LEARNING FORGETTING RATES ON A HALF-LIFE GRID

Anonymous authors
Paper under double-blind review

ABSTRACT

Recurrent language models such as GatedDeltaNet compress context into a fixed-size state and use a learned forget gate to control its decay. In the GatedDeltaNet models we train, the learned gates favor short retention times: the median head half-life is only 9–16 tokens. We propose Mixture Half-Life Decay Gates (MHLG), in which a token-conditioned router sets each head's decay rate to a weighted average of the rates corresponding to a small, ordered grid of learned half-lives. Each head applies a single scalar decay to a single state, so the state size and update rule are unchanged, while the resulting half-life stays between the shortest and longest learned half-lives and the router can retain weight on slow timescales. In a study of capacity-matched 134M-parameter models trained on 262M tokens, MHLG lowers WikiText-103 test perplexity by 11.5 points, with all five paired seeds favoring it, and by 11.8 points on the full test set. The WikiText-103 improvement is consistent across all evaluated seeds at 400M parameters, at about 1B training tokens, and on a second training corpus. Over three seeds, MHLG reduces perplexity by 9–12 points relative to a wider continuous gate, a trained continuous router, and a control initialized on the same half-lives, and, per layer, its router places 33–85% of its weight on the slowest half-life. A variant that fixes half of the router weight on the slowest half-life raises the gain to 14.5 points. These controls point to a bounded forgetting rate with a maintained slow end as the source of the improvement, and to a learned half-life grid as an inspectable parameterization of this constraint.

1 INTRODUCTION

Efficient recurrent sequence models compress the observed sequence into a fixed-size state, so at every token they must decide how much of that state to forget before knowing which future queries will need it (Sun et al., 2023; Peng et al., 2023; Qin et al., 2024). In gated linear attention (GLA), GatedDeltaNet (GDN), and Mamba (Yang et al., 2024b;a; Gu & Dao, 2023; Dao & Gu, 2024), a free, input-dependent *forget gate* makes this decision: it computes a positive decay rate from the current token and attenuates the state by that rate. We interpret each rate through its *half-life*, the number of tokens after which an unrefreshed memory decays to half its initial magnitude. The GDN gates we train use short half-lives: in five control checkpoints, the median head half-life is 9–16 tokens (Figure 1a). We hypothesize that fast forgetting provides an immediate benefit for next-token loss, whereas the benefits of slow retention are sparse and delayed, so an unconstrained gate underuses the long timescales that help on out-of-domain text.

Existing architectures supply slow timescales in other ways. RetNet fixes one decay per head, HGRN2 learns lower bounds on forgetting, structured state-space models and linear recurrent units (LRU) set their timescales through initialization, and log-linear attention adds states with fixed timescales as the context grows (Sun et al., 2023; Qin et al., 2024; Gu et al., 2021; Orvieto et al., 2023; Guo et al., 2025). Each design changes the timescales together with the rate's input dependence, its initialization, or the state size. It therefore remains open whether constraining a learned, input-dependent rate by itself, with the state size and write rule held fixed, improves language modeling, and which property of the constraint is responsible; addressing this question requires controls for gate capacity and initialization.

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.

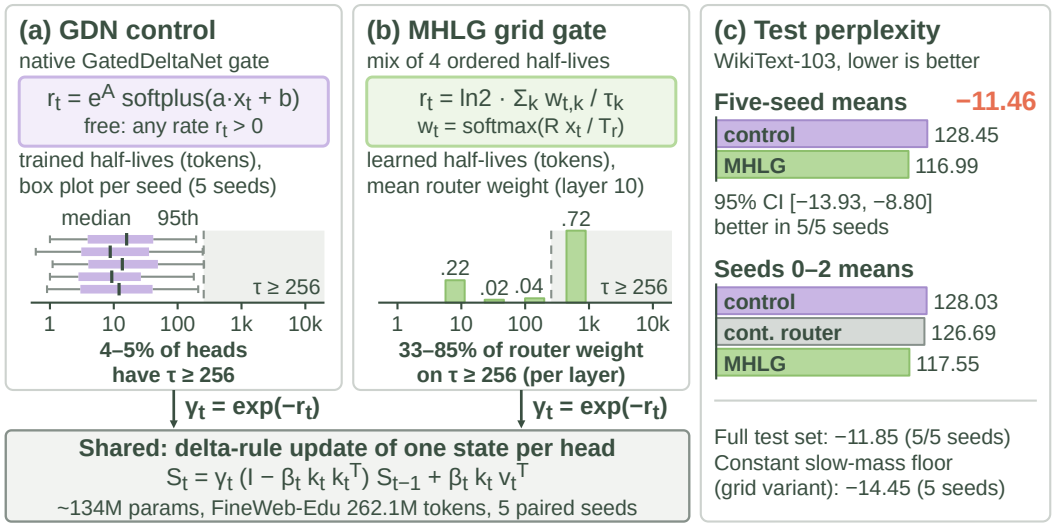


Figure 1: MHLG changes only how GatedDeltaNet forgets. (a) The GDN control computes a free decay rate $r_t > 0$ per head from each token. Box plots: one summary half-life per head (not per token) in each of five trained control checkpoints (box: 25th–75th percentile; whiskers: 5th–95th; tick: median). The median head half-life is 9–16 tokens and the 95th percentile 179–259 tokens; only 4–5% of heads have half-lives of at least 256 tokens (shaded); token-level rates have a heavier slow tail (Appendix I). (b) MHLG mixes $K = 4$ learned, ordered half-lives with token-routed convex weights (T_r: router temperature). Green bars: learned knots and their mean router weight in layer 10 (three seeds); per layer, the slowest knot receives 33–85% of the weight (Table 4). Both arms feed $\gamma_t = \exp(-r_t)$ to the same delta-rule update of one state per head, at matched capacity (134.06M vs. 133.83M parameters) and data (262.1M tokens). (c) Mean WikiText-103 head-window test perplexity over five seeds, where MHLG gains 11.46 points (paired 95% CI $[-13.93, -8.80]$), and over seeds 0–2 with the continuous router, the strongest continuous-rate control of Table 2a. Bottom of (c): the full-test-set difference and the constant slow-mass floor variant (Section 4.3).

We study such a constraint, *Mixture Half-Life Decay Gates* (MHLG; Figure 1b). Each head uses a small, ordered grid of $K = 4$ learned half-lives, which we call *knots*, and a token-conditioned router assigns convex weights to them; the weighted average of the knot rates is the head's decay rate. The rate still varies continuously with the input, but it remains within the range spanned by the knots, and the router can allocate part of its weight on a slow knot. The key design choice is that MHLG changes only the parameterization of forgetting: the projections, the delta-rule write (GDN's key-dependent erase-and-write), and the single state per head remain those of GDN, so the state does not grow with K.

We insert MHLG into GDN and compare it with a capacity-matched GDN control (134M parameters, 262.1M FineWeb-Edu tokens, five paired seeds). Our primary WikiText-103 evaluator, the *head window*, scores the first 23.9% of the test positions. Under this protocol, MHLG lowers test perplexity by 11.46 points (paired 95% seed-bootstrap confidence interval (CI) $[-13.93, -8.80]$, all five seeds); on the *full test set*, the gain is 11.85 points (Figure 1c). Four alternative gates remain within 2 perplexity points of the control (Table 2a): a wider continuous gate, a hard slow clamp, a trained *continuous router* that mixes continuous rates instead of knot rates, and a control initialized on the same half-lives. The clamp acts on one of four channels that are averaged before the update. On held-out WikiText-103, MHLG places about 1.7 times as many tokens as the continuous router at half-lives of at least 256 tokens (Table 3). Its WikiText-103 improvement is consistent across all evaluated seeds at 400M parameters, at about 1B tokens, on a second training corpus, and at a tuned learning rate (Table 6).

The controls also indicate which constraint underlies the improvement. The *constant slow-mass floor*, a variant of MHLG that fixes half of the router weight on the slowest knot, raises the WikiText gain to 14.45 points with a shorter training step time than MHLG ($1.086\times$ vs. $1.125\times$ the control). A hybrid that clamps half of the heads at a 1,024-token half-life and routes continuous rates on the

rest ("Slow clamp + routed fast" in Table 2b) gains 9.33 points. What these gates share with MHLG, and what the native gate and the continuous-rate controls do not guarantee, is a *bounded forgetting rate with a maintained slow end*. MHLG provides it with an inspectable per-layer allocation of weight across half-lives.

Contributions. This paper makes four contributions:

- **A constrained forget-gate parameterization.** MHLG replaces GDN's free decay with a token-routed convex mixture over learned, ordered half-lives. It keeps one state per head, and we characterize the knot allocations that realize the same rate, showing that knot weights are coordinates of one scalar rate rather than separate memories (Sections 2–3).
- **A controlled five-seed comparison.** At matched capacity, data, and write rule, MHLG lowers WikiText-103 perplexity by 11.46 points (11.85 on the full test set), and comparisons with alternative gates point to a bounded forgetting rate with a maintained slow end as the property that produces the gain (Sections 4.2–4.3).
- **An analysis of the learned forgetting mechanism.** The router places 33–85% of its weight on the slowest knot, whose learned half-life is 303–790 tokens. Removing the slow knots costs 39.7–45.1 perplexity points, while a rate-preserving merge of two knots avoids 94.5–96.4% of the perplexity increase caused by deleting them (Section 4.4).
- **Robustness and compute-matched evidence.** The WikiText-103 improvement is consistent across all evaluated seeds in scale, token-budget, corpus, and learning-rate checks, and at seed 0 it holds against a control trained for 15% more steps, roughly MHLG's training time. Full-test-set WikiText-103 and full-shard PG19 results are also reported (Sections 4.2 and 4.5).

2 METHOD: A LEARNED HALF-LIFE GRID

A scalar rate from an ordered grid. For each layer and head h, let $0 < \tau_{h,1} < \cdots < \tau_{h,K}$ be learned half-lives. Positive log-increments enforce the order:

$$\log \tau_{h,1} = b_h, \qquad \log \tau_{h,c} = b_h + \sum_{j<c} \mathrm{softplus}(\delta_{h,j}). \tag{1}$$

A linear router produces $w_{t,h,g} = \mathrm{softmax}(R_{h,g} x_t / T_r)$ for channel group g. The main gate is

$$r_{t,h,g} = \ln 2 \sum_{c=1}^{K} \frac{w_{t,h,g,c}}{\tau_{h,c}}, \qquad \gamma_{t,h,g} = e^{-r_{t,h,g}}, \qquad \tau_{\mathrm{eff},t,h,g} = \frac{\ln 2}{r_{t,h,g}}. \tag{2}$$

Thus the decay is a geometric mean of knot decays, and the rate is a convex combination of knot rates. Its output is continuous. The default $K = 4$ grid starts at $\{16, 64, 256, 1024\}$ tokens; these are initialization values, and the knots are learned with the rest of the model (Table 4 reports the trained half-lives). Grid parameters are shared across groups within a head, not across all heads or layers.

Integration with the recurrent update. The main implementation, MHLG, replaces GatedDeltaNet's native smooth decay while preserving its projections, convolutions, value gate, and delta-rule update. In key-by-value orientation, the update is

$$S_t = \gamma_t (I - \beta_t k_t k_t^\top) S_{t-1} + \beta_t k_t v_t^\top, \qquad o_t = q_t^\top S_t, \tag{3}$$

with normalized keys. The gate uses one scalar per head ($G = 1$), and recurrent-state size does not grow with K. An auxiliary implementation on GLA, the GLA-side grid (MHL), uses four key-channel groups ($G = 4$). A grouped delta variant instead groups value channels. These are explicit experimental factors, and the comparison within GatedDeltaNet isolates the rate parameterization.

Training and controls. The main models have 12 layers of width 768 and eight heads (key/value head dimensions 80/120). MHLG has 134.06M parameters versus 133.83M for the primary GDN control (+0.17%). Both train on FineWeb-Edu sample/10BT (GPT-2 tokenizer) for 16,000 steps, eight 2,048-token sequences per step (262.1M tokens). The main paired comparisons use seeds 0–4 with matched data order and training configuration. Section 4.5 reports step time and a compute-matched control. Most mechanism controls use seeds 0–2; a two-seed 400M comparison, later

LAMBADA seeds, additional token budgets, and tuned recipes are reported separately from the primary estimates.

3 WHAT THE GATE REPRESENTS

Fix a checkpoint, head, and channel group, with knot rates $a_c = \ln 2/\tau_c$. Equation (2) confines r_t to $[a_{\min}, a_{\max}]$. Every interior rate can be expressed by a mixture of the two endpoint rates; the additional knots and router affect the parameterization and optimization within that scalar range.

