye, 2024). Our continuation argument characterizes the value information retained by a fixed-query summary.

State structure and attention shape. Finite translation spaces yield exponential polynomials (Engert, 1970); positive finite-rank Hankel theory identifies positive exponential sums (Andersson & Carlsson, 2017). We derive an attaining merge rule, a matching value-state lower bound and finite mixture-exclusion certificates. Sigmoid attention studies alternative weights and efficient execution (Ramapuram et al., 2025); AdaSplash accelerates sparse attention (Gonçalves et al., 2025); Differential attention subtracts softmax maps (Ye et al., 2025a). Unlike these maps, coupled moments retain exact normalized merging. Single-location regression links activations to estimation risk (Duranthon et al., 2025); our controlled tasks show learned maps crossing certified positive-mixture risk bounds.

Compact sequence models use feature maps (Katharopoulos et al., 2020; Qin et al., 2022; Han et al., 2023), random features (Choromanski et al., 2021; Peng et al., 2021), low-rank approximations (Lu et al., 2021), gated recurrence (Peng et al., 2023; Sun et al., 2023; Ma et al., 2023; Yang et al., 2024b; Qin et al., 2023), and selective state spaces (Gu & Dao, 2023; Dao & Gu, 2024).

7 Conclusion

Exact attention merging extends beyond independent softmax components. Coupled moments attain the minimum value-state count, and a finite weight signature certifies their non-mixture structure. The learned instances turn this distinction into predictive gains in controlled estimation and language modeling, with practical blockwise execution. State count and state coupling jointly define an attention-design space.

Limitations

The classification concerns exact, fixed-query, value-linear summaries with real rates. Numerical conditioning at high degree or closely spaced rates, and evaluation across larger training scales and additional corpora, remain open directions.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlícek, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big - data-centric training of a small language model. *CoRR*, abs/2502.02737, 2025. doi: 10.48550/ARXIV.2502.02737. URL https://doi.org/10.48550/arXiv.2502.02737.

Fredrik Andersson and Marcus Carlsson. On the structure of positive semi-definite finite rank general domain hankel and toeplitz operators in several variables. *Complex Analysis and Operator Theory*, 11(4):755–784, 2017.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

O. Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. *CoRR*, abs/2509.21936, 2025. doi: 10.48550/ARXIV.2509.21936. URL https://doi.org/10.48550/arXiv.2509.21936.

Martin Engert. Finite dimensional translation invariant subspaces. *Pacific Journal of Mathematics*, 32(2):333–343, 1970.

Nuno Gonçalves, Marcos V. Treviso, and André F. T. Martins. Adasplash: Adaptive sparse flash attention. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/goncalves25a.html.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752, 2023. doi: 10.48550/ARXIV.2312.00752. URL https://doi.org/10.48550/arXiv.2312.00752.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *IEEE/CVF International Conference on Computer*

Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 5938–5948. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00548. URL https://doi.org/10.1109/ICCV51070.2023.00548.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis, 2nd Ed*. Cambridge University Press, 2012. ISBN 9780521548236. doi: 10.1017/CBO9781139020411. URL https://doi.org/10.1017/CBO9781139020411.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020. URL http://proceedings.mlr.press/v119/katharopoulos20a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Hao Liu and Pieter Abbeel. Blockwise parallel transformers for large context models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1bfd87d2d92f0556819467dc08034f76-Abstract-Conference.html.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *CoRR*, abs/2310.01889, 2023. doi: 10.48550/ARXIV.2310.01889. URL https://doi.org/10.48550/arXiv.2310.01889.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. SOFT: softmax-free transformer with linear complexity. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 21297–21309, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b1d10e7bafa4421218a51b1e1f1b0ba2-Abstract.html.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: Moving average equipped gated attention. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=qNLe3iq2El.

Maxim Milakov and Natalia Gimelshein. Online normalizer calculation for softmax. *CoRR*, abs/1805.02867, 2018. URL http://arxiv.org/abs/1805.02867.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan S. Wind, Stanislaw Wozniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: reinventing rnns for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 14048–14077. Association for Computational Linguistics, 2023. doi:

10.18653/V1/2023.FINDINGS-EMNLP.936. URL https://doi.org/10.18653/v1/2023.findings-emnlp.936.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. Random feature attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=QtTKTdVrFBB.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Bl8CQrx2Up4.

Zhen Qin, Songlin Yang, and Yiran Zhong. Hierarchically gated recurrent neural network for sequence modeling. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/694be3548697e9cc8999d45e8d16fe1e-Abstract-Conference.html.

Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russell Webb. Theory, analysis, and best practices for sigmoid self-attention. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=Zhdhg6n2OG.

Barna Saha and Christopher Ye. I/O complexity of attention, or how optimal is flashattention? In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43024–43042. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/saha24a.html.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7ede97c3e082c6df10a8d6103a2eebd2-Abstract-Conference.html.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *CoRR*, abs/2307.08621, 2023. doi: 10.48550/ARXIV.2307.08621. URL https://doi.org/10.48550/arXiv.2307.08621.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523. PMLR / OpenReview.net, 2024b. URL https://proceedings.mlr.press/v235/yang24ab.html.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=OvoCm1gGhN.

Zihao Ye, Lequn Chen, Ruihang Lai, Wuwei Lin, Yineng Zhang, Stephanie Wang, Tianqi Chen, Baris Kasikci, Vinod Grover, Arvind Krishnamurthy, and Luis Ceze. Flashinfer: Efficient and customizable attention engine for LLM inference serving. In Matei Zaharia, Gauri Joshi, and Yingyan (Celine) Lin (eds.), *Proceedings of the Eighth Conference on Machine Learning and Systems, MLSys 2025, Santa Clara, CA, USA, May 12-15, 2025*. OpenReview.net/mlsys.org, 2025b. URL https://openreview.net/forum?id=RXPofAsL8F.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark W. Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

Reproducibility statement

The appendices give full proofs, learning-task definitions, complete method and transfer grids, language-model results, and execution settings. Editable figure sources use frozen arrays and model-parameter certificates. The accompanying manifest records input hashes, model states, token sets and execution configurations.

AI-use disclosure

AI systems assisted with theoretical arguments, computation, literature discovery, manuscript drafting, figures and rendered-page inspection. The authors are responsible for the scientific content and its supporting evidence.

A Controlled Learning Tasks

A.1 Data model and optimization

Each example has a scalar target $y \sim \mathcal{N}(0,1)$ and tokens $x_{j,i} = y + \epsilon_{j,i}$ in score group j, with n_j tokens and independent noise $\epsilon_{j,i} \sim \mathcal{N}(0, \sigma_j^2)$. Its exact group mean is $X_j = y + \eta_j$, where

$\eta_j \sim \mathcal{N}(0, D_j)$ and $D_j = \sigma_j^2/n_j$. A normalized generator induces group weights

$$q_j = \frac{n_j h(t_j)}{\sum_k n_k h(t_k)}, \qquad \widehat{y} = b \sum_j q_j X_j. \tag{15}$$

The population MSE is $(b-1)^2 + b^2 \sum_j D_j q_j^2$. We train on sampled group means and labels, fitting generator parameters and a scalar readout. The group means are distributionally exact sufficient statistics of the independent Gaussian token model.

For the two-state task, the training length is 32768, the three scores are $(-2, -1, 0)$, and the noise law is $\sigma^2(t) \propto \exp(-4t + 0.6t^2)$. Counts are proportional to noise variance with deterministic rounding. For the three-state task, scores are $(-8, -4, -2, -1, 0)$, the length is 32768, and $\sigma^2(t) \propto [e^t(1 + 0.64t^2)]^{-1}$. Largest-remainder rounding assigns integer counts. A common variance scale makes $\sum_j 1/D_j = 1$, giving Bayes MSE 0.5.