Window attenuation is an endpoint identity. Over T tokens, the product of the scalar forget factors is

$$\prod_{t=1}^{T} \gamma_t = \exp\left(-\sum_t \sum_c w_{t,c} a_c\right) = \exp(-T\bar{r}_T), \qquad \bar{r}_T = \sum_c \bar{w}_c a_c. \tag{4}$$

This is exact for arbitrary routing paths. It says that a fixed window's endpoint attenuation equals one effective exponential, whose rate $\bar{r}_T$ varies with the endpoint. Because γ_t is a scalar, this attenuation multiplies the key-dependent erasure factors of Equation (3) in the full delta-rule transition. The extended window analysis is in Appendix B.

A knot mixture retains a single state. A hypothetical bank that writes separately to K states would give a fixed write the survival factor $\sum_c w_c \gamma_c^T$. The implemented gate instead gives $(\prod_c \gamma_c^{w_c})^T$ under constant allocation. These are different operators, and the models here train the latter. Heterogeneous rates across heads or windows can generate a mixture when their attenuation curves are averaged, while each head keeps a single evolving state.

Rate-equivalent allocations. Any feasible perturbation δw with $\mathbf{1}^\top \delta w = 0$ and $a^\top \delta w = 0$ leaves the gate unchanged. For distinct rates and an interior allocation, this family has dimension $K - 2$. The weights therefore parameterize the applied scalar rate. Empirically, roughly one third to one half of measured allocation variation is rate-invisible. The rate-preserving knot-merge experiment in Section 4.4 tests the practical importance of this redundancy.

Alternative arithmetic gate. We also evaluate $\gamma_t^{\text{arith}} = \sum_c w_{t,c} e^{-a_c}$. Jensen's inequality gives $\log \gamma_t^{\text{arith}} \geq -\sum_c w_{t,c} a_c$, with equality at a point mass. This comparison holds for identical allocations. Proofs, the bank comparison, and the survival-curve estimator's approximation results are in the appendix.

4 EXPERIMENTS

4.1 EVALUATION PROTOCOLS AND STATISTICS

We evaluate held-out WikiText-103 (WT-103), LAMBADA, and PG19, separately from training-corpus validation. A post-hoc 13-gram overlap scan is reported in Appendix Q.4. The primary WikiText evaluator (the *head window*) scores the first 65,536 of 273,882 test positions (23.9%); the *full-test-set* evaluation scores all of them. LAMBADA perplexity is a teacher-forced stream measure, and we report the benchmark's last-word span accuracy separately. The primary PG19 evaluator (the *capped* protocol) limits the input to 5M tokens and the number of scored windows to four. The *full-shard* protocol uses sliding windows at half-window stride, scores 10,559,488 tokens at context 16k, and removes both caps. We report each protocol separately.

Signed differences are method minus control, so negative perplexity differences favor the method. The main uncertainty intervals are percentile bootstraps of the five paired seed differences (10,000 resamples, fixed random seed). They quantify variation over the paired seeds, and exact sign-flip tests are reported in Appendix F. Sample standard deviations use $n - 1$ normalization. The plots show every observed seed, and analyses with other sample sizes state them.

Table 1: Mean test perplexity over seeds 0–4 (lower is better); the attention row is a single-seed (seed 0) reference model. WikiText-103 uses the head-window evaluator (on the full test set, the difference to the control is -11.85), LAMBADA scores the token stream, and PG19 is reported under the capped and full-shard protocols. Δ rows are MHLG minus the named comparator (negative favors MHLG). Bold: best recurrent model per column (attention is a reference and is not ranked). Dashes: not evaluated under that protocol.

	WT-103	LAMBADA	PG19		
Model	head window	stream	4k, capped	16k, capped	16k, full shard
GDN control	128.45	293.83	**48.27**	**43.45**	**64.23**
GLA-side grid (MHL)	124.13	293.29	53.33	49.10	—
MHLG	**116.99**	**274.12**	48.42	43.67	67.10
Attention (seed 0, ref.)	103.60	226.96	143.78	448.64	—
Δ MHLG − GDN control	$-11.46^{\dagger}$	$-19.70^{\ddagger}$	$+0.16$	$+0.21$	$+2.88^{\S}$
Δ MHLG − MHL	$-7.14^{\dagger}$	$-19.16^{\dagger}$	$-4.90^{\dagger}$	$-5.43^{\dagger}$	—

†The 95% paired seed-bootstrap interval excludes zero in favor of MHLG. ‡Interval $[-38.84, +7.32]$; four of five seeds favor MHLG. §Same sign in all five seeds; descriptive, no interval.

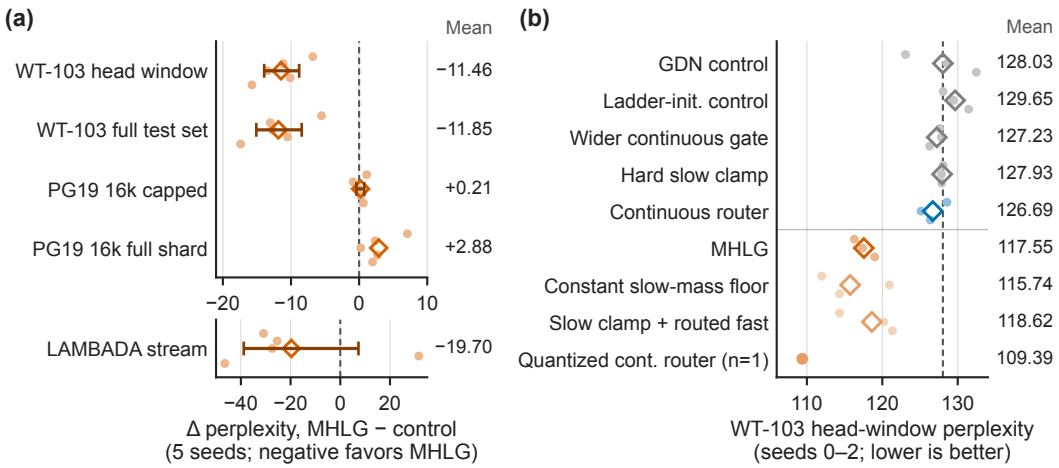


Figure 2: Language-model outcomes of MHLG and of alternative gates. (a) Paired MHLG minus GDN-control perplexity differences over seeds 0–4 (negative favors MHLG): dots are the five per-seed differences, open diamonds their means (printed at right), and capped whiskers 95% paired seed-bootstrap intervals; full-shard PG19 shows point estimates. LAMBADA (bottom) has its own x-scale; dashed lines mark zero. (b) WikiText-103 head-window perplexity: dots are seeds 0–2, open diamonds their means (printed at right, as in Table 2a), and the dashed line is the GDN-control mean. Grey and blue: the controls of Table 2a (Ladder-init.: the native gate initialized on MHLG's initial half-lives). Vermillion: MHLG (dark) and the variants with a maintained slow end (light; Section 4.3). The filled light-vermillion dot shows seed 0 for post-hoc quantization of the continuous router. MHLG and every variant with a maintained slow end lie well below the control line, while the continuous-rate controls stay close to it.

4.2 Primary language-model comparison

Table 1 compares MHLG with the GDN control at a fixed write rule, matched capacity, and an equal token budget. On WikiText-103, the MHLG-minus-control difference is -11.46 perplexity (95% CI $[-13.93, -8.80]$), with improvements in all five seeds (Table 1; Figure 2a shows every seed). On the full test set, the difference is -11.85 (CI $[-15.09, -8.43]$, all five seeds). LAMBADA's stream-perplexity difference is -19.70, with four of five seeds favoring MHLG; with two additional seeds, six of seven pairs favor it (all-pair mean -23.33; Appendix Q.2). Last-word span accuracy is 6.66% for MHLG and 6.12% for the control under the same context protocol.

Table 2: Alternative forget gates on the same GatedDeltaNet backbone, recipe, and evaluators (Section 4.3). (a) Mean WikiText-103 head-window test perplexity over seeds 0–2 (the means in Figure 2b; Ladder-init.: the native gate initialized on MHLG's initial half-lives). (b) Five-seed means of paired differences Δ to the same-seed GDN control (negative favors the variant) in WikiText-103 head-window and PG19 16k (capped and full-shard) perplexity, and training step time relative to the control (median throughput after step 2,000). Bold: best value per column (lowest perplexity, most negative Δ, lowest step time). Per-seed values: Appendix E and Table 15.

(a) Seeds 0–2	WT-103 perplexity
GDN control	128.03
Ladder-init. control	129.65
Wider continuous gate	127.23
Hard slow clamp	127.93
Continuous router	126.69
MHLG	117.55
Constant slow-mass floor	**115.74**
Slow clamp + routed fast	118.62

(b) Paired Δ, seeds 0–4	WT-103	PG19 16k		Step time
	head	capped	full	($\times$ ctrl.)
MHLG	-11.46	$\mathbf{+0.21}$	$\mathbf{+2.88}$	1.125
Constant slow-mass floor	$\mathbf{-14.45}$	$+0.36$	$+3.35$	1.086
Slow clamp + routed fast	-9.33	$+0.57$	$+5.42$	**1.065**

Constant slow-mass floor: half of the router weight is fixed on the slowest knot, and the rest is routed over the faster knots. Slow clamp + routed fast: half of the heads are clamped at a 1,024-token half-life, and continuous rates are routed on the rest.

Table 1 also lists PG19 under the capped and full-shard protocols and a one-seed, same-recipe full-attention reference model; Appendix C discusses these long-context measurements together with training-corpus validation. The grid gate also lowers full-test-set WikiText perplexity on the GLA-side implementation (MHL, four key-channel groups) relative to the GDN control (-3.23, CI $[-5.66, -0.79]$, four of five seeds; Appendix A), and MHLG improves on MHL in turn (-7.14 on the head window, Table 1, last row; -8.62 on the full test set, CI $[-10.89, -6.35]$, all five seeds; Appendix F).

4.3 Which gate property produces the gain?

Table 2 and Figure 2b compare MHLG with alternative gates trained on the same backbone and recipe. Over seeds 0–2, MHLG reaches 117.55 perplexity, against 128.03 for the control and three continuous-rate controls at router-comparable parameter count that stay close to it: a wider continuous gate with four fixed channels (127.23), a hard clamp that holds one channel at a slow rate without routing (127.93), and a trained input-dependent router that mixes continuous rates instead of knot rates (126.69). The clamp acts on one of four channels that are averaged before the state update (Appendix E). Initializing the native gate on the grid's half-life ladder leaves perplexity at the control level (paired difference $+1.62$, CI $[-1.01, +4.97]$; Appendix I) and 12.10 points above MHLG (CI $[+11.77, +12.47]$, all three seeds), so the gain is not explained by the initialization prior or by additional gate parameters alone.

Variants with a bounded forgetting rate and a maintained slow end recover the gain. Post-hoc quantization of the continuous router's rates to the nearest initial knot reaches 109.39 WikiText perplexity at seed 0. The constant slow-mass floor fixes half of the router weight on the slowest knot and routes the remaining half over the faster knots; over five seeds it increases the mean WikiText improvement to 14.45 points (11.46 for MHLG) with a shorter training step time ($1.086\times$ versus $1.125\times$ the control). A hybrid that clamps half of the heads at a 1,024-token half-life and routes continuous rates on the rest gains 9.33 points; Appendix E compares the three variants seed by seed on WikiText, LAMBADA, and capped PG19. What MHLG, the floor, and the hybrid share is a bounded forgetting rate with a maintained slow end, which the native gate and the continuous-rate controls do not guarantee (Section 4.4); among the three, MHLG has the smallest full-shard PG19 change (Table 2b).

Most of the gain is already present with a fixed grid: freezing the grid at its initialization while continuing to train the router costs MHLG 1.47 perplexity points at seed 0 (about a fifth of its seed-0 gain) and the GLA-side grid 0.71. Knot-count, arithmetic-decay, grouped-rate, RetNet, recipe-tuning, and longer-training configurations are reported in Appendices E, G, Q.3, and Q.12.

Table 3: Realized per-window effective half-lives τ of MHLG and of the continuous-router control (three seeds each; 8×2048 held-out WikiText-103 windows, gates computed on each layer's actual input). Each entry is the mean over the twelve layers of a per-layer statistic: the median τ, the standard deviation of $\ln\tau$ pooled over all tokens and heads, the token shares with $\tau \geq 256$ and $\tau \leq 16$, and the standard deviation across windows of the window-mean $\ln\tau$.

Arm	Seed	Median τ (tokens)	SD of $\ln\tau$ (pooled)	Share $\tau \geq 256$	Share $\tau \leq 16$	SD of window-mean $\ln\tau$
MHLG	0	203.2	1.607	0.480	0.250	0.093
	1	185.3	1.653	0.455	0.314	0.083
	2	211.7	1.629	0.455	0.296	0.088
Continuous router	0	103.8	3.006	0.283	0.552	0.098
	1	114.1	2.883	0.263	0.563	0.104
	2	103.4	2.761	0.279	0.563	0.086

Table 4: Where the MHLG router puts its weight, per layer (seed mean over three MHLG checkpoints, full WikiText-103 test set; weights rounded to two decimals). $K_{\text{eff}} = \exp(-\sum_k \bar{w}_k \ln \bar{w}_k)$ is the exponentiated entropy of the layer's mean router weights $\bar{w}_k$. The full learned grids and per-knot weights are in Table 13 (Appendix L).

Layer	0	1	2	3	4	5	6	7	8	9	10	11
Slowest-knot half-life (tokens)	456	303	307	315	361	379	790	560	494	536	628	507
Weight on slowest knot	0.85	0.36	0.33	0.45	0.72	0.61	0.76	0.65	0.58	0.62	0.72	0.68
Weight on fastest knot	0.01	0.45	0.49	0.44	0.23	0.28	0.16	0.25	0.31	0.30	0.22	0.22
Effective knots K_{eff}	1.59	3.20	3.01	2.80	2.16	2.62	2.11	2.43	2.65	2.46	2.15	2.40

4.4 Behavior of the trained gate

How does MHLG forget differently from a gate with continuous rates? Table 3 compares the per-window effective half-lives realized by MHLG and by the continuous-router control on held-out WikiText-103 windows. MHLG's median effective half-life is 185–212 tokens against 103–114 for the continuous router; it places 46–48% of tokens at half-lives of at least 256 tokens (26–28% for the router), 25–31% at 16 tokens or less (55–56%), and the standard deviation of $\ln\tau$ across tokens and heads is about half as large (1.61–1.65 vs. 2.76–3.01). The bounded grid thus shifts the realized rates toward slow timescales and narrows their spread. This is consistent with Figure 1a, where GDN's native gate, free to shorten its retention times, keeps the median head half-life at 9–16 tokens.

Table 4 shows how the router allocates its weight. In every layer the slowest knot, with learned half-lives of 303–790 tokens, receives 33–85% of the weight; in layers 1–11 the rest concentrates on the fastest knot, and the effective number of knots is 1.6–3.2. The learned grids also have a simple placement rule. Because each window's endpoint attenuation is a single exponential (Equation (4)), placing the knots amounts to quantizing the interval of rates that the router reaches, and under relative (log-domain) error the minimax K-point quantizer of an interval is geometric. Training keeps the grids close to this placement (median adjacent half-life ratio about 3.7, range 3.29–5.27, across the 36 seed–layer grids, versus 4 at initialization) while moving the knots to bands of 7–10, 26–40, 86–191, and 286–1,004 tokens (Appendix Q.12). By contrast, on synthetic multi-query associative recall (MQAR), the four solving delta-backbone runs (gaps 32 and 128) place 77–100% of late-layer weight on the slowest knot, with effective knot counts of 1.0–1.5 and slow half-lives of up to 6.4×10^5 tokens (Appendix E), a single-timescale regime, whereas the language model spreads its weight more broadly.

The slow knots matter, and they act through the realized rate. Table 5 applies four evaluation-time interventions to the trained gate. Removing the knots with half-lives of at least 256 tokens degrades the model by $+39.7$ to $+45.1$ perplexity across three seeds, with the degradation concentrated in deeper layers. Merging each layer's least-cost adjacent pair of knots into a single knot whose rate preserves the realized effective rate costs only 0.35–1.39 points and avoids 94.5–96.4% of the perplexity increase caused by deleting the same knots (Appendix M). The realized rate is therefore what the model relies on. Routing every token to its layer's dominant knot, which is the slowest

Table 5: Evaluation-time knot interventions on the three MHLG checkpoints, applied in all twelve layers. Columns 3–6 give the change in WikiText-103 head-window perplexity relative to the unmodified model (positive is worse). *Drop slow*: zero the weight on knots with half-lives of at least 256 tokens and renormalize. *Dominant only*: route every token to its layer's dominant knot. *Delete* and *merge*: remove each layer's least-cost adjacent pair of knots, or replace the pair by one knot whose rate preserves the realized effective rate. Last column: share of the deletion damage that merging avoids (Appendix M).

	Unmodified	Δ perplexity after the intervention				Damage
Seed	perplexity	Drop slow	Dominant only	Delete pair	Merge pair	avoided (%)
0	116.28	+45.1	+6.2	+25.4	+1.39	94.5
1	117.38	+39.7	+10.7	+7.0	+0.35	95.0
2	118.99	+43.0	+10.8	+19.0	+0.69	96.4

Table 6: Robustness of the MHLG advantage: MHLG minus the capacity-matched GDN control at matched recipe, in test perplexity (negative favors MHLG). The main setting (134M parameters, 262.1M training tokens) is the five-seed mean; each other row changes one factor and lists every seed (s0–s2). PG19 uses the capped protocol at 16k context; the ≈1B-token pilot trains for 61k steps; the learning-rate row trains both arms at 10^{-3}, the best of four rates for both arms at seed 0. Dashes: not measured, or not reported for every seed; blank: seed not run. Details in Appendices R, Q.3, H, and G.

	Δ WT-103			Δ LAMBADA			Δ PG19 16k		
Setting	s0	s1	s2	s0	s1	s2	s0	s1	s2
Main (5-seed mean)		−11.46			−19.70			+0.21	
400M parameters	−5.60	−13.79		—	—		−2.12	−1.87	
≈1B tokens	−2.28	−7.15	−11.71	−7.95	−28.28	−25.56	+1.49	−0.25	−0.18
OpenWebText corpus	−24.15	−9.99		−36.53	−44.59		−16.26	+20.63	
Learning rate 10^{-3}	−5.97	−10.08	−3.90	—	—	—	—	—	—

knot in 10 to 11 of the 12 layers, costs +6.2 to +10.8 points, about a quarter or less of the perplexity increase from removing the slow knots, so the dominant knot carries most of the slow knots' contribution. Appendix A adds further rate and recall measurements: a survival-curve analysis that detects the spread of the realized rates across tokens, MQAR recall at gap 32, where moving the gate to the delta backbone closes 98.6–103.2% of the GLA-side model's recall deficit (three seeds), a token-bucket analysis of the GLA-side grid (MHL) that places its gains over the GDN control on high-surprisal tokens, and heterogeneous decay rates in a pretrained Mamba2-130M checkpoint.

4.5 Robustness and compute matching

Table 6 collects the robustness checks. The WikiText improvement is consistent across all evaluated seeds at 400M parameters, at about 1B training tokens, on OpenWebText, and at a tuned learning rate. Across these checks, LAMBADA improves in every measured seed, and capped PG19 improves in both 400M seeds.

Compute-matched control. The router adds 0.17% parameters, and training step time is 1.125× that of the control (median throughput after step 2,000, five seeds). A control trained for 15% more steps, approximately matching MHLG's training time, stays at 123.36 seed-0 WikiText perplexity (123.08 at the standard budget; Appendix E), so the WikiText improvement persists at matched training time for seed 0.

5 Related work

Forgetting within recurrent attention. Linear attention and fast-weight interpretations replace token caches with recurrent sufficient statistics (Katharopoulos et al., 2020; Choromanski et al.,

2020; Schlag et al., 2021). GLA learns input-dependent decay; DeltaNet and GatedDeltaNet add key-dependent state correction (Yang et al., 2024b;c;a). DeltaProduct and test-time-training layers enrich the update itself (Siems et al., 2025; Sun et al., 2024). Our intervention preserves the chosen write rule and parameterizes only its real forgetting rate. Concurrent spectral analyses of GDN also find that long-context behavior depends on slow decay rates, but they reparameterize the free gate when initializing continued pretraining (Liu et al., 2026), whereas our gate is trained from scratch as a token-routed convex mixture over a monotone half-life grid. Selective RoPE adds a distinct rotational degree of freedom (Movahedi et al., 2026), complementary to the half-life grid, which acts on the real forgetting rate at fixed state size.

Timescales and constrained gates. HiPPO and structured state spaces encode memory through structured dynamics; LRU and clockwork recurrences use complementary timescale designs (Gu et al., 2020; 2021; Orvieto et al., 2023; Koutník et al., 2014). RetNet, RWKV, and HGRN2 provide fixed or learned decay structures, including learned lower bounds on forgetting (Sun et al., 2023; Peng et al., 2023; Qin et al., 2024); the constant slow-mass floor of Section 4.3 is close in spirit to such bounds. Mamba and its state-space duality, Griffin, Hyena, and xLSTM further couple recurrence, selectivity, and local mixing (Gu & Dao, 2023; Dao & Gu, 2024; De et al., 2024; Poli et al., 2023; Beck et al., 2024). Nearest to us is log-linear attention, which segments the cache into a logarithmically growing bank of fixed-timescale states (Guo et al., 2025). Our finite grid is a continuous scalar-rate parameterization within this design space: recurrent-state size stays constant, and the grid parameterizes the realized rate directly.

Retrieval and evaluation. Recall benchmarks and recall–throughput comparisons motivate the distinction between language modeling and long-range retrieval (Arora et al., 2024a;b); concurrent work reports strongly seed-dependent recall at a fixed state budget (Boesch & Wee, 2026). We test transfer from FineWeb-Edu to WikiText-103 and PG19 under explicitly stated coverage (Penedo et al., 2024; Merity et al., 2016; Rae et al., 2020), because the set of scored targets determines which changes an evaluation can measure. Estimator- and systems-related work is discussed in Appendix A.

6 Discussion and conclusion

The central result is a five-seed WikiText-103 improvement from changing only the rate parameterization within GatedDeltaNet, confirmed on the full test set and consistent in sign across a larger model, a larger token budget, a second training corpus, and a tuned learning rate. The mechanism is a reparameterization of forgetting: each head keeps one recurrent state, the router produces a continuous scalar rate, and the realized rate carries the effect. The controls point to a bounded forgetting rate with a maintained slow end as the relevant constraint: MHLG, its constant slow-mass floor variant, and a slow-clamp hybrid all obtain WikiText gains of 9.3–14.5 points, whereas the continuous-rate controls stay close to the control. In practice, the floor variant reaches the largest WikiText gain with a shorter training step time than MHLG, while MHLG has the smallest full-shard PG19 change of the three and an inspectable allocation. A learned half-life grid is therefore a useful, inspectable parameterization of forgetting with measured transfer gains.

6.1 Limitations and future work

We see four directions for future work; Appendix C discusses each in detail. *Scale and breadth.* Our models are small and trained for limited token budgets on one main corpus; the checks at larger scale, longer budgets, and on a second corpus use few seeds, and the cross-architecture runs are incomplete. Wider sweeps would show how the gain scales. *Transfer without in-domain cost.* The held-out gain comes with higher training-corpus validation perplexity and a long-context penalty under full-coverage evaluation; gates that keep the gain without these costs are a natural next step. *Retrieval.* Neither model retrieves reliably at long gaps at this scale, so whether maintaining a slow rate improves retrieval remains open. *Mechanism.* Our controls cover specific constructions; whether a finite grid is necessary, how many knots are best, and how allocations relate to stored content call for broader parameterizations and causal probes. Rate-equivalent allocations are not separately identifiable, and the channel-level clamp control does not guarantee a slow effective rate after channel averaging.

Ethics Statement

This work trains and evaluates language models using publicly available corpora, benchmarks, and model checkpoints. We did not conduct human-subject experiments. The work shares the risks common to language-model research, including bias and potentially harmful content in public corpora and model outputs.

Reproducibility Statement

The supplementary ZIP archive provides code, configurations, launch and analysis scripts, and recorded results, together with instructions for reproducing the experiments and generating the reported tables and figures. Experimental settings are specified in the main text and appendix.

AI Use Statement

During the project, vibe coding tools (Codex and Claude Code) were used to help develop the code for the methods and experiments. All such work was conducted under human supervision and underwent human review. During manuscript preparation, AI tools were used to polish and revise the text and figures.

References

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *ICLR*, 2024a. arXiv:2312.04927.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, et al. Simple linear attention language models balance the recall-throughput tradeoff. In *ICML*, 2024b. arXiv:2402.18668.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, et al. xLSTM: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024. URL https://arxiv.org/abs/2405.04517.

Julian Boesch and Andrew Wee. Anatomy of associative recall in fixed-state recurrences: A matched-state decomposition, an interference wall, and a curriculum that breaks it. *arXiv:2609.16183*, 2026.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. URL https://arxiv.org/abs/2009.14794.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. URL https://arxiv.org/abs/2307.08691.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. arXiv:2405.21060.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*, 2022. URL https://arxiv.org/abs/2205.14135.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024. URL https://arxiv.org/abs/2402.19427.

William Feller. *An Introduction to Probability Theory and Its Applications, Vol. II*. Wiley, 2 edition, 1971.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. arXiv:2312.00752; COLM 2024.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *NeurIPS*, 2020. arXiv:2008.07669.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. URL https://arxiv.org/abs/2111.00396.

Han Guo, Songlin Yang, Tarushii Goel, Eric P. Xing, Tri Dao, and Yoon Kim. Log-linear attention. *arXiv:2506.04761*, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. *arXiv preprint arXiv:2006.16236*, 2020. URL https://arxiv.org/abs/2006.16236.