Each task uses 20 independent training seeds. The five sample sizes are nested prefixes of 8192 examples within a seed, and all methods share those examples. For each nonlinear map, eight predetermined initializations feed L-BFGS-B with at most 400 iterations and 1600 function evaluations per start. Raw parameters lie in $[-12, 12]$; the function and gradient tolerances are 10^{-12} and 10^{-9}. The readout is profiled by least squares, and the lowest training objective selects the fitted parameters. Population risk is then computed from Equation (15). A separate 32768-example sample supplies the held-out MSE.

The two-state task compares softmax, learned positive temperature, Ray, positive two- and eight-exponential mixtures, and an unrestricted linear readout over the three group means. The three-state task compares softmax, temperature, Ray, the quadratic map, positive three- and eight-exponential mixtures, and softmax with a learned four-feature Chebyshev score transformation. All attention maps share the scalar readout.

A.2 Risk certificate

For equally spaced scores u, v, w and any positive exponential sum,

$$A^2 h(u) - 2Ah(v) + h(w) = \sum_k c_k e^{a_k u} \left(A - e^{a_k (v-u)}\right)^2 \geq 0. \tag{16}$$

After including group counts, this is a linear constraint $c^\top q \geq 0$ on normalized group weights. The nonzero entries of c are proportional to A^2/n_u, $-2A/n_v$ and $1/n_w$. Define

$$S_0 = \sum_j D_j^{-1}, \quad S_1 = \sum_j c_j D_j^{-1}, \quad S_2 = \sum_j c_j^2 D_j^{-1}. \tag{17}$$

When $S_1 < 0$, the inverse-variance optimum violates the halfspace. Minimizing $q^\top D q$ over the affine constraints $\mathbf{1}^\top q = 1$ and $c^\top q \geq 0$ yields

$$q_j^\star = \frac{\nu + \mu c_j}{D_j}, \quad \nu = \frac{S_2}{S_0 S_2 - S_1^2}, \quad \mu = \frac{-S_1}{S_0 S_2 - S_1^2} > 0. \tag{18}$$

For every feasible q,

$$q^\top D q - \nu = \sum_j D_j (q_j - q_j^\star)^2 + 2\mu c^\top q \geq 0. \tag{19}$$

Optimizing the scalar readout gives the risk bound

$$\mathrm{MSE}_{\mathrm{mixture}} \geq \frac{1}{1 + S_0 - S_1^2/S_2}. \tag{20}$$

The three-state task uses the witness $(-8, -4, 0)$, and the unseen grid uses $(-6.5, -3.5, -0.5)$. Setting $A = h(v)/h(u)$ from the specified task gives the evaluation certificate. The training procedure uses sampled inputs and labels. The training-grid bound is 0.5044, and the learned quadratic mean is 0.5002.

The two-state task uses the complete three-score positive-mixture projection, with its feasible stationary points and boundary checked numerically. Its attained risk is 0.5118; lower and upper evaluations agree within 10^{-8}. The learned Ray mean is 0.5002. These comparisons translate the finite-score shape constraint into an estimation-error gap.

A.3 Complete transfer grids

All generator parameters and readouts remain fixed after fitting. The two-state grid varies sequence length, count allocation and noise scale. The three-state grid adds a uniform allocation and two new score grids. Its nine-score grid contains $(-7.5, -6.5, \ldots, -0.5, 0)$; the seventeen-score grid is equally spaced on $[-8, 0]$. Figure 8 reports all methods and conditions at 8192 training examples. The doubled-noise setting also changes the optimal readout.

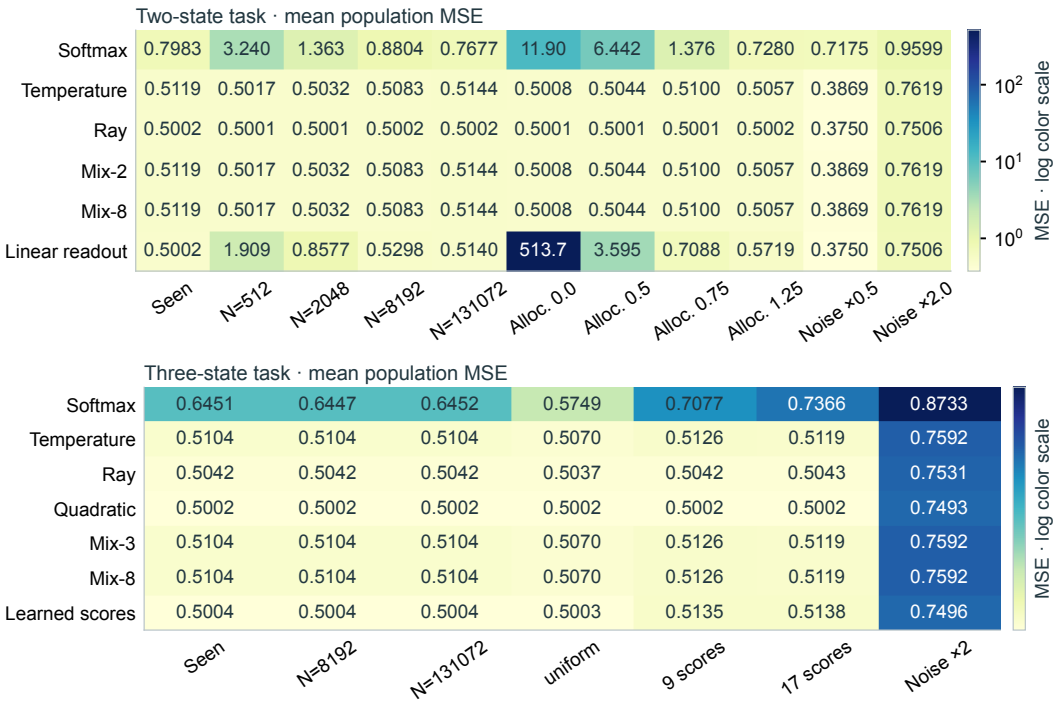


Figure 8: Complete transfer grids for two- and three-state tasks. Cells show mean population MSE over 20 seeds; color is logarithmic. Every fitted generator and readout is frozen.

B Language Modeling and Learned Geometry

Training. The decoder uses a Qwen2-style architecture with hidden size 256, intermediate size 768, four layers, four query/key/value heads, tied embeddings, and rotary position embeddings. The vocabulary has 8192 entries and is learned only from training articles. Article titles and hashes separate training, calibration and the held-out evaluation. Each run has 16384 updates at 8192 target tokens per update, using length 512 and microbatch 16. Paired seeds are 42, 137 and 2026.

AdamW (Loshchilov & Hutter, 2019) uses coefficients $(0.9, 0.95)$, a backbone learning rate of 6×10^{-4}, a generator learning rate of 0.003, and matrix weight decay 0.1. Bias, normalization and generator parameters have zero weight decay. Gradients are clipped at 1; 512 warm-up steps precede cosine decay to 0.1 of the peak learning rate. Training keeps FP32 weights and scores with BF16 linear/value products. Softmax and temperature start from identical functions; Ray starts at $(a, \lambda) = (1.5, 0.5)$ and Mix-2 at rates $(0.5, 1.5)$ with equal coefficients. Every method uses its final checkpoint.

Coxian variant. Writing $d = -t \geq 0$, the additional two-state family is

$$h(-d) = e^{-ad} + c\,\frac{e^{-bd} - e^{-ad}}{a - b}, \qquad a \geq c \geq 0, \quad b > 0, \tag{21}$$

with the continuous equal-rate limit $e^{-ad}(1 + cd)$. This sequential two-state form contains Ray at equal rates. We learn positive c, positive $a - c$ and the log ratio b/a; initialization matches the Ray function.

Evaluation. All methods use the same 48 held-out articles, target positions and context plans. The 262097 targets in each cell are evaluated through the same FP16 CUDA-graph route used for cost measurements. Softmax and temperature have official FlashAttention-2, PyTorch Flash and shared fused implementations; the other maps use their fused implementations. There are 81 quality cells. Table 2 gives every seed and method with the primary routes. Figure 5 displays paired differences without rounding the underlying NLLs. Article-bootstrap intervals resample the 48 articles using their paired loss contributions.