Jan Koutník, Klaus Greff, Faustino Gomez, and Jürgen Schmidhuber. A clockwork RNN. In *ICML*, 2014.

Zian Liu, Yiwen Hu, Zican Dong, Tian Xie, Wayne Xin Zhao, Yucheng Ding, Ran Tao, and Bryan Dai. SpectralShift: Effective context window extension of Gated DeltaNet via spectral reparameterization. *arXiv:2609.14320*, 2026.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher. Selective rotary position embedding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/113e6f1d94af5df4f306fbcb3f82339f-Abstract-Conference.html.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023. URL https://arxiv.org/abs/2303.06349.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *NeurIPS D&B / arXiv:2406.17557*, 2024.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of EMNLP*, 2023. arXiv:2305.13048.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *ICML*, 2023. arXiv:2302.10866.

Daniel Potts and Manfred Tasche. Parameter estimation for exponential sums by approximate Prony method. *Signal Processing*, 90(5):1631–1642, 2010.

Gaspard Riche de Prony. Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures. *Journal de l'École Polytechnique*, 1(2):24–76, 1795.

Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. HGRN2: Gated linear RNNs with state expansion. In *COLM*, 2024. arXiv:2404.07904.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. *ICLR*, 2020.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. *arXiv preprint arXiv:2102.11174*, 2021. URL https://arxiv.org/abs/2102.11174.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. *arXiv preprint arXiv:2502.10297*, 2025. URL https://arxiv.org/abs/2502.10297.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, et al. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024. URL https://arxiv.org/abs/2407.04620.

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, 2023. arXiv:2307.08621.

Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, 1939.

David Vernon Widder. *The Laplace Transform*. Princeton University Press, 1941.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule, 2024a. arXiv:2412.06464; ICLR 2025.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *ICML*, 2024b. arXiv:2312.06635.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the Delta Rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024c. URL https://arxiv.org/abs/2406.06484.

Morris Yau, Sharut Gupta, Valerie Engelmayer, Kazuki Irie, Stefanie Jegelka, and Jacob Andreas. Sequential-parallel duality in prefix scannable models. *arXiv preprint arXiv:2506.10918*, 2025. URL https://arxiv.org/abs/2506.10918.

A ADDITIONAL IMPLEMENTATION, RATE, AND RECALL DETAILS

This section gives the full versions of material that Sections 4 and 5 summarize in a sentence or two. The paragraphs on MQAR accounting, the survival-curve estimator, and token buckets and Mamba2 hold the rate and recall measurements summarized in the last paragraph of Section 4.4; the gap-128 recall and needle-in-a-haystack results are in Appendices J and K.

Code names. The appendix identifies models by code name. MHLDN is MHLG, MHL is the GLA-side grid, and GDN_CM (also written GDNCM) is the primary GDN control. GDNCAP, GDNBSN, GDNRM, and GDNLAD are the wider continuous gate, the hard slow clamp, the continuous router, and the ladder-initialized control of Table 2a. RFLONG is the constant slow-mass floor and HYBCS the hybrid ("Slow clamp + routed fast") of Table 2b, and QUANT is the post-hoc quantization of the continuous router. DROPLONG and DOM are the *drop slow* and *dominant only* interventions of Table 5. The remaining variants (MHLDNH, STATICH, G4, A1B, ARITHDN, RETNET, RETNETM) are defined in Appendix E and in the notes to Table 15; Appendix N collects all arms in one grid. The appendix also calls MHLG the *combined arm*, because it combines the router of the GLA-side grid with the delta-rule backbone. GDN_FS and BAND_FS are the two earlier checkpoint families of the survival-curve study (a GatedDeltaNet and a variant with a softened delta-rule erase); Appendix D.1 defines them together with the terms of that study (single-rate and erase-softened class, certified estimator, T=512 re-run).

Conventions used in the appendix. The appendix uses the following conventions. An *arm* is a gate variant trained under the common recipe, and a *cell* is one arm at one seed (for the survival-curve estimator, one (layer, seed) pair); s0–s6 denote seeds and L0–L11 (also written L00–L11) denote layers. *MHLG side* and *GLA side* refer to the two write rules (delta rule and GLA). The *head window* (also *nominal-32*, for its 32 windows) is the primary WikiText-103 evaluator of Section 4. Many single-seed controls are summarized by the recovery statistic $s = (123.080 - o)/6.799$, where o is the arm's WT-103 head-window perplexity, 123.080 is the seed-0 GDN control (the *anchor*; the full-attention reference model is called the *attention anchor*), and 6.799 is MHLG's seed-0 gain over it; $s = 1$ recovers MHLG's full seed-0 gain and $s = 0$ none of it. Each such statistic comes with a decision rule fixed before the run, usually with three outcomes separated by thresholds (for example $s \geq 0.75$, $s \leq 0.25$, and a partial band in between); hyphenated or capitalized outcome names (for example bq-persist2 or PARTIAL) are the labels under which the outcomes are stored in the result files. A *penalty* is a perplexity increase relative to the control; ppl abbreviates perplexity, and nats are per-token NLL (negative log-likelihood) differences in natural-log units. *Zero-compute* analyses reuse stored checkpoints, training logs, or per-token evaluation outputs without new training, a *byte-identical recipe* changes nothing but the stated factor, and a *parameter stamp* is the measured parameter count (tied embedding counted once; Appendix D) checked before a run. The *family* of an arm is its set of seeds under the common recipe; used alone ("family", "family side", "family control"), the word denotes the GDN control, and a *family band* is the range of a measure over the family's seeds. A detector *fires* when it reports a second timescale. File names in typewriter font (for example `evals/h2_lm_bq_agg.json`) refer to result files and scripts in the supplementary material.

GLA-side implementation. The GLA-side implementation has weaker outcomes: the five-seed full-test-set WikiText difference is -3.23 (CI $[-5.66, -0.79]$), while validation worsens by $+3.03$ and capped PG19 by $+5.1$/$+5.6$ at 4k/16k. The delta implementation improves all four primary evaluation measures over that implementation (Table 1, last row), but the write rule and grouping granularity both change. The controlled within-GatedDeltaNet comparison supplies the cleaner evidence for the rate parameterization.

MQAR accounting. On MQAR-style associative recall, both tested families stay near 10^{-4} accuracy at gap 1,024 after 12,000 steps; gaps of at least 256 remain at the floor in the measured rate sweep. At gap 32, the reported terminal-accuracy means are 98.98% for the delta control ($n = 3$) and 52.27% for the GLA grid model ($n = 4$). The reported aggregate excludes a 97.82% run with a repeated seed identifier. Including all five recorded runs gives 61.38%, a descriptive run average rather than five independent seeds. This favorable high-tail run limits a general trainability claim; none of these accuracy ratios measures training speed. Rehosting the gate on the delta backbone

closes this gap (98.6–103.2% of the difference over three seeds; Appendix E), but that comparison also changes grouping. Gap-128 results depend on learning rate and seed for both families (Appendix J).

PG19 sensitivity calculation. Capped PG19 differences ($+0.16$ at 4k and $+0.21$ at 16k) are described as small observed gaps, not equivalence. The 80%-power sensitivity calculation is neither a bound on the true effect nor a guarantee that a larger effect would be detected.

Survival-curve estimator. A survival-curve estimator, which fits one- and two-component exponential decays to the realized attenuation curves, originally motivated the timescale study. Its two-component criterion fires in $34/36$ (layer, seed) cells in each of two trained families (GDN_FS and BAND_FS; Appendix D.1), but shuffling rate order while preserving marginals reproduces detection at mean share 0.96, with no cell above its null. The estimator therefore measures marginal rate dispersion and supplies no evidence of temporally organized memory spectra in those checkpoints. Calibration and failed detector variants remain in Appendix D.1.

Token buckets and Mamba2. Token-bucket measurements, taken on the GLA-side grid (MHL) against the GDN control rather than on MHLG, locate a related tradeoff: gains concentrate on high-surprisal tokens, whereas repeat-heavy tokens often incur losses (Figure 3). This is descriptive localization, not proof of a causal copy mechanism. A pretrained Mamba2-130M checkpoint also exhibits heterogeneous decay rates, but the separate from-scratch Mamba2 runs stopped around 6.1k/6.6k of 16k steps (Appendix R). Their intermediate measurements do not constitute a completed cross-architecture replication.

Additional related work: representations, kernels, and evaluation. Classical exponential-mixture and approximation results support the auxiliary survival-curve estimator (Feller, 1971; Widder, 1941; Szegő, 1939; Potts & Tasche, 2010); they do not identify what content a trained gate stores. Efficient scan and attention algorithms address execution separately from that representation question (Yau et al., 2025; Dao et al., 2022; Dao, 2023). Recall benchmarks and recall–throughput comparisons motivate the distinction between language modeling and long-range retrieval (Arora et al., 2024a;b). A decomposition of masked recall at a fixed 1,024-element state budget attributes failures at long distances to interference under sparse supervision, with strongly seed-dependent outcomes (Boesch & Wee, 2026). This is consistent with the seed dependence of our gap-128 recall results, although the tasks differ and our claim concerns maintained slow rates in language modeling, not a decomposition of recall capacity.

B Extended retention-rate analysis

Setup: the trained object is a scalar rate with a guaranteed grid. Let fixed half-lives $0 < \tau_1 < \cdots < \tau_K$ parameterize rates $\gamma_i = 2^{-1/\tau_i}$, with a token-dependent allocation $w(x) \in \Delta^{K-1}$ produced by a softmax router. Every trainable in this paper is one of the two scalar-gate parameterizations

$$\gamma^{\mathrm{conv}}(x) = \sum_{i=1}^{K} w_i(x)\,\gamma_i, \qquad \log \gamma^{\log-\mathrm{conv}}(x) = \sum_{i=1}^{K} w_i(x)\,\log \gamma_i, \tag{5}$$

each computing *one scalar per token* applied to a single recurrent accumulator per (head, channel-group). No K-fold memories exist in any trained checkpoint (Proposition 1 below); what the grid fixes is a *retention-rate constraint*: the realized log-rate is confined to the knot-rate hull $[a_{\min}, a_{\max}]$, $a_c = \ln 2/\tau_c$, and the allocation is an explicit parameter coordinate (Remark 1). Two readings of the multi-timescale intuition follow from exactly where multi-atom structure can be realized.

What a scalar gate can and cannot represent, window by window.

Lemma 1 (Mean-rate collapse)**.** *If the allocation is constant over a window, the realized survival is one exponential with per-step decay factor $\bar\gamma = \sum_i \bar w_i \gamma_i$ (convex form); under the log-convex form* (5) *the collapse is exact for* arbitrary *allocation paths: a window's gate product is a single exponential at the window-mean allocation.*

Lemma 2 (Aggregation mixture). *Heads/features with heterogeneous fixed rates yield an aggregate curve $S(t) = \sum_h \nu_h \bar{\gamma}_h^t$, a finite discrete completely monotone mixture. This curve-level statement requires $\bar{\gamma}_h$ to be fixed across the plotted lag t. Substituting a different window-mean rate at each endpoint does not by itself establish complete monotonicity. The empirical surplus measurement can therefore be explained by fixed-rate heterogeneity without identifying temporally organized memory content.*

Lemma 3 (Allocation-swing retention channel). *For the convex gate in* (5)*, $g_j := \log \mathbb{E}_w[\gamma] - \mathbb{E}_w[\log \gamma]$ satisfies $0 \le g_j \le \mathrm{Var}_w(\gamma)/(2\gamma_{\min}^2)$ (Taylor remainder against $|(\log)''| \le 1/\gamma_{\min}^2$). The log-convex gate realizes exactly the $\mathbb{E}_w[\log \gamma]$ term and forgoes this bonus; the two forms coincide at point-mass allocations and their per-token gap is exactly g_j, so on identical routers the variant-difference isolates the swing channel (the swing bound certified over* 200000 *random allocations: the worst draw reaches* 0.942 *of the* $\mathrm{Var}(\gamma)/(2\gamma_{\min}^2)$ *ceiling,* `evals/mhlg_theory_cert.json`*).*

Where retained structure actually lives: the window-kernel classification.

Window-kernel classes under switching allocation. The collapse lemma (Lemma 1) states what each form realizes on one window; we now characterize both exactly (full proofs in Appendix Q).