Table 2: Complete held-out NLL: five maps, three paired seeds and three context lengths.

	Context 128			Context 256			Context 512		
Map / seed	42	137	2026	42	137	2026	42	137	2026
Softmax	3.781	3.778	3.779	3.666	3.664	3.662	3.587	3.586	3.584
Temperature	3.777	3.775	3.779	3.663	3.661	3.662	3.584	3.582	3.584
Mix-2	3.777	3.776	3.791	3.665	3.661	3.675	3.588	3.583	3.595
Coxian	3.776	3.776	3.780	3.662	3.663	3.662	3.583	3.585	3.581
Ray	3.777	3.776	3.779	3.662	3.663	3.661	3.582	3.584	3.581

Shape coordinate. Let $\kappa = (\log h)'(0) > 0$ and $u = \kappa(m - s)$. In this coordinate, a slope-matched temperature has log weight $-u$. Figure 5 plots $(\log h)''(0)/\kappa^2$. For Ray it equals $-\lambda^2/(a-\lambda)^2$. The trajectory contains all 48 heads at the seven saved training steps. The heatmap uses the final checkpoint.

Value-space intervention. For the seed-42 model, we collect native score rows, values and output projections from 16 fixed training windows, four layers, four heads and four query positions. Each Ray distribution is replaced by its KL-optimal positive-temperature projection. We multiply both distributions by the same native values and output projection and measure relative output RMS. The FP16 reference is the reconstruction error, compared in the same output units. This intervention measures a representation change with the learned values held fixed.

C Execution and Existing-Backbone Results

C.1 Compact-model execution

The CUDA graph includes embedding lookup, all Transformer layers, final normalization and a vocabulary readout at one selected position. Every route uses the common feed-forward buffer-reuse implementation. Each seed/length/backend has three independent-process timing rounds, yielding 243 cells. Table 3 reports median wall time and peak allocated memory for all nine routes. The ratio panel in Figure 6 first takes a median across process rounds within each seed, then summarizes the paired seed ratios.

C.2 Qwen and SmolLM2

The FP32 study computes full model forwards, the language-model head and cross-entropy. Qwen uses a trained seed-42 Ray parameter file; SmolLM2 uses the fixed pair $(a, \lambda) = (1.1, 0.1)$ in all heads. The head-dimension-64 fused kernel is unchanged between backbones. Warm-up and compilation precede ten recorded timing repeats in each of 18 cells.

All 336 trained Qwen heads have positive λ and monotonicity margin $a - \lambda$. Their three-score normalized obstructions, i.e. $w_1^2 - w_0 w_2$ for the weights w_0, w_1, w_2 on scores $(0, -1, -2)$, range from 3.867×10^{-4} to 5.584×10^{-4}. SmolLM2's 270 head instances share one pair with obstruction 4.930×10^{-4}. Figure 9 gives the head-wise measurements.

The Qwen continuation study compares four maps on WikiText-2 with 2097152 targets per seed. Generator learning rates are selected on a calibration prefix. Across three paired seeds, all four methods have mean perplexity 13.17; Ray's mean NLL difference from softmax is -7.256×10^{-5}. Figure 10 shows the paired results.

Table 3: All compact-model forward routes by context length. Median latency and peak allocated memory over three seeds and three process rounds.

	128		256		512	
Forward route	ms	MiB	ms	MiB	ms	MiB
Softmax / FA2	0.3457	38.40	0.3800	39.02	0.3862	40.53
Softmax / Torch Flash	0.3514	38.40	0.3936	39.02	0.3884	40.53
Softmax / fused	0.3326	38.40	0.3577	39.02	0.3805	40.53
Temperature / FA2	0.3648	38.40	0.4007	39.02	0.4075	40.53
Temperature / Torch Flash	0.3713	38.40	0.4135	39.02	0.4100	40.53
Temperature / fused	0.3331	38.40	0.3593	39.02	0.3815	40.53
Ray / fused	0.3331	38.40	0.3587	39.02	0.3809	40.53
Mix-2 / fused	0.3340	38.40	0.3605	39.02	0.3853	40.53
Coxian / fused	0.3364	38.40	0.3646	39.02	0.4017	40.53

Table 4: Full-model FP32 forward time in milliseconds. Speed ratios use the fastest dense route in each row. Best and second-best times are bold and underlined.

Model	Length	Dense	Blocked	Fused	Speed ratio
Qwen 0.5B	512	<u>29.84</u>	44.23	**27.62**	1.080
Qwen 0.5B	2048	<u>127.0</u>	129.4	**91.93**	1.382
Qwen 0.5B	8192	<u>1360</u>	1404	**521.1**	2.610
Smol 135M	512	<u>29.34</u>	45.04	**26.55**	1.105
Smol 135M	2048	<u>62.74</u>	96.02	**45.94**	1.366
Smol 135M	4096	<u>200.7</u>	300.7	**108.1**	1.856

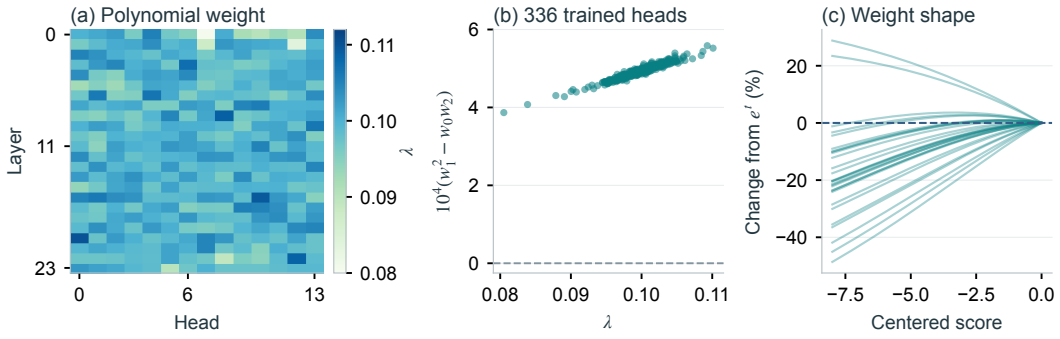


Figure 9: Ray in the continued-trained Qwen model. Left: polynomial coefficient by layer and head. Center: positive three-score mixture obstruction for every head. Right: generator change from a unit-rate exponential, sampling every twelfth head in layer-major order.

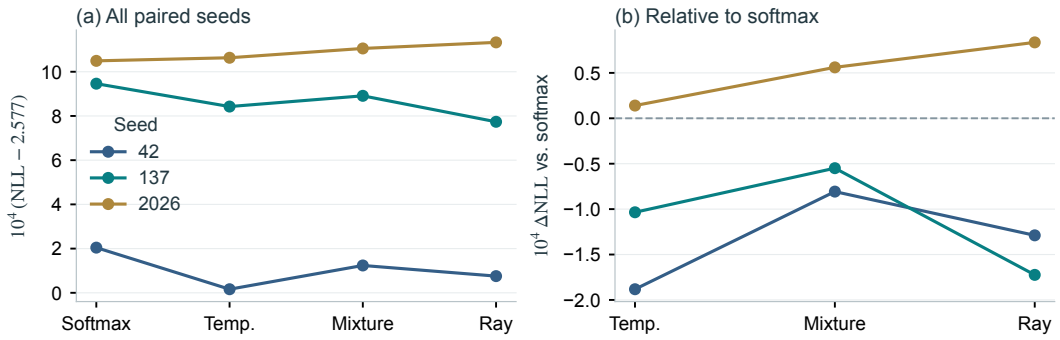


Figure 10: Qwen continued training. Left: NLL from every method and paired seed. Right: changes from ordinary softmax. Colors retain seed identity.