Theorem 1 (Window-kernel classes, complete). *Consider freely chosen allocations in the closed simplex for a fixed (head, channel-group) and a window of T gates with allocation path $w(1..T)$, and write $a_c = \ln 2/\tau_c \in [a_{\min}, a_{\max}]$. (i)* Log-convex (implemented). *The window survival equals $e^{-T\bar{r}}$ with $\bar{r} = \sum_c \bar{w}_c a_c$, $\bar{w} = T^{-1}\sum_t w(t)$; conversely every $r \in [a_{\min}, a_{\max}]$ is realized by a constant two-knot path: the within-window class is* exactly $\{e^{-Tr} : r \in [a_{\min}, a_{\max}]\}$*, a scalar endpoint class. Finite softmax logits realize only interior rates, with the displayed set giving their closure; a shared learned router need not realize arbitrary allocation paths. (ii)* Convex. *The window survival equals $\mathbb{E}[e^{-X_T}]$ with $X_T = \sum_{t=1}^T A_t$, A_t independent, $\Pr(A_t = a_c) = w_c(t)$ — equivalently, $F(z) = \mathbb{E}[e^{-zX_T}]$ is completely monotone in a common rate multiplier $z \ge 0$, with the measured product at $z = 1$. Its auxiliary mixing measure has at most $\min\big(K^T, \binom{T+K-1}{K-1}\big)$ atoms inside the T-fold sumset of $\{a_c\}$. This is not a completely-monotone classification in window length T under arbitrary switching, nor a physical memory-bank realization.*

Proof sketch. (i) Telescoping log of the window product plus a two-knot interpolation. (ii) Factorization under independent coordinates, Bernstein's sufficient direction, a bounded-composition count (full detail with the proofs). □

Three further statements live with the proofs: achievable trajectories of the implemented gate — the single-atom within-window family, independent cross-window products, transient-only readout path-dependence with a $(K-2)$-dimensional null family (Proposition 3); the log-domain knot-cost conversion (K linear in LT at the exponentially small tail, Corollary 2); and the self-averaging statement that stationary routers yield no heavy-tailed long-window survival in either form (Remark 2); a non-worst-case counterpart — exact four-point identifiability of a two-atom truth inside an explicit engagement threshold — is Corollary 4.

Corollary 1 (Channel accounting). *(a) Convex $\ge$ log-convex on every allocation path, with equality iff every allocation is a point mass; the log-excess is $\sum_t g_t$, bounded per Lemma 3. (b) Under the implemented log-convex form every* fixed *window's* endpoint *is single-rate, $S(T) = e^{-\bar{r}_T T}$ — an endpoint identity, not a curve classification: consecutive gates $(0.9, 0.1)$ give $S(0)S(2) - S(1)^2 = -0.72$, outside every completely monotone class. Averaging fixed heterogeneous rates yields the mixture in Lemma 2; varying endpoint rates across window lengths does not itself imply such a curve. Only the arithmetic arm collects the swing channel (Lemma 3).*

The bank contrast: what would be multi-atom memory and why this paper does not train it.

Proposition 1 (Two realizations of one knot family). *The knot family admits two realizations that must not be conflated: (i) a* parallel bank *(write-time routing) of K per-component memories $m_i(t) = \gamma_i m_i(t-1) + w_i(x_t)\, W(x_t)$, a write's impulse response exactly $\sum_i w_i(x_s)\, \gamma_i^{t-s}$ (finite spectrum, token-dependent masses); and (ii) the* scalar grouped rate trained here *— one recurrent*

accumulator gated by the per-token scalar (5)*, window kernels that are* products of per-token mixtures*, one atom per window under the implemented log-convex form (Theorem 1(i)), not a K-atom mixture. At fixed allocation $\bar{w}$, with $M_t := \left(\sum_i \bar{w}_i \gamma_i^t\right)^{1/t}$ the power means, the three kernels in the same order satisfy, for $t \geq 1$,*

$$\sum_i \bar{w}_i \gamma_i^t \;=\; M_t^t \;\geq\; M_1^t \;=\; \left(\sum_i \bar{w}_i \gamma_i\right)^t \;\geq\; \left(\prod_i \gamma_i^{\bar{w}_i}\right)^t, \tag{6}$$

by power-mean monotonicity and AM–GM; the first inequality is equality by definition at $t{=}1$ and strict for $t > 1$ whenever the support spans two knots of distinct rate, and the second is strict whenever the γ_i take two distinct values on the support — in particular every allocation in the interior of the simplex over a grid with distinct knot rates makes both inequalities strict (numeric certification `evals/mhlg_theory_cert.json`*: zero strictness violations over* 50000 *random allocations, $t{\in}[2, 48)$).* Everything trained in this paper is (ii)*: the GLA-side mixture arm and the combined variant run the grouped scalar rate inside the backbone's ordinary recurrent state (Sec. 2); no K-fold memories exist in any trained checkpoint, and* (6) *gives the exact sense in which the bank spectrum upper-bounds the trained gate's.*

Proposition 2 (Fixed-size recurrent state)**.** *Both realizations have fixed state size. The bank carries $K \cdot d_v$ per head and every future readout is a linear functional of $\{m_i(t)\}_{i \leq K}$ (components evolve autonomously). The trained scalar form carries the backbone's single recurrent state per (head, channel-group) — strictly lighter — with all later readouts linear in it.*

Remark 1 (The trained budget coordinate)**.** *The allocation $w(x_t)$ is a parameter coordinate trained end-to-end by the LM loss. Different allocations can induce the same rate, so this coordinate is not an identifiable memory partition.*

Approximation theory behind the survival-curve estimator. The survival-curve presence measurement (Appendix A) classifies realized survival curves using the completely-monotone (CM) kernel machine: a survival function is CM iff it is a Laplace mixture of exponentials (Feller, 1971; Widder, 1941); on finite windows a K-term exponential sum fits a bounded CM curve with the super-exponential Gauss– Christoffel rate $M(eTL/2K)^{2K}$, and the stated two-atom local analysis relates recovery to noise, separation, mass, and engagement. This apparatus is the classifier's hypothesis class — the $\geq 20\%$ held-out *surplus* criterion detects realized multi-timescale structure in a *curve*, regardless of which architectural object produced it. All statements and proofs live in Appendix Q; none of them claims the trained scalar gate *stores* such a spectrum (Proposition 3).

Scope and open items. The representation and ergodic results are classical. The approximation results characterize auxiliary curve fitting, and the local identifiability conditions do not imply global recovery for arbitrary overparameterized fits.

C Extended discussion of limitations

This appendix gives the detailed accounting behind the four future-work directions of Section 6.1. The first four paragraphs follow those directions in order and give the underlying measurements with their numbers; the last two add notes on evaluation coverage and experimental scope.

Scale and breadth. The models are small and briefly trained (134M parameters, 262.1M tokens, one main corpus). The two- and three-seed robustness checks vary strongly in size (2.28–24.15 points), capped PG19 is near parity at 1B tokens except seed 0 and changes sign on OpenWebText, and validation worsens at 400M ($+0.07$/$+1.62$) and on OpenWebText ($+1.59$, seed 0); these checks bound rather than extend the main result and establish no scaling trend. The from-scratch Mamba2 runs stopped at 6.1k/6.6k of 16k steps and do not form a cross-architecture replication. The one-seed attention reference leads at short context (WikiText 103.60) but reaches 448.64 at 16k, beyond its 2,048-token training context, and says nothing about modern long-context attention. The GLA-side implementation has larger validation and capped-PG19 penalties, and cross-backbone contrasts change write rule and grouping together. Seed intervals are descriptive (exact sign-flip $p \geq 0.0625$ at $n = 5$) and do not cover datasets or architectures. Training was not rerun on a decontaminated corpus.

Transfer and in-domain costs. The WikiText gain is a transfer effect with measured costs: training-corpus validation perplexity rises by $+1.89$ (CI $[+0.29, +3.69]$, four of five seeds; on low-surprisal tokens, Appendix P), and full-shard PG19 at 16k by $+2.88$ (64.23 to 67.10) in all five seeds, whereas the capped differences ($+0.16$ at 4k, $+0.21$ at 16k) are small observed gaps, not equivalence. The floor variant's larger WikiText gain comes with a larger full-shard PG19 penalty ($+3.35$; hybrid $+5.42$), so the variants form a tradeoff rather than a unique winning construction, and the evidence favors the bounded, slow-biased rate rather than soft routing itself. Post-hoc quantization raises PG19@16k from 43.5 to 62.3 at seed 0 (seeds 1–2 alike); hard top-1 training avoids this with seed-dependent outcomes. The grid gate trains at $1.125\times$ the control's step time (about 17% more on the GLA side), and the compute-matched control, though flat on WikiText, improves LAMBADA (293.49 to 257.00) and all PG19 windows.

Retrieval. No retrieval advantage is shown: the five-seed LAMBADA interval $[-38.84, +7.32]$ crosses zero (excluding the one unfavorable pair is a sensitivity analysis only), span accuracy changes by a small amount at a low level, MQAR recall is at floor for both families at gaps of at least 256, gap-128 results depend on learning rate and seed ($+0.026$, CI $[-0.198, +0.326]$), and a needle-in-a-haystack probe is at floor for both models (Appendix K). Token buckets also show losses on repeat-heavy tokens, a descriptive localization rather than a causal copy mechanism.

Mechanism and interpretation. The controls are specific to their constructions: whether a finite knot grid is necessary among continuous parameterizations remains open, the $K = 4$ optimum at one seed reverses on a second ($K = 2$ best), and the appendix configurations support no general ranking of rate families. The gate is not a new memory architecture, and knot count or allocation complexity does not indicate stored content: knot half-lives are learned rate parameters, not guaranteed retrieval scales; allocation weights are not unique fractions of stored memory; the ablations do not show that each knot stores separate content; and query-side associations are correlational. The continuous router reaches slower rates than the grid in its token-level tail (Appendix E), and the native gate's token-level rates have a heavier slow tail than the per-head summary in Figure 1a (Appendix I). The window identity fixes endpoint attenuation only, the Jensen comparison assumes identical allocations, and the survival-curve estimator measures marginal dispersion, not temporally organized memory spectra.

Evaluation coverage. Evaluation coverage materially changes the conclusion: capped PG19 shows a small gap, whereas full-shard evaluation exposes a penalty in all five seeds (see the paragraphs on transfer costs and retrieval above). Two additional LAMBADA seeds give six favorable pairs out of seven and an all-pair mean of -23.33 perplexity. A separately specified outlier-exclusion rule removes the sole unfavorable pair and yields a narrower favorable interval; we retain the all-pair distribution and treat the excluded analysis as sensitivity evidence, not the primary result (Appendix Q.2). A non-significant interval is not an equivalence result, and pre-specifying an exclusion does not eliminate its effect on the estimand.

Experimental scope. The experiments use one main training corpus, modest token budgets, and a limited set of small backbones. The 400M comparison has two seeds, the attention reference has one, and the Mamba2 training comparison is incomplete. Equal token budgets and approximate parameter matching do not ensure equal wall time; the constant-floor control is cheaper, and the GLA-side grid incurs a measured step-time cost. The post-hoc overlap scan quantifies exact 13-gram matches but does not rule out broader contamination or prove cancellation in paired comparisons. The rate bounds are learned ranges at a checkpoint, not fixed lower guarantees on retrieval distance, and the null detection result is specific to its estimator and measured checkpoints. The main gate produces one scalar rate per head ($G{=}1$) while the GLA-side variant groups key channels four ways ($G{=}4$), so the delta-backbone to GLA-side contrast is a multi-factor comparison rather than a one-factor intervention.

D Training and evaluation protocols

Matched-recipe identity. Token budget 262.1M; model width $d{=}768$ with $L{=}12$ layers and 8 heads per layer (in the GatedDeltaNet-based arms, key dimension 80 and value dimension 120 per head, i.e. value expansion 1.5); batch 8; 16k steps; the same optimizer schedule; the same evaluation

entry points (WT-103 val/test, LAMBADA, PG19 windows), with the fixed-convention validation axis (run-log minimum).

The primary 32-window evaluator (head window). The primary WT-103 evaluator of the main results uses 32 contiguous windows of length 2048 over the test shard, window starts $(i \cdot 2048) \bmod (n-2049)$, teacher-forced token negative log-likelihood (NLL) summed then exponentiated (the same procedure for LAMBADA; PG19 uses the sliding-window convention that counts the trailing stride, first 5M tokens). It covers the first $\sim$24% of the shard because it was fixed early as a light-weight primary evaluator; the pre-specified full-test-set paired result (-11.85 over 5 seeds, CI below zero; Appendix F) is reported alongside it throughout this paper. Units: the PG19 windows at $W \in \{4096, 8192, 16384\}$ score different target regions, LAMBADA is a concatenated-stream perplexity (not last-token accuracy), and the full-test-set WT-103 evaluation filters records to `len`>100 before pooling; every comparison in this paper is quoted in these units.

Embedding and output head; parameter counts. Input embedding and output projection are tied in every language-model arm, and every parameter count in this paper counts the tied $50{,}257 \times d$ matrix once (38.6M parameters at d=768). The 134M label is therefore the models' total parameter count (control 133.835M, MHLG 134.057M), of which $\approx$95.2M lie outside the embedding; counting the tied matrix separately for input and output would give 172.432M and 172.654M. The same convention applies to every language-model arm, including the 400M pair. Robustness of the results to untying or to vocabulary-size variation was not tested.

D.1 Response of the survival-curve estimator

Setting and terms. The survival-curve study (Appendix A) measures two families of Gated-DeltaNet language models trained from scratch at an earlier stage of this work, three seeds each (seeds 0–2; 12 layers, hence 36 (layer, seed) cells per family). GDN_FS ("from scratch") uses the native gate and the standard delta-rule erase; we call it the *single-rate class* and its cells the *gdn side*. BAND_FS multiplies the delta-rule erase strength β_t by a per-layer scalar $u < 1$; we call it the *erase-softened class* and its cells the *band side*. In each cell the realized survival curve is the logarithm of the mean, over heads and eight WT-103 test windows, of the running product of the erase factors $1 - u\beta_t \cos^2 \theta_t$, where θ_t is the angle between the key k_t and the window-mean key direction ($u = 1$ for GDN_FS). The estimator fits one- and two-component exponential decays to the first half of the curve and validates both fits on the second half; the *surplus* is the relative reduction in validation error from adding the second component, and the two-component criterion *fires* when the surplus is at least 0.20. The pre-specified *presence rule* required firing in at least two thirds of the 36 BAND_FS cells. The first measurement used positions 256–2047 of 2,048-token windows and a dyadic grid of candidate half-lives; the T=512 *re-run* ("redux" in the supplementary file names) repeats it on 512-token windows (positions 2–510). At that geometry, a *certification ladder* of four estimator configurations was tested: V0, V1, and V2 fit on grids of 10, 19, and 28 log-spaced half-lives, and V3 adds continuous (golden-section) refinement of the fitted half-lives to the 28-point grid. Each configuration was scored on a synthetic *stage-1 battery* of six single-rate curves (half-lives 13–400 tokens, on which any firing is a false fire) and 27 injected two-component curves, and it counted as *certified* if its power on a pre-specified subset of 12 injected curves was at least 0.8 with at most one false fire. Only V3 passed (power 1.0, 0/6 false fires); it is the *certified estimator* (also V3 estimator) used below, and its stage-2 measurement of the 36+36 realized cells gives the 34/36 firing count per class.

Estimator development. Earlier versions of the estimator with a mismatched fitting objective, mishandled small-amplitude components, or incorrectly paired sorted components failed synthetic validation. All comparisons below use the certified estimator; the failed versions are recorded in the supplementary material (`evals/h1_synth_validate_iter{1,2,3}.json`).

Threshold scan (zero-compute). Sweeping the firing threshold t over $[0.01, 0.50]$ on the stage-2 cells leaves the 34/36 count of both classes unchanged for every t up to 0.288. The two non-firing cells of each class have negative surplus (-2.59 and -0.53 on the gdn side, -0.20 and -0.11 on the band side) and fire at no positive threshold, and the third-smallest surplus is 0.3026 on the gdn side and 0.2879 on the band side. Above these values the count declines slowly, to 31/36 (gdn side)

and $30/36$ (band side) at $t = 0.50$. Detection is therefore robust for every threshold in $(0, 0.288]$ and is not an artifact of placing the threshold at 0.20; the *amount* of heterogeneity is measured by the knot-mass shares, not by this scan.

Response to rate dispersion. The 0.20 threshold is a presence detector, not a graded heterogeneity meter. Its sensitivity depends on the time scale, even non-monotonically, through three specified conditions of the certified estimator: the log-spaced grid (adjacent half-lives a factor $2^{1/3}$ apart, a step of 23% in log half-life), the refinement limited to one grid interval, and the unmasked log-domain validation. For synthetic populations of single-rate heads with median half-lives of 13, 15.3, 30, and 53 tokens (15.3 is the single-rate class's realized scale), it fires in essentially 100% of replicates once the coefficient of variation (CV) of the head rates reaches 0.5%, or for two-cluster splits with half-life ratio 1.03. At the erase-softened class's realized scale (21.0 tokens) it *underfires* below a CV of 6% and reaches 90% only at a CV of about 10%. At the $T{=}512$ geometry, the uncertified dyadic-grid configuration keeps power 0.92 on the injected two-component curves. Two consequences follow. At the single-rate class's scale, the $34/36$ firing count carries about the same information as the head heterogeneity of Lemma 2; at the erase-softened class's scale, firing is *not* produced by infinitesimal heterogeneity (it requires a CV of about 0.31). The firing fractions of the uncertified configuration at the $T{=}512$ geometry support no claim, and the outcome of the presence rule at the first geometry is not reinterpreted. On the certification ladder, densifying the grid alone increased the false fires ($5/6 \to 6/6$), whereas continuous refinement (V3) reached power 1.0 with $0/6$ false fires; the ladder and its protocol are included in the supplementary material.

A test for temporally organized structure. Because the certified estimator fires in nearly 100% of cases on very small dispersions (CV $\geq 0.5\%$) in some regimes, firing alone does not distinguish marginal rate dispersion from temporally organized multi-timescale structure. The test for the latter compares each cell with a *rate-shuffle null*: the same within-window multiset of rates with its temporal order shuffled (marginal distribution preserved, serial structure destroyed), run through the identical estimation protocol. A cell that fires at the level of its null carries only marginal-rate heterogeneity (Lemma 2); firing above the null is what temporally organized multi-timescale structure would produce. The shuffle convention and the decision threshold were fixed before the null was computed.

Rate-shuffle null calibration (fixed before the computation; marginal-only). For each realized cell we permute the per-position rate trajectory (the log erase factors) uniformly *within* each (sequence, head): every marginal rate multiset is preserved and serial organization is destroyed; the cumulative endpoint is invariant, and only the curve's shape is resampled (200 permutations per cell, random seed $48000{+}7919\,i$ for cell index i, identical geometry and 0.20 threshold). A cell counts as temporally organized if it fires while at most 5% of its shuffles fire. The certified V3 estimator is re-certified within the same run (power 1.0, false fires $0/6$), and the realized cells reproduce the recorded $34/36$ firing count per class exactly ($68/72$ cells fire; non-firing cells: GDN_FS s0-L4 and s1-L1, BAND_FS s0-L8 and s1-L11). Across the 68 firing cells the null fires at a mean share of 0.964 (median 1.000; per-class means 0.982 and 0.946), and **no** cell counts as organized; the outcome is *marginal-only*. After within-window organization is destroyed and every marginal kept, the detector still fires almost everywhere it fired before: firing is carried by marginal rate dispersion alone, in *both* classes. *Below-null check* (does any cell look *anti*-organized?): $2/72$ cells fall below their own null (BAND_FS s0-L8, surplus -0.107 against an 88.0% shuffle-fire share; s1-L11, -0.196 against 64.5%). On these two cells the realized serial arrangement is slightly *less* two-component than most of its within-cell shuffles, a weak anti-organized tail; the reverse never occurs (no firing cell has a null that mostly does not fire: the minimum shuffle-fire share among the 68 firing cells is 0.240). The marginal-only outcome is unaffected. What survives this calibration: the contrast in dispersion *level* (the erase-softened class genuinely requires a marginal CV of about 0.31), the certification ladder, the response curve, and the recorded failed iterations. What does not survive: any reading of the $34/36$ firing count as temporally organized structure for this estimator. A first iteration of this computation mistakenly used the grid-densification configuration; it lies outside the specified protocol, is kept under its own name, and is not the result (`evals/h1_rateshuffle_null_v3.json`; excluded iteration `evals/h1_rateshuffle_null.json`). *Solver check.* The certified estimator's mass fits used least squares with a nonnegativity clip, not true non-negative least squares (NNLS). An exact two-variable NNLS with explicit face enumeration (`evals/h1_`

recert_nnls_v2.py; it agrees with scipy.nnls on 300 solves, 258 of them with strictly smaller residuals than the clip) re-runs the stage-1 battery to the identical V3 certificate (power 1.0, false fires $0/6$). The realized-cell $34/36$ counts and the shuffle calibration above were measured with the clip fit. Re-measured with NNLS on the same protocol, the certificate is unchanged (power 1.0, false fires $0/6$), the identical 68 cells fire ($34 + 34$), and the outcome is again marginal-only, with mean null shares $0.9806/0.9534$ against the clip fit's $0.9821/0.9465$; the solver changes none of the conclusions of this section.

Power-curve derangement probe. Does the detector register the *strength* of organization once marginals are held exactly fixed? On four cells (per class, the two seed-0 layers whose firing surplus is nearest the class median: GDN_FS L8/L9; BAND_FS L0/L7), each realized per-(sequence, head) rate trajectory is reorganized under exact marginal preservation: (a) 200 uniform shuffles (the null band, stored); (b) block-wise value-descending orderings with block size $B \in \{16, 128\}$; (c) a fully value-descending ordering, $B{=}T$; and (d) a block-shuffle control ($B{=}16$ permutation within blocks: same locality, same marginals, no value order). Each variant is compared with its cell's null by r_q, the share of null surpluses at least as large as the variant's. Fine-grained value order is indistinguishable from the null (stored outcome org-undetectable-16; $r_q \in \{0.830, 0.075, 0.720, 0.095\}$), whereas coarse value order *inverts* the power band (org-inverted-128: $4/4$ cells with $r_q \geq 0.995$; org-inverted-full: $4/4$ with $r_q = 1.000$, three of the four surpluses negative), and the block-shuffle control does not collapse consistently ($r_q = 0.940/0.180/1.000/0.460$), which attributes the collapse to value ordering itself rather than to the loss of locality. Organization strength is therefore detectable only in the negative direction: the stronger the global monotone reorganization, the lower the surplus, and no minimal detectable organization strength exists. These measurements used an identity check repaired before evaluation (cross-process GPU re-measurement differences at most $9.6{\times}10^{-5}$ on this estimator, one cell bit-exact; every outcome stable under a worst-case shift of $\pm 2{\times}10^{-4}$; evals/h1_powercurve_v3.json).

D.2 Other measurements and allocation meters

Descriptive measurements. Each of the following was evaluated once and is reported per result: the first survival-curve measurement on the trained models (36 (layer, seed) cells, positions 256–2047, prefix fit and suffix validation, dyadic grid), which did not meet the presence rule ($6/36$ firing cells); the LM family (six checkpoints across two arms under the matched recipe, with paired tolerance checks against both controls); the allocation-meter replay on the six mixture checkpoints of both write sides (Tables 7 and 8; per-layer seed means with ranges); the on-task MQAR gap suite, run under the pre-specified endpoint-floor stop rule (24 cells planned; execution stopped as the protocol specifies once the endpoint floor and the three-seed $g{=}32$ replication were complete); the 16-bucket positional profile on 16k windows; the token-bucket analysis on 32k-token slices per domain; and the training-throughput extraction from run logs. This subsection gives protocols and descriptive evidence only; pre-specified decision rules and their outcomes are reported with the corresponding experiments in Appendix E and the appendices that follow it.

Parameterization note. The arithmetic-form arm (Section 3) changes one line of the gate, a log-sum-exp over log-weight plus log-decay; the code and its pointwise test against Jensen's inequality are in the supplementary material.

E Delta-backbone controls and complete results

Construction for MQAR (sole difference from the delta-rule control). For the MQAR recall experiments (paragraph "MQAR accounting" in Appendix A), the combined arm takes the small GDN_CM recall model (hidden 256, 8 heads, head dim 64, expand_v $= 1.0$, short convolutions, output gate), not the 134M language model, and replaces only the native smooth rate gate $g = -A_{\log} \exp(\cdot)\, \mathrm{softplus}(a_{\mathrm{proj}} x + \mathrm{dt}_{\mathrm{bias}})$ with this paper's token-routed mixture $g_{t,h} = -\ln 2 \sum_c w_{t,h,c}/\tau_{h,c}$: $K{=}4$ knots, learnable monotone grid initialized at $\{16, 64, 256, 1024\}$, router width and initialization identical to the GLA-side mixture arm (std 10^{-3} weight, staggered bias ladder $+2.0$ on knot $h \bmod K$), geometric form, per-head scalar granularity ($G{=}1$, a kernel constraint of the delta rule). Trainable parameters 6.949632M vs. the control's 6.924864M and

Table 7: Realized allocation meters of the MHLG combined arm MHLDN (three seeds, held-out WT-103 windows; per-layer seed-mean with min–max range; source `evals/h2_meters_dn.json`). The two pre-specified heterogeneity thresholds read asymmetrically on this family: the threshold on the interquartile range (IQR) of τ_{eff} (10 tokens) passes in all 36 cells; the router-entropy threshold (0.10 nats) passes in $33/36$ — all three exceptions are layer 0, whose router is single-expert on this family (85% of its corpus mass on one knot, Table 13); the GLA-side grid's meters ($36/36$ pass) stand in Table 8.