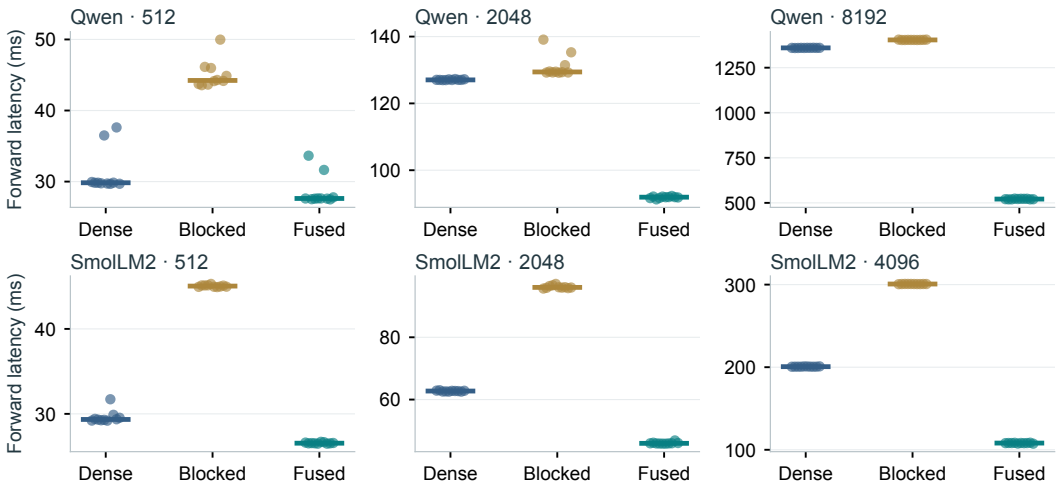


Figure 11: All 180 FP32 full-forward timing samples: two backbones, three lengths, three implementations and ten repeats. Dots are individual wall times; horizontal strokes are medians.

C.3 Numerical fidelity

The initial Qwen calibration contains 20 backend pairs, the compiled-control calibration 24, and the SmolLM2 calibration 24. Fused logit normalized RMS errors are of order 10^{-6}. Full held-out Qwen evaluation uses 262144 targets at each context length, with dense, blocked and fused routes at 512 and 2048, and blocked and fused routes at 8192. Table 5 reports the paired numerical differences.

Table 5: Paired Qwen test losses. Deltas are candidate minus reference; RMS is over per-token loss differences.

Length	Reference	Candidate	Mean NLL Δ	Max abs.	RMS
512	Dense	Blocked	0	0	0
512	Dense	Fused	3.833×10^{-9}	6.819×10^{-5}	4.259×10^{-6}
2048	Dense	Blocked	-2.934×10^{-9}	9.179×10^{-5}	4.978×10^{-6}
2048	Dense	Fused	-1.179×10^{-8}	1.831×10^{-4}	5.005×10^{-6}
8192	Blocked	Fused	-1.163×10^{-8}	7.749×10^{-5}	5.541×10^{-6}

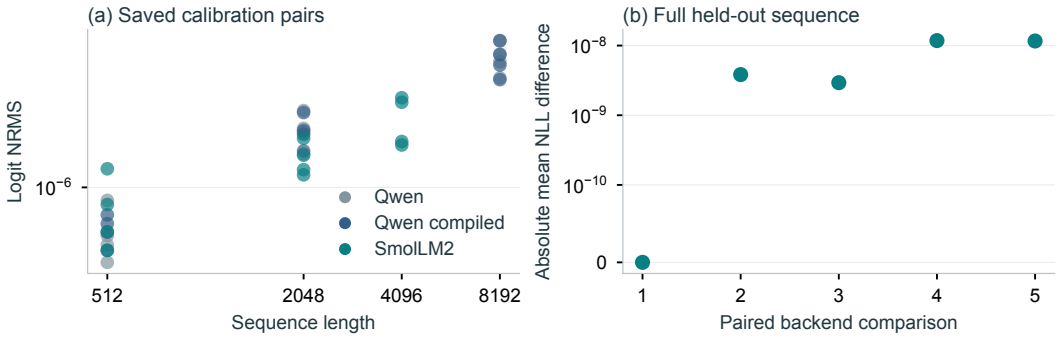


Figure 12: Numerical fidelity. Left: all saved fused-versus-dense calibration logit errors. Right: absolute mean NLL differences for the paired held-out routes.

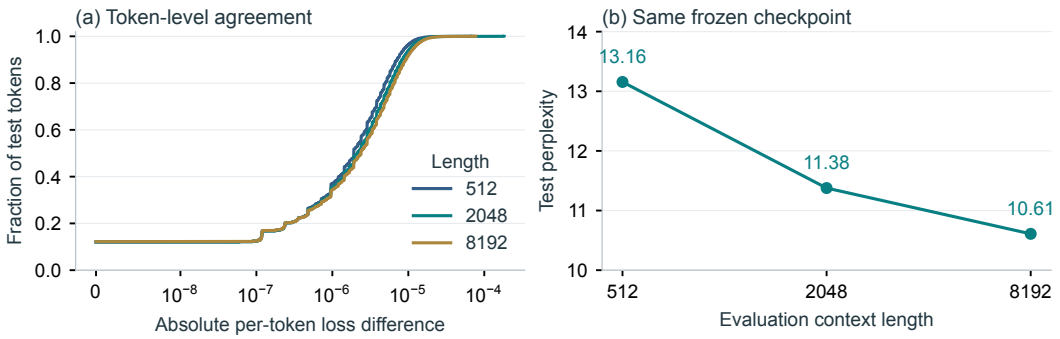


Figure 13: Frozen Qwen checkpoint at longer contexts. Left: cumulative distributions of paired token-loss differences. Right: full-test perplexity at lengths 512, 2048 and 8192.

D Additional constructive details

Algorithm 1 Merge two maximum-centered moment summaries

Require: Nonempty block states (m_A, S^A, U^A) and (m_B, S^B, U^B)

$M \leftarrow \max(m_A, m_B)$; set every merged $S_{k,p}, U_{k,p}$ to zero
for $C \in \{A, B\}$ **do**
  $\delta \leftarrow m_C - M$
  for each rate k and degree $p = 0, \dots, d_k$ **do**
    $S_{k,p} \leftarrow S_{k,p} + e^{a_k \delta} \sum_{j=0}^{p} \binom{p}{j} \delta^{p-j} S^C_{k,j}$
    $U_{k,p} \leftarrow U_{k,p} + e^{a_k \delta} \sum_{j=0}^{p} \binom{p}{j} \delta^{p-j} U^C_{k,j}$
  end for
end for
return (M, S, U); decode using Equation (6) when requested

D.1 Coefficient Block Inertia

The signature can be computed before evaluating any eigenvalues. Set $u_k = e^{-a_k x}$ and $V_{i,(k,p)} = i^p u_k^i$. Expanding the polynomial in $i + j$ gives the congruence

$$H_R = V \left(\bigoplus_k G^{(k)} \right) V^\top, \qquad G^{(k)}_{pr} = c_{k,p+r} (-x)^{p+r} \binom{p+r}{p}, \tag{22}$$

Coefficients beyond d_k are zero. Distinct real rates make V nonsingular for every positive spacing. Sylvester's law of inertia then transfers the positive and negative counts of the coefficient blocks to the matrix built from weights.

A degree-d coefficient block has determinant

$$\det G = c_d^{d+1} x^{d(d+1)} (-1)^{d(d+1)/2} \prod_{p=0}^{d} \binom{d}{p} \neq 0. \tag{23}$$

Shrinking all lower coefficients continuously to zero leaves this determinant unchanged, so no eigenvalue crosses zero. The remaining anti-diagonal matrix has one positive and one negative direction per off-center pair. When d is even, its central direction has sign c_d. This proves Equation (12): changing lower coefficients cannot remove the negative directions contributed by a nonconstant block.