Layer	Router entropy (nats)	τ_{eff} IQR (tokens)	Allocation dispersion
0	0.066 (0.039–0.093)	216 (181–236)	0.005 (0.005–0.005)
1	0.412 (0.396–0.420)	279 (238–314)	0.008 (0.007–0.009)
2	0.385 (0.358–0.431)	284 (166–345)	0.010 (0.009–0.012)
3	0.321 (0.261–0.404)	339 (298–359)	0.008 (0.005–0.012)
4	0.210 (0.137–0.299)	372 (363–388)	0.007 (0.005–0.008)
5	0.355 (0.310–0.408)	390 (324–497)	0.012 (0.010–0.013)
6	0.300 (0.263–0.360)	1222 (631–1549)	0.013 (0.011–0.016)
7	0.317 (0.295–0.354)	803 (389–1301)	0.015 (0.014–0.016)
8	0.402 (0.333–0.455)	620 (516–706)	0.018 (0.015–0.021)
9	0.321 (0.303–0.333)	632 (450–847)	0.020 (0.019–0.022)
10	0.263 (0.225–0.299)	771 (593–1125)	0.021 (0.019–0.023)
11	0.357 (0.301–0.404)	487 (446–519)	0.021 (0.018–0.023)
All layers	0.309 (0.039–0.455)	534 (166–1549)	0.013 (0.005–0.023)

Table 8: GLA-side grid (MHL) allocation meters (three seeds, same convention; source `evals/h2_meters_a1.json`). All 36 cells pass the pre-specified heterogeneity thresholds (IQR and router entropy) of the MHLG counterpart, Table 7.

Layer	Router entropy (nats)	τ_{eff} IQR (tokens)	Allocation dispersion
0	0.227 (0.225–0.230)	353 (322–406)	0.013 (0.013–0.013)
1	0.575 (0.532–0.610)	143 (133–153)	0.013 (0.012–0.013)
2	0.706 (0.643–0.757)	200 (184–229)	0.013 (0.012–0.014)
3	0.680 (0.547–0.795)	198 (165–230)	0.014 (0.011–0.015)
4	0.685 (0.578–0.759)	224 (185–246)	0.014 (0.010–0.019)
5	0.784 (0.731–0.828)	195 (184–209)	0.015 (0.012–0.017)
6	0.740 (0.720–0.757)	234 (215–248)	0.016 (0.015–0.016)
7	0.611 (0.542–0.722)	325 (272–364)	0.016 (0.015–0.017)
8	0.532 (0.453–0.626)	368 (360–383)	0.017 (0.016–0.018)
9	0.480 (0.364–0.594)	388 (352–436)	0.021 (0.020–0.023)
10	0.561 (0.535–0.598)	321 (296–334)	0.030 (0.027–0.034)
11	0.558 (0.541–0.568)	179 (127–244)	0.033 (0.030–0.036)
All layers	0.595 (0.225–0.828)	261 (127–436)	0.018 (0.010–0.036)

the GLA-side mixture's 5.450752M. Ladders, clipping, schedules, and all probe conventions are identical to those of the pre-specified control runs.

Pre-specified read-outs and obtained numbers. The g=32 batch (three seeds, serial chain): confirmed iff mean ≥ 0.90 and $3/3 \geq 0.75$; obtained $\{0.9867, 0.9938, 0.9999\} \Rightarrow$ write-side-confirmed (mean 0.9935 ± 0.0066; per-seed closure of the $0.522 \rightarrow 0.990$ gap $[103.2, 98.6, 100.1]\%$; paired delta vs. control $[+0.016, -0.006, +0.001]$; distance-bucket profile of the best seed 1.000 on every learnable bucket; replayed evaluation reproduces the logged accuracies to all reported digits). The g=128 batch (at the pre-specified rate 3×10^{-3}): solve / instance-level (> 0.10) / floored (≤ 0.105), fixed before training; obtained $\{0.9952, 0.4590, 0.7024\}$ (mean 0.719 ± 0.269) $\Rightarrow$ instance-level, sharing the control replica's regime ($\{0.937, 0.759, 0.770\}$; exactly one seed above the solve threshold in both families) while the GLA-side mixture reads $\leq 7\times10^{-4}$ at every rate. One seed-1 run stopped mid-training because of a transient GPU error and was re-run with an identical configuration; every completed run passed its configuration and parameter-count checks.

Budget semantics on the solving models (replay probe, checkpoint cadence 3k). All four solving seeds route 77–100% of late-layer mass onto the longest knot, whose grid half-life migrates far beyond both task and window (1.5×10^5 to 6.4×10^5 tokens); the readable budget of the combined arm therefore spells out *quasi-frozen store with key-addressed selective erasure* — the implicit accounting the delta-rule control performs without exposing it. Acquisition dynamics: deepest-layer long-knot share $0.471 \to 0.697 \to 0.749 \to 0.760$ at steps 3k/6k/9k/12k (final 0.760), seed-mean; K_{eff} settles 1.0–1.5 on this single-timescale task, the concentration the task's fixed-gap structure warrants (compare the mixed-task LM allocations, Appendix L).

Allocation as predictive state for retrieval (descriptive, pre-specified read). On the combined arm's obtained $g{=}128$ checkpoints, we score gate features at value and query positions as classifiers of per-instance retrieval success (Mann–Whitney AUC over 28,818 supervised positions, seed 2; base rate 0.704; AUC: area under the ROC curve). At the deepest layers the query-position signals invert far below chance: long-knot mass AUC 0.230 and effective half-life AUC 0.219 at L03 (L02: 0.398/0.351), while router entropy reads 0.574 mildly above; value-position features sit at chance. Successful retrieval is marked by *de*-commitment of retention at the query — the write side does the key-addressed read; the gate's readability exposes that hand-off to downstream analysis without probing internals. *Causal grading (clamp replay).* Re-running the same checkpoints with the long knot's mass clamped to zero *only at query positions* moves accuracy by at most 0.04 points across three seeds ($-0.00/+0.01/-0.04$pt), while clamping it at *every* position collapses the instance-level seed by 15.5pt (0.459 → 0.304; the solved seeds move $\leq$0.2pt). The de-commitment is therefore graded *correlational at the query*: long-knot capacity acts at non-query (encoding/write) positions; the query-time allocation marks, but does not mediate, the read (`evals/h2_mqar_clamp_replay.json`; stored outcome correlational-at-query).

LM-scale instantiation (MHLDN). The GLA-side router design is re-hosted on GDN_CM at full LM configuration: $d{=}768$, $L{=}12$, $H{=}8$ heads of key dimension 80 (value expansion 1.5), the per-head scalar $G{=}1$ kernel constraint, router width and staggered ladder unchanged; 134.057M parameters with the tied embedding counted once (172.654M if it is counted separately for input and output; control 133.835/172.432M, +0.17%). Five seeds of the byte-identical 262.1M-token recipe; every run passed configuration and parameter-count checks. Results are taken verbatim from `evals/mhldn_lm_family5x3.json` (Table 1); training integrity 16,000/16,000 steps per seed (best validation among the five combined-arm runs 41.30, seed 1). *Two-stratum streams (seed 2, full-corpus convention).* Paired against the stored seed-2 per-token streams of both LM arms: vs. control -0.0908 nats full-corpus (-0.0855 on the 208,346 non-head positions, amplitude 94%), per-octile $\{+0.126, +0.082, +0.028, -0.009, -0.052, -0.159, -0.315, -0.426\}$; vs. the GLA-side arm -0.0414 nats, per-octile $\{-0.112, -0.111, -0.077, -0.064, -0.045, -0.038, -0.023, +0.138\}$. Cross-read: the octile-0 copy penalty drops from the GLA arm's $+0.238$ to $+0.126$ (write-side component 0.112 nats = 47%; residual gate-side), the octile-7 gain keeps 75% amplitude ($-0.565 \to -0.426$), and the between-write-side gap on that octile is $+0.138$ — the novel-tail advantage is gate-majority, its amplification host-dependent. *Parity is write-side.* On PG19 the combined arm's per-seed paired deltas against the control are sign-mixed with every CI spanning zero, and the control itself shows no copy-window penalty: once the write side is delta, the gate is indistinguishable from inert on this surface — the copy locus is handled by the delta rule regardless of gate. *Non-head slice (five seeds, positions $\geq$65,536).* The GLA arm's own full-corpus octile trend re-sliced outside the head region: octile-7 delta -0.470 nats mean, negative in $5/5$ seeds (per-seed -0.416–-0.550); octile-0 penalty $+0.198$–$+0.320$ persists; the head concentration of the GLA-side WikiText result is thus an amplitude statement, not a location statement (`evals/h2_lm_nonhead_slice.json`). *LAMBADA exclusion sensitivity.* The seed-3 pair — whose control cell (262.9) falls outside that family's own 293–311 band — is recorded, not excluded; excluding it leaves the verdict intact: $n{=}4$ LAMBADA mean -32.53 ppl, bootstrap 95% CI $[-41.70, -26.40]$, $4/4$ below zero. The WT-103 head-window verdict is likewise robust to the same exclusion (-11.79, CI $[-14.61, -8.47]$). Source: `evals/mhldn_lm_lambada_excl.json` (single evaluation).

The completed 2×2 (gate $\times$ write side). The missing cell — GLA write side under the native single-rate gate at LM scale — was trained for this comparison (three seeds, byte-identical recipe, native GLA gate, 133.897M parameters, counts checked); the single-rate checkpoint families of the survival-curve study (Appendix D.1) come from an earlier stage of this work and are not reused as controls. WT-103 (nominal-32, seeds 0–2 mean): GLA$\times$single 134.27, GLA$\times$MHL

122.80, delta×single 128.03, delta×MHL 117.55. Gate effect on the GLA write side -11.46 (CI $[-15.03, -5.60]$), on the delta side -10.48 (CI $[-13.48, -6.80]$) — same sign, overlapping: the two factors carry jointly. The write side's effect likewise holds on both gates (-6.23 single, -5.25 mixture): the LM result decomposes *nearly additively* into a general gate gain (≈ -11 ppl on either write side) and a general write-side gain (≈ -5–6 ppl on either gate). Secondary axes per seed are recorded in the single-evaluation source `evals/h2_gla_cell_agg.json`.

Protocol compatibility and the interaction term (zero-compute re-read). The factorial cells use the head window over seeds 0–2, whereas the main comparison of Section 4.2 uses five seeds and reports the full test set alongside the primary head window. The two analyses differ in both evaluator and seed pool, and we report them separately rather than pooling them. Where they overlap they agree: on the shared seeds, re-scoring the same checkpoints on the full test set changes no per-seed sign of either comparison (Appendix F, reading (ii)), and the seed spread of reading (iii) covers the factorial seeds. The interaction term itself: the gate effect differs across write sides by $+0.982$ ppl at seed means (-11.464 GLA-side vs. -10.482 delta-side), per-seed signs $\{+6.96, -5.56, +1.55\}$ — sign-mixed, and at n=3 the bootstrap range is the seed range itself ($[-5.56, +6.96]$): a true interaction is neither visible nor excludable at this n, and "nearly additive" stands as the three-seed bound, not a smoothness claim. On the full test set over five seeds, the two write sides differ in how the gate effect carries over: on the delta side it carries over one-to-one (-11.85, CI $[-15.1, -8.4]$, $5/5$, vs. -10.48 at the pre-specified head-window design point), whereas on the GLA side the gain is concentrated in the head region and shrinks at full coverage (-3.23, CI $[-5.66, -0.79]$, $4/5$, vs. -11.46). The factorial decomposition is therefore restricted to its pre-specified head-window design point, and only the delta side (MHLG) keeps the gate effect under both evaluators of Section 4 (result files: `mhldn_lm_fullcorpus5.json` and the seed extension of Appendix P).

Gate-capacity isolation (GDNCAP). The mixture gate's input-dependent router carries 24,640 parameters per layer against the capacity gate's 6,160 ($4\times$). The capacity-matched control GDNCAP (shared `A_log`/`dt_bias`, fixed-uniform C=4 softplus channels, no trainable routing, 134.0564M vs. 134.0566M parameters, bit-exact at initialization) records WT-103 127.605, LAMBADA 296.299, PG19 47.492/46.467/43.967 (vs. the capacity control 123.080/293.489/46.865/45.467/43.023): no recovery on any axis — the WT-103 recovery is -4.53 ppl against the 6.799 ppl seed-0 gain, $r = -0.67$ (rule regimes pre-specified: ≤ 0.25 constraint-carries / ≥ 0.75 capacity-carries / mixed). The $4\times$ gate-path capacity asymmetry is ruled out as the combined arm's account; the gain depends on what the router routes. *Initialization correction.* The channel-projection initialization, specified as std 10^{-3}, was reset by the backbone's global initialization to its default (0.0195); the evaluated cells therefore trained at the backbone default, and the $9/9$ result of the multi-seed paragraph below describes the arm as trained. A re-trained cell with the specified initialization (verified std 9.60×10^{-4}) reads WT-103 126.825 (pre-specified statistic $s = -0.551$), LAMBADA 282.20, PG19 43.76: below the 0.25 threshold, closer to the family control yet still without recovery, so the $9/9$ capacity-rejection result survives the correction. *Anchor-seed sensitivity* (Appendix G): the re-trained cell changes outcome at both alternative anchors ($-0.551 \rightarrow +0.253$, partial, at the seed-1 anchor; $+0.83$, carries, at the seed-2 anchor), so its insufficiency is relative to the anchor within the recorded ± 0.8–1.4 anchor-seed spread; the pre-specified anchor convention stands, and this cell joins RETNET and A1B among the anchor-sensitive results.