This proof also constructs a witness. The principal block on indices $0, d$ has determinant $-c_d^2 x^{2d} < 0$. Embed one of its negative directions as z in coefficient space and solve $V^\top y = z$. Then $y^\top H_R y < 0$ excludes every mixture in Equation (10). The full signature is stronger than a determinant sign, which is positive when the number of negative eigenvalues is even.

For the converse, let $h(t) = \sum_k A_k e^{a_k t}$ with $A_k > 0$, and set $Z_k = \sum_i e^{a_k t_i}$. The list-level gates

$$\pi_k(s) = \frac{A_k Z_k}{\sum_\ell A_\ell Z_\ell} \tag{24}$$

are nonnegative, sum to one and reproduce the attention weights. This proves mixture membership and completes the theorem. The PSD/positive-exponential-sum correspondence itself is classical (Andersson & Carlsson, 2017); the block calculation here gives its explicit inertia and realizes the obstruction on a finite attention list.

D.2 Irregular Score Spacing

The signature does not depend on equal spacing. At any R distinct nonnegative offsets $y_0, \dots, y_{R-1}$, define

$$K_{ij} = h(-(y_i + y_j)), \qquad \widetilde{V}_{i,(k,p)} = y_i^p e^{-a_k y_i}. \tag{25}$$

Expansion yields $K = \widetilde{V}(\bigoplus_k G^{(k)}(1))\widetilde{V}^\top$. A nonzero real-rate exponential polynomial of representation length R has at most $R-1$ distinct real zeros. Thus $\widetilde{V}$ is nonsingular and K has the same inertia. Appendix H.3 proves this real-rate Chebyshev property by differentiating away one rate block and applying Rolle's theorem.

To read K from attention weights, place the scores $-(y_i + y_j)$ on one list and add a zero-score anchor if needed. Each entry of K is then a listed weight times the common positive normalizer. A locked mixture again gives a sum of PSD outer products. The anchor matters: without it, strictly positive offsets would induce a different maximum and hence different generator arguments.

The matrix size remains R in either construction; an arithmetic progression reduces the required score positions to $2R-1$. Irregular offsets use the distinct pairwise sums and, when necessary, the anchor. Spacing changes conditioning and eigenvalue magnitudes, but not the exact signature within the real-rate class.

E Legal generators outside the mixture class

Table 6 gives positive, own-score-monotone attention families in the non-mixture region on the entire negative half-line. The examples isolate different reasons why state cost and mixture structure separate; Appendix J verifies their legality.

Table 6: Legal non-mixture generators. State count R and inertia $(n_+, n_-, 0)$ are exact for every $x > 0$. All parameter bounds suffice; the bounds on a (rows 1–2) and the upper bounds on c (rows 3–4) are exact monotonicity boundaries. In the first row, $r_+ = \lceil (p+1)/2 \rceil$ and $r_- = \lfloor (p+1)/2 \rfloor$.

Generator $h(t)$	Conditions	R	n_+	n_-
$e^{at}(1-\lambda t)^p$	$p \in \mathbb{N}_{>0},\ \lambda > 0,\ a \geq p\lambda$	$p+1$	r_+	r_-
$e^{at}(1+\beta t^2)$	$\beta > 0,\ a \geq \sqrt{\beta}$	3	2	1
$e^t - ce^{2t}$	$0 < c \leq 1/2$	2	1	1
$e^t + e^{2t} - ce^{3t}$	$0 < c \leq 1$	3	2	1
$e^t + te^{2t}$	No free parameter	3	2	1

Polynomial degree and state cost. For the polynomial ray $e^{at}(1-\lambda t)^p$, the logarithmic derivative $a - p\lambda/(1-\lambda t)$ is minimized at zero. Hence $a \geq p\lambda$ is the exact monotonicity boundary, and the degree forces $p+1$ channels. Setting $\lambda = 0$ collapses the polynomial to a constant and the rank to one. This degeneration explains why minimality must be checked before counting degrees.

The quadratic family is constrained away from the maximum. Its logarithmic derivative reaches $a - \sqrt{\beta}$ at $t = -1/\sqrt{\beta}$. A tied-maximum check would miss this bottleneck. Thus rank follows from the minimal representation, while legality requires a separate half-line calculation.

E.1 Three-Point Tests

A necessary mixture inequality on the three scores $0, -x, -2x$ is

$$\Delta_2(x) := h(-x)^2 - h(0)h(-2x) \le 0. \tag{26}$$

This is the negative determinant of the leading 2×2 Hankel matrix. A positive value rules out a mixture; a nonpositive value leaves the question open. Figure 14 separates failures caused by spacing from failures caused by matrix size.

For $h(t) = e^t(1 + t^2)$, $\Delta_2 = e^{-2x}x^2(x^2 - 2)$. Spacings up to $\sqrt{2}$ therefore fail to detect an obstruction, even though the full H_3 is indefinite for every $x > 0$. Replacing $1 + t^2$ by $61 + t^2$ moves the threshold to $\sqrt{122}$: integer spacings one through eleven miss it, and twelve detects it. Sampling several small spacings cannot make the three-point test complete.

The signed three-rate family fails for a different reason. With $v = e^{-x} \in (0, 1)$, expansion gives

$$\Delta_2 = v^2(1 - v)^2\big[-(1 - c) + 2cv + 2cv^2\big], \tag{27}$$

$$\det H_3 = -c\, v^8(v - 1)^6(v + 1)^2 < 0. \tag{28}$$

For $0 < c \le 1/5$, the bracket is nonpositive for all spacings. Refining or widening the grid cannot help an anchored three-point test, but every five-token equally spaced list gives a negative direction in H_3. The missing resource is witness size, not sampling resolution. Increasing c can make three points informative without changing the full rank or inertia.

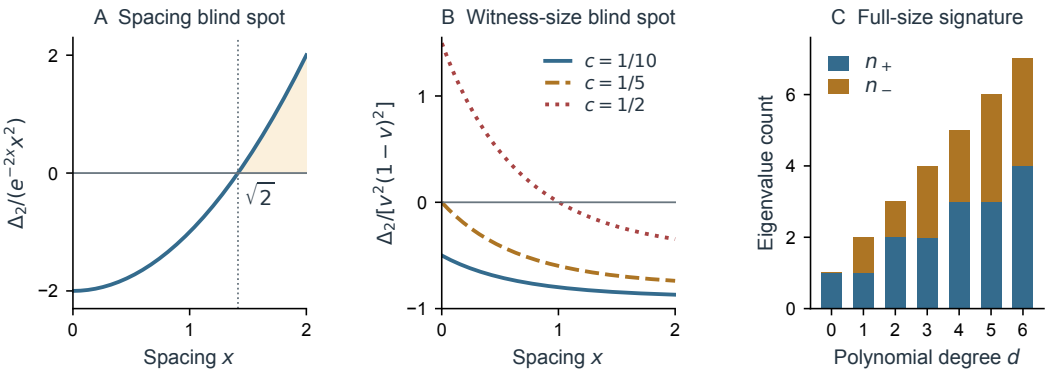


Figure 14: Finite mixture certificates. (A) Quadratic obstruction above spacing $\sqrt{2}$. (B) Signed three-rate family, $v = e^{-x}$. (C) Exact Hankel inertia for one degree-d polynomial block; every $d \ge 1$ contributes a negative direction.

Five-token certificate. At unit spacing the quadratic generator passes the leading three-point inequality, with $\Delta_2 = -e^{-2} < 0$. Its larger coefficient block nevertheless supplies an explicit negative direction:

$$G = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad z = (1, 0, -1)^\top, \qquad z^\top G z = -1. \tag{29}$$

Solving $V^\top y = z$ gives $y = (1/2, e, -e^2/2)^\top$ and $y^\top H_3(1) y = -1$. Form the Hankel matrix from the five normalized weights on $(0, -1, -2, -3, -4)$, and write $Z = \sum_{j=0}^{4} h(-j)$. The same vector gives quadratic form $-1/Z < 0$, which every locked nonnegative mixture forbids. Thus five observable weights give a certificate even though their first three entries pass the smaller inequality. The quadratic form is negative while all token weights are positive.

Signed components. The generator $h(t) = e^t + te^{2t}$ makes the distinction visible at the component level. Its faster term is negative for $t < 0$, but $h = e^t(1 + te^t) > 0$ since $te^t \ge -1/e$. Also $h' = e^t[1 + e^t(1 + 2t)] > 0$, since the bracket is at least $1 - 2e^{-3/2}$. It is a legal rank-three map.

F DIAGNOSTICS, SCOPE AND DESIGN IMPLICATIONS

Algebraic checks. The CPU checks agree with the coefficient factorizations, continuation determinants, Pascal composition, generalized-offset kernels and lower-degree identities in integer or rational arithmetic. Five additional constructor cases cover two continuation determinants, one coefficient-block determinant, one Hankel factorization and one nonuniform-offset kernel. These checks exercise the formulas used by both the merge construction and the finite-list certificate.

Numerical implementation checks. The saved direct-sum implementation check uses five fixed generator families and three-dimensional values. For each family, eight deterministic lists include singletons, ties and separated maxima; the seven-token list also compares both parenthesizations of a three-block merge. Four pseudorandom length-eight lists per family compare moment decoding with dense evaluation. The largest discrepancy across 70 scalar maximum-error comparisons is 2.665×10^{-15}. Appendix L specifies their shared inputs and exact accounting.

Rank diagnostics and complex rates. Finite translation spaces also include complex spectra. Trigonometric polynomials have exponents on the imaginary axis, and a sampling grid can alias their oscillations. The positive, own-score-monotone generator $h(t) = e^t(2 + \cos t)$ gives an explicit example: its oscillatory block collapses to a single exponential on every full-period grid. Appendix K gives the identity. This distinguishes the real-rate hypothesis of the all-spacing signature theorem from finite rank, positivity and monotonicity, and separates spectral aliasing from the witness-size effects above.

Generator and state selection. For a generator already in the real-rate class, first check positivity and any desired monotonicity on the whole centered domain. Combine repeated rates and remove zero polynomials; the remaining degrees give the exact value-state budget, and Algorithm 1 attains it. The block signature then decides the separate mixture question. A larger budget can retain coupled moments of one polynomial factor.

Relation to exact-attention algorithms. FlashAttention-2 retains an unnormalized value sum during online updates and normalizes at readout (Dao, 2024). FlashInfer composes partial outputs with their log-normalization scales (Ye et al., 2025b). Both make scalar-exponential merging a concrete computational primitive. Here the same fixed-query operation is carried by multiple, possibly coupled moments. Saha and Ye count cache transfers under explicit computational models (Saha & Ye, 2024); our lower bound counts the value information retained by a fixed-query summary. These resource models address complementary levels of exact attention computation.

Classical finite-dimensional translation theory includes exponential-polynomial representations, including complex spectra (Engert, 1970). The continuous finite-rank PSD Hankel classification also already gives positive exponential sums and polynomial obstructions (Andersson & Carlsson, 2017). The operational result here is the matching storage requirement exposed by continuations; the explicit finite-list congruence supplies its complementary mixture test.

G OWN-SCORE MONOTONICITY, INCLUDING TIES

Let $g = h'/h$ and $Z = \sum_j h(t_j)$. If token i is not a maximum, changing only s_i leaves the centering maximum unchanged locally. The quotient rule gives

$$\frac{\partial w_i}{\partial s_i} = \frac{h'(t_i)(Z - h(t_i))}{Z^2} = w_i(1 - w_i)g(t_i). \tag{30}$$

If it is the unique maximum, its own centered argument is zero and all other arguments decrease at unit rate. Hence

$$\frac{\partial w_i}{\partial s_i} = \frac{h(0)\sum_{j\neq i} h'(t_j)}{Z^2} = w_i \sum_{j\neq i} w_j g(t_j). \tag{31}$$

At a tie for the maximum, a sufficiently small decrease leaves another token at the maximum, so the left derivative is the nonmaximum formula with $t_i = 0$. A sufficiently small increase makes i the unique maximum, so the right derivative is the second formula, including the other tied tokens

among $j \neq i$. Both are nonnegative if $h' \geq 0$. For a fixed list of other scores, the real line is partitioned at their finitely many distinct values; on each smooth piece the derivative is nonnegative, and weights are continuous across the partition points. Thus w_i is globally nondecreasing. A singleton has constant weight one.

For necessity, choose the two-token list $(t, 0)$ with $t < 0$ and vary the first score. Its derivative is $h'(t)h(0)/(h(t) + h(0))^2$, whose sign is the sign of $h'(t)$. Monotonicity therefore forces $h'(t) \geq 0$ for every negative t. Continuity of h' gives the same at zero, completing the proof of Lemma 1.

H Translation algebra and the exact lower bound

H.1 Pascal composition and decoding

For $0 \leq j \leq p \leq d$, multiplication of two Pascal matrices gives

$$[T_d(\delta_1)T_d(\delta_2)]_{pj} = \sum_{r=j}^{p} \binom{p}{r}\binom{r}{j}\delta_1^{p-r}\delta_2^{r-j} \tag{32}$$

$$= \binom{p}{j}\sum_{r=j}^{p}\binom{p-j}{r-j}\delta_1^{p-r}\delta_2^{r-j} = \binom{p}{j}(\delta_1+\delta_2)^{p-j}. \tag{33}$$

Entries above the diagonal vanish. Combining this identity with $e^{a\delta_1}e^{a\delta_2} = e^{a(\delta_1+\delta_2)}$ proves the shift composition rule for every moment block. Each token's contribution in an arbitrary merge tree is shifted by the sum of changes in its ancestors' maxima, which telescopes to its leaf-block maximum minus the final maximum, so its centered argument becomes its score minus the final maximum. Addition then gives exactly Equation (4) for the concatenated list, independently of tree shape. Substituting the polynomial coefficients into those sums gives Equation (6). A separate empty flag provides a two-sided identity for the merge operation.

H.2 Confluent determinant with explicit factors

Use the order of rates and increasing degree within each rate for all columns. Set $n_k = d_k + 1$, $C_k = n_k(n_k-1)/2$, $z_k = e^{-a_k}$, and choose $x_j = -j$, $b_\ell = \ell + 1$ for $j, \ell = 0, \ldots, R-1$. Taylor expansion of the polynomial is exact:

$$q_k(x-b) = \sum_{p=0}^{d_k} x^p \frac{q_k^{(p)}(-b)}{p!}. \tag{34}$$

Thus $P = BX^\top$ as in Equation (9). The determinants are

$$\det X = \prod_k \left[(-z_k)^{C_k}\prod_{p=0}^{d_k} p!\right]\prod_{k<\ell}(z_\ell - z_k)^{n_k n_\ell}, \tag{35}$$

$$\det B = \prod_k \left[c_{k,d_k}^{n_k} z_k^{C_k+n_k}\prod_{p=0}^{d_k}\binom{d_k}{p}p!\right]\prod_{k<\ell}(z_\ell - z_k)^{n_k n_\ell}, \tag{36}$$

$$\det P = \prod_k \left[c_{k,d_k}^{n_k}(-1)^{C_k} z_k^{n_k^2}\prod_{p=0}^{d_k}\binom{d_k}{p}(p!)^2\right]\prod_{k<\ell}(z_\ell - z_k)^{2n_k n_\ell}. \tag{37}$$

To verify every factor, start with the derivative-column confluent Vandermonde $D_{j,(k,p)} = \partial_{z_k}^p z_k^j = (j)_p z_k^{j-p}$. Its determinant is $\prod_k \prod_{p=0}^{d_k} p!$ times the cross-node product in Equation (35). There is no additional power of z_k in this derivative-basis determinant. For example, one block of size two gives $\left(\begin{smallmatrix}1 & 0\\ z & 1\end{smallmatrix}\right)$, of determinant one.

The conversion from falling factorials to ordinary powers is triangular:

$$j^p z^j = \sum_{r=0}^{p}\left\{{p \atop r}\right\} z^r \partial_z^r z^j, \tag{38}$$

where braces denote Stirling numbers of the second kind. The diagonal factors are z^p, contributing $z_k^{C_k}$ per block. Replacing j^p by $(-j)^p$ contributes $(-1)^{C_k}$ and proves Equation (35).

For B, first use the basis $b^r z_k^b$ on $b = 1, \ldots, R$. Shifting the row variable from j to $j+1$ induces a triangular polynomial change of determinant one and one factor z_k in each of its n_k columns. This supplies $z_k^{C_k+n_k}$, together with the same factorial and cross-node products. The polynomial in its degree-p column is $q_k^{(p)}(-b)/p!$, whose leading term is $c_{k,d_k}\binom{d_k}{p}(-1)^{d_k-p}b^{d_k-p}$. Reverse the degree order inside this block. The reversal sign is $(-1)^{C_k}$ and the product of the leading signs is also $(-1)^{C_k}$, so they cancel. Lower-degree coefficients do not affect the determinant of this triangular change of basis. This proves Equation (36). Multiplication and $2C_k + n_k = n_k^2$ give Equation (37).

Every factor is nonzero: real distinct rates give distinct positive z_k, and minimality gives $c_{k,d_k} \neq 0$. Hence P is invertible. For a positive generator, Lemma 2 transfers this rank to the state lower bound. Positivity is used for the attention normalization; the determinant identity itself is polynomial-algebraic and remains valid for coefficient choices that do not define attention.

H.3 Evaluation on arbitrary distinct real nodes

A nonzero exponential polynomial with distinct real rates and representation length R has at most $R-1$ distinct real zeros. Here is a direct induction. The single-rate case reduces to a polynomial after multiplication by a nonvanishing exponential. For more rates, multiply by $e^{-a_1 t}$ and differentiate $n_1 = d_1 + 1$ times. The first polynomial vanishes. Every other block becomes $e^{(a_k-a_1)t}(\mathrm{D} + a_k - a_1)^{n_1} q_k(t)$, whose polynomial has the same degree and nonzero leading coefficient because $a_k - a_1 \neq 0$. If the original function had R distinct zeros, repeated Rolle's theorem would give at least $R - n_1$ zeros of this derivative, contradicting the inductive bound $R - n_1 - 1$. Therefore any square evaluation matrix of this basis at R distinct real nodes is nonsingular. Real rates are necessary for this argument; oscillatory complex exponentials have different zero behavior.

I Hankel factorization and inertia

For one block, expand by total degree:

$$q(-x(i+j)) = \sum_{m=0}^{d} c_m (-x)^m \sum_{p=0}^{m} \binom{m}{p} i^p j^{m-p}. \tag{39}$$

Reading off the coefficient of $i^p j^r$ gives G_{pr} in Equation (22). Multiplying the i and j basis terms by u_k^i, u_k^j and summing over rates proves the full factorization. The determinant of V is

$$\det V = \prod_k \left[u_k^{C_k} \prod_{p=0}^{d_k} p! \right] \prod_{k<\ell} (u_\ell - u_k)^{n_k n_\ell}, \tag{40}$$

which is nonzero for all $x > 0$.

Within G, entries below the anti-diagonal vanish and anti-diagonal entries equal $c_d(-x)^d\binom{d}{p}$. Thus its determinant is the product of those entries times the reversal sign, yielding Equation (23). The lower coefficients only affect entries strictly above that anti-diagonal. Set these coefficients to $(1-\theta)c_p$, $0 \leq \theta \leq 1$, keeping c_d fixed. The determinant is nonzero along the entire path, so no eigenvalue crosses zero. At $\theta = 1$, positive diagonal rescaling turns the absolute binomial anti-diagonal into the exchange matrix of size $d+1$. Each off-center pair contributes one positive and one negative direction. For even d, the center contributes the sign of c_d; for odd d, there is no center. This proves all three cases of Theorem 2, including the semisimple case $d = 0$.

The same signature holds without equal spacing. For distinct nonnegative offsets $y_0, \ldots, y_{R-1}$,

$$K = \left[h(-(y_i + y_j))\right]_{i,j=0}^{R-1} = \widetilde{V}\left(\bigoplus_k \Gamma^{(k)}\right)\widetilde{V}^\top, \quad \widetilde{V}_{i,(k,p)} = y_i^p e^{-a_k y_i}, \tag{41}$$

where $\Gamma^{(k)}_{pr} = c_{k,p+r}(-1)^{p+r}\binom{p+r}{p}$ is the block at $x = 1$. Appendix H.3 makes $\widetilde{V}$ nonsingular, so the inertia is unchanged. On the list of pairwise negative sums with an added zero anchor when necessary, the normalization again scales the matrix positively. Equal spacing is especially economical because it needs only $2R - 1$ score positions.

Negative directions versus determinant signs. For $d \geq 1$, the block on indices $0, d$ is $\begin{pmatrix} c_0 & c_d(-x)^d \\ c_d(-x)^d & 0 \end{pmatrix}$, which has determinant $-c_d^2 x^{2d}$. If there is a negative constant-coefficient block instead, its coordinate vector is already a negative direction. Embedding either vector in the full coefficient space and solving $V^\top y = z$ supplies the negative quadratic form claimed in the main text. For a positive generator, $H_{00} = h(0) > 0$, so a negative direction makes H indefinite. Without positivity of the generator, a matrix with only negative semisimple coefficients could be negative definite; the signature formula still applies, although the matrix no longer comes from attention weights.

The example $h(t) = e^t - \frac{1}{10}e^{2t} - \frac{1}{10}e^{3t}$ illustrates why the full inertia matters. With $u = e^t \in (0, 1]$, its positive factor is $1 - u/10 - u^2/10 \geq 4/5$ and its derivative factor is $1 - u/5 - 3u^2/10 \geq 1/2$. It is therefore legal. Its representation-sized Hankel inertia is $(1, 2, 0)$, so its determinant is positive despite its two negative directions.

J Legality and finite witnesses for the examples

Polynomial ray. For $p \geq 1$ and $\lambda > 0$, $(1 - \lambda t)^p > 0$ on $t \leq 0$. Its logarithmic derivative after multiplication by e^{at} is $a - p\lambda/(1 - \lambda t)$, minimized at zero. Lemma 1 gives the necessary and sufficient condition $a \geq p\lambda$. For any $x > 0$,

$$\Delta_2 = e^{-2ax}\left[(1 + \lambda x)^{2p} - (1 + 2\lambda x)^p\right] > 0, \tag{42}$$

because $(1+\lambda x)^2 > 1+2\lambda x$. This family is detected already by three points. Its leading coefficient has sign $(-1)^p$, which gives the table's inertia: for odd p the two counts are equal, and for even p the leading coefficient is positive.

Quadratic family. For $h = e^{at}(1+\beta t^2)$ and $\beta > 0$, the logarithmic derivative is $a+2\beta t/(1+\beta t^2)$. Setting $u = -\sqrt{\beta}\, t \geq 0$, the second term is $-2\sqrt{\beta}\, u/(1 + u^2)$, whose minimum is $-\sqrt{\beta}$ at $u = 1$. Thus legality is exactly $a \geq \sqrt{\beta}$. More generally, $h = e^{at}(L + t^2)$ is legal when $L > 0$ and $a \geq 1/\sqrt{L}$. Its three-point quantity is $e^{-2ax}x^2(x^2 - 2L)$. At $L = 61$, the polynomial factor at $x = 12$ is $144(144 - 122) = 3168$, while the factor is negative for integer $x = 1, \ldots, 11$. Its full rank is three and its full signature is $(2, 1, 0)$ at every positive spacing.

Signed two-rate family. For $h = e^t - ce^{2t}$ and $0 < c \leq 1/2$, writing $u = e^t$ gives $h = u(1-cu) > 0$ and $h' = u(1 - 2cu) \geq 0$. The witness is $\Delta_2 = c(e^{-x} - e^{-2x})^2 > 0$. The two distinct rates have coefficient signs $+, -$, so rank two and inertia $(1, 1, 0)$ follow directly.

Signed three-rate family. For $h = e^t + e^{2t} - ce^{3t}$ with $0 < c \leq 1$, the factors are $1 + u - cu^2 \geq 1$ and $1 + 2u - 3cu^2 \geq 0$ for $u \in [0, 1]$. The first inequality uses $u - cu^2 \geq 0$. The second polynomial is concave, and its endpoint values are 1 and $3 - 3c \geq 0$. Thus $h > 0$ and $h' \geq 0$. Expansion gives Equations (27)–(28). The bracket in Equation (27) increases with $v \in (0, 1)$ and has supremum $-1 + 5c$. Consequently, $c \leq 1/5$ prevents a positive anchored three-point witness at any spacing. The determinant in Equation (28) is the product of coefficients and the squared Vandermonde on v, v^2, v^3; it is strictly negative for every $c > 0$ and $x > 0$.

A faster component with a negative polynomial factor. For $h = e^t + te^{2t}$, minimizing te^t on the half-line gives $-1/e$ at $t = -1$. Minimizing $e^t(1 + 2t)$ gives $-2e^{-3/2}$ at $t = -3/2$. Both yield strict positivity of h and h'. The rate-one block contributes one positive direction, and the degree-one rate-two block contributes one positive and one negative direction. This yields $(2, 1, 0)$ although the faster polynomial component is negative away from zero.

K OSCILLATORY SPECTRA AND GRID ALIASING

The real-rate restriction cannot be recovered from a finite-rank test. A trigonometric polynomial is a finite sum of $e^{i\omega t}$ terms, so its translation span is finite-dimensional despite its imaginary-axis spectrum. For a positive and own-score-monotone example with an exponential envelope, take

$$h(t) = e^t(2 + \cos t) = 2e^t + \tfrac{1}{2}e^{(1+i)t} + \tfrac{1}{2}e^{(1-i)t}. \tag{43}$$

Its derivative is $e^t(2 + \cos t - \sin t) \geq e^t(2 - \sqrt{2}) > 0$, and $h(t) \geq e^t > 0$. For every positive integer k and grid spacing $x = 2\pi k$,

$$h(-jx) = 3e^{-jx}, \qquad j = 0, 1, \ldots. \tag{44}$$

Every Hankel matrix on that grid is therefore rank-one PSD, and the normalized grid weights equal a single softmax distribution. This is an exact aliasing identity. The real-rate Vandermonde and zero-count arguments used in the main theorems are precisely what exclude this collapse within their stated class.

L IMPLEMENTATION CHECKS AND FIGURE DERIVATION

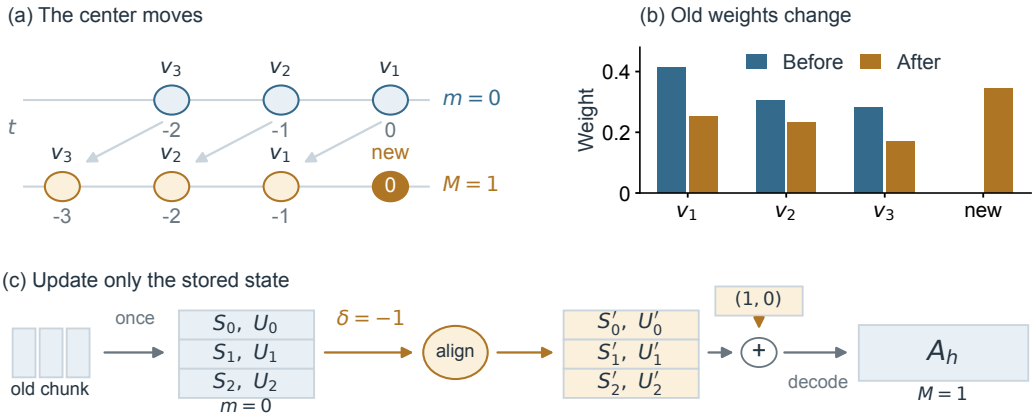


Figure 15: A zero-valued token changes the centering maximum and the weights on previous values. Three paired moments retain the required information for the quadratic generator.

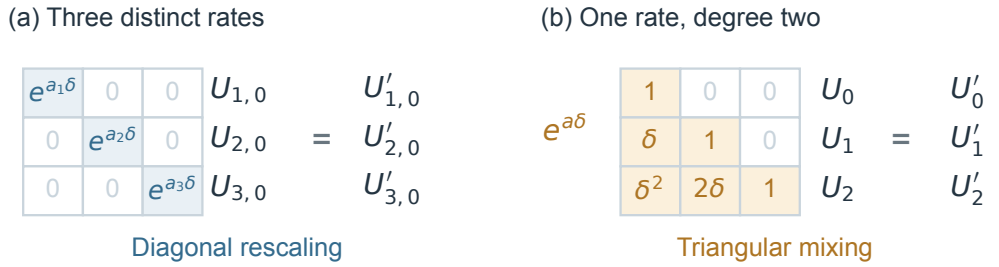


Figure 16: Diagonal and triangular alignment at three value states. Three independent exponentials rescale diagonally; the quadratic generator couples its zeroth, first and second moments.

The exact-arithmetic suite covers the binomial coefficient identity for the Gram matrix, its determinant and negative principal minor, invariance to lower coefficients, confluent continuation factorization and determinant, the equal-spacing and generalized-offset kernel factorizations, and Pascal composition. Rational node variables stand in for exponentials in these identity tests. These algebraic identities apply to arbitrary rational node variables. The five additional constructor cases comprise degrees up to three in a two-rate Hankel test, two continuation tests of lengths six, a degree-four coefficient determinant, and a length-five kernel at offsets $(0, 2, 3, 7, 11)$. All identities hold.

For floating-point comparison with dense barycenters, the five fixed generators are

$$h_1(t) = e^t(1-t), \qquad h_2(t) = e^t(1+t^2),$$
$$h_3(t) = e^t(1-t) + 2/5, \qquad h_4(t) = e^t(1-t) + e^{2t}(1+t^2),$$
$$h_5(t) = 1 + e^t(1-t) + \tfrac{1}{5}e^{2t}.$$

All are positive and own-score monotone by the component calculations above. The deterministic lists are (0), $(0,0)$, $(2,-1)$, $(0,0,-3)$, $(5,1,-2,-4)$, $(0,-1,-2)$, $(3,3,-1,-1,-5)$ and $(4,1,0,-1,-2,-3,-6)$. For a list of length n, deterministic values are the entries of $0,1,\ldots,3n-1$ reshaped into an $n \times 3$ array. The seven-token list additionally supplies a three-block association comparison. Four length-eight Gaussian score/value lists per family use the fixed NumPy generator initialized with seed one.

Each family contributes eight direct moment/dense comparisons, two further comparisons from the parenthesized merge case, and four pseudorandom moment/dense comparisons, totaling fourteen. The reported maximum 2.665×10^{-15} is taken over seventy such scalar maximum-error comparisons across the five families. Comparisons on the same list share input data; the association checks reuse the deterministic seven-token case.

Figure 15 evaluates Equation (1) for $h = e^t(1+t^2)$ on the old scores $(0,-1,-2)$ and the new list $(0,-1,-2,1)$. Its token positions show centered arguments. Bar heights are the exact-formula normalized weights; the new token's value is zero. Figure 16 displays the diagonal and degree-two Pascal alignment matrices, including their coefficients. Figure 14A evaluates x^2-2; B evaluates $-(1-c)+2ce^{-x}+2ce^{-2x}$ for $c \in \{1/10, 1/5, 1/2\}$; C counts the exact positive/negative directions in Equation (12) for positive leading coefficients and degrees zero through six. Removing the positive factors named on the axes preserves the witness sign.