Maintained-basin clamp counter-arm (GDNBSN). The *maintained* slow basin without any routing: a parameter-frozen GDNCAP subclass whose sole change surface clamps one fixed channel per group (25% of C=4) to a slow floor, $\mathrm{chan}_{\mathrm{slow}} = \min(\mathrm{chan}_{\cdot,3},\ (1/256)\ /\ \exp(\mathtt{A_log}))$ with the other three channels left native-continuous and zero new parameters (identical parameter count, 134.0564M); the clamp is identity-active in 99.16% of sites at initialization. Under the pre-specified three-outcome rule ($s = (123.080 - o)/6.799$: ≥ 0.5 basin-suffices / ≤ 0.25 routing-load-bearing / partial), a single evaluation of the full evaluator suite returns WT-103 128.069, hence $s = -0.734$ (≤ 0.25) — **routing-load-bearing**: the routing-free control recovers nothing and ends 4.99 ppl above the family control, matching the capacity control ($r = -0.67$) and the rejected ladder init-prior ($+1.62$, CI spanning zero). Descriptive axes (no thresholds): LAMBADA 289.444, PG19 46.414/45.385/42.950. A GPU replay probe of the trained state measures the clamp still binding at 73.11% mean (per-layer 0.378–0.978). All three non-routing variants with the router's degrees of

freedom (capacity, initialization prior, maintained clamp) fail: the gain depends on what the router routes, at router-comparable parameter count. *Implementation note*: the specified floor was computed as a per-unit rate $1/256$ — the clamped channel's realized guarantee is half-life ≥ 177.45 ($\ln 2 \cdot 256$), not $\tau \geq 256$; and the four channels are averaged *before* the shared state update, so one clamped channel maintains no independent slow subspace (a direct test of the realized combined half-life reads 1.33). This arm is therefore not a maintained-25%-basin witness: its three-seed failure stands as a routing-free capacity-class reading, and the clamp-limit reading below transfers to its slow-clamped routed variant HYBCS.

Trained continuous-mixture control (GDNRM). This control tests whether *any* trainable continuous gate suffices when the router stays input-dependent, same-width, and trained. A parameter-frozen GDNCAP subclass whose sole change surface reuses each (token, head, group) router logit vector twice: softmaxed into mixture weights w and softplused into continuous per-channel rates; the applied gate is the w-weighted mean over channel rates instead of GDNCAP's fixed uniform mean over channels — $g = -A_{\exp} \cdot \sum_c w_c \, \mathrm{chan}_c$. Identical tensor shape, width, and parameter count as GDNCAP (24,608 per layer); its initialization is specified in the supplementary code. Under the pre-specified three-outcome rule (same anchor form $s = (123.080 - o)/6.799$, thresholds ≥ 0.75 trained-mixing-suffices / ≤ 0.25 gating-alone-insufficient / partial), a single evaluation of the full evaluator suite returns WT-103 128.563, hence $s = -0.806$ (≤ 0.25) — **gating-alone-insufficient**, the weakest of the four non-discrete variants (capacity -0.67, maintained clamp -0.73, trained mixture -0.81, CI-crossing). With the continuous-routing control also negative, the evidence goes against gating per se: the gain travels with the discrete knot grid and the rate-hull routing structure. Descriptive axes (no thresholds): LAMBADA 284.004, PG19 47.020/46.139/43.461.

Multi-seed replication of the three negative controls. Each negative control was re-trained at seeds 1 and 2 byte-identical to the seed-0 cell (same recipe, same decision rule $s = (123.080 - o)/6.799$, same thresholds), so that the paper's sharpest attribution rests on three independent draws per class. WT-103 (s) per seed: GDNRM 128.563(-0.806), 125.131(-0.302), 126.373(-0.484); GDNCAP 127.605(-0.666), 127.811(-0.696), 126.267(-0.469); GDNBSN 128.069(-0.734), 127.846(-0.701), 127.872(-0.705). Counting the already-evaluated seed-0 cells, 9/9 replication cells sit at or below the 0.25 threshold, and each family's t-based 95% confidence band on the seed-mean s lies fully below it — GDNRM $[-1.166, +0.104]$, GDNCAP $[-0.916, -0.304]$, GDNBSN $[-0.758, -0.669]$ (seed sd 0.256/0.123/0.018). The gating-alone-insufficient reading is therefore three-seed supported on all three independently implemented classes; the clamp arm additionally exhibits the tightest seed dispersion of the suite (seed sd 0.018 in s, 0.12 ppl), consistent with its capacity-reallocation failure channel being deterministic rather than seed-dependent. Descriptive axes per seed, no thresholds (LAMBADA; PG19@16k): GDNRM 284.00/273.51/268.08; 43.46/43.46/43.35 — GDNCAP 296.30/276.90/266.64; 43.97/44.59/42.71 — GDNBSN 289.44/268.37/272.82; 42.95/44.51/45.17.

Rate distribution of the trained continuous gate (GDNRM vs. MHLDN). This paragraph measures the router-matched continuous gate's *converged* per-window rate distribution directly. The MHLG-side reading replays both final-checkpoint classes on the pre-specified meter convention (8×2048 held-out WT-103 windows, realized gates taken on each layer's exact forward input through the layer's own gate path — for MHLG, the gate function `compute_gate_mhl` of the supplementary code; sources `evals/h2_ratedist_dn.json` and `evals/h2_ratedist.json`, one reading each, rules fixed before computation) and displays it directly (Table 3); the majority-of-layers median rule holds (10/12), the narrower-dispersion rule does not. (The same statistics measured on the GLA-side grid MHL are recorded separately as measurements of MHL; the orderings they suggest do *not* transfer across write sides, as each axis below shows.) The two reachable distributions differ on three measured axes, whose directions the MHLG-side read reverses relative to the GLA-side grid's. (i) *Median side*: the short shift dominates by median and strengthens — the continuous router's layer-mean median statistic runs 103.4–114.1 (three seeds) against MHLG's 185.3–211.7, the pre-specified majority rule holding in 10/12 layers, and its fast share is roughly doubled (55.2–56.3% vs. 25.0–31.4% at $\tau \leq 16$); the continuous router's early layers converge to sub-token-to-few-token medians (τ_{q50} 0.40–2.05 at L1–L3), against MHLG's 8.8–90.2 there. (ii) *Tail side*: the slow mass does *not* collapse on the continuous router (26.3–28.3%

at $\tau \geq 256$) — but on MHLG the routed hull keeps *more* slow mass, not less (45.5–48.0%, $\approx 1.7\times$ the continuous router's in every seed pair): the slot-anchored hull occupies the slow band more, not merely more stably. (iii) *Width*: the continuous router's pooled log-rate dispersion is the wider one in all three seed pairs (2.76–3.01 vs. 1.61–1.65) — on the MHLG-side read the hull constraint compresses the realized distribution (the opposite ordering holds for the GLA-side family, whose finer $G{=}4$ channel-group granularity and GLA write side change the pooled statistic; the router-vs-grid dispersion ordering is therefore write-side dependent, and on MHLG it is the continuous router that realizes the wider set). What separates the arms on the MHLG-side read is therefore not a trade between slow-occupation level and slot guarantees — the hull leads on both — but the continuous router's reach into a structurally unattainable joint configuration: trained input-dependent mixing keeps a heavy token-conditional slow tail (τ_{q90} reaching 1.2×10^5 tokens — 33/36 *gate-side* (seed, layer) pairs at $\geq 10^2$, well past the grid cap) and still loses the gain ($s \leq -0.30$, three seeds). The floor comparison sharpens accordingly: the constant slow-mass floor pins a simplex coordinate at 0.5 by construction — occupation *stability*, undrifting by design at the coordinate level (construction in the RFLONG paragraph below) — whereas the continuous gate's slow share is an emergent tail: substantial in level, unconstrained in structure. Level, stability, and trajectory are distinct axes, and the write side moderates the first. This record is descriptive: it displays the distribution and changes no threshold-counted claim. The distribution itself is Table 3 in Section 4.4 (sources `evals/h2_ratedist_dn.json` for the grid gate and `evals/h2_ratedist.json` for GDNRM, one reading each). For comparison, the same statistics on the GLA-side grid MHL (seeds 0/1/2) are: median 86.1/95.0/93.6, dispersion 1.351/1.346/1.365, slow 0.148/0.180/0.173, fast 0.266/0.248/0.239, win. 0.074/0.067/0.069; they are recorded at `evals/h2_ratedist.json`.

Pinned multi-scale structural control (RETNET). The antecedent class named in the introduction (RetNet), as a direct ablation: fixed multi-scale decay with no learnable grid and no token routing (MultiScaleRetention layers trained under the common recipe, seed 0, byte-identical recipe and evaluator). Measured parameter count 152.013M (+13.6% over the combined-arm class — inside the pre-specified 15% band; capacity direction recorded, no equal-capacity claim made for this arm). Under the pre-specified three-outcome rule ($s = (123.080 - o)/6.799$, thresholds ≥ 0.75 fixed-grid-suffices / ≤ 0.25 fixed-grid-insufficient / partial), a single evaluation returns WT-103 122.426, hence $s = +0.096$ — **fixed-grid-insufficient**: at this scale and recipe the pinned implementation recovers $\sim$10% of the anchored gain. This says a measured difference exists for this configuration; it does not by itself establish which components are individually necessary (the routing and the learnable grid are jointly on the positive side of every negative control we have), and a strong showing here would not have subsumed the contribution, only relocated it. Descriptive axis: the same arm posts the strongest LAMBADA of any arm measured (256.072 vs. 262.62 combined arm, 293–311 family control) while its PG19 windows sit at the GLA-side penalty level (50.5/53.4/54.2 vs. delta-side 43–49). We keep it as a simplified/hybrid-method candidate on the LAMBADA axis and a cost account on the long-context axis, with no claim beyond the measured cells. At seed 1, however, the same recipe returns WT-103 114.748 ($s = +1.23$) — at one seed the pinned antecedent recovers beyond the anchor ratio and posts the strongest non-attention WT-103 cell we measured, while its LAMBADA again matches the combined arm (262.12) and PG19 stays penalized (52.96/51.29/48.62). We therefore record the pinned arm as a simplification/hybrid candidate with a seed-fragile gain, and the seed-0 insufficiency has no standalone reading — it is one configuration inside a mixed family, not a component-level verdict (parameter count 152.013M at both seeds).

Matched-capacity pinned antecedent (RETNETM; insufficient at $n{=}3$). The remaining isolation trains an equal-parameter version of the antecedent, training the same MultiScaleRetention class at `expand_v` = 1.125 (RETNETM), whose measured parameter count is 133.400M — 0.5% *below* the combined-arm family's 134.057M, i.e. the comparison is conservative for a positive showing by the antecedent. Both seeds of the byte-identical 16k recipe collapse: WT-103 132.217/134.099, hence $s = -1.34/-1.62$ — both far below the 0.25 threshold — with the long axes equally collapsed (LAMBADA 306.36/297.43; PG19 53.98/53.22/50.65 and 57.06/56.17/53.48 at $4k/8k/16k$). At matched capacity the antecedent class holds neither the WT-103 gain nor either long axis at either seed, so the parameter count is not what separates the families. Boundary: this verdict does not touch the seed-mixed antecedent result — the 152M original-parameterization s1 cell (+1.23) stays recorded as the strongest simplification/hybrid candidate — and it changes no reading of the main result; it pins only the matched-capacity cell. A third

seed under the pre-specified stability rule (single-use) yields the same class again: WT-103 126.381 ($s = -0.49$), LAMBADA 283.48, PG19@16k 51.30 — three of three matched-capacity cells sit far below the partial threshold (retnetm-insufficient-n3; `evals/h2_retnetm_n3.json`), so the instability resolves at matched capacity as stable insufficiency, not seed dependence.

Long-context cell (context-independent). The mechanism question: with the training window at 2048, what gradient signal could hold learned knots at 1.5–640×10^3 tokens? The natural second cell re-trains MHLDN s0 at T=8192, 4000 steps — a byte-equal token budget ($8\times8192\times4000 = 8\times2048\times16000 = 262{,}144{,}000$), same recipe otherwise (param stamp inside the 0.6% band of 134.057M; gradient accumulation 2×4 microbatching, effective batch unchanged). The pre-specified statistic compares mean-over-layers $\log_2$ of the per-layer largest knot half-life: pre-training reference value 9.708 (anchor domain $[3.98\times10^2, 4.20\times10^3]$ tokens, not recomputed). Evaluation (`evals/h2_lm_longctx_grid.json`): long-context per-layer maxima run 809–1302 tokens — every layer inside the anchor domain — mean $\log_2$ 10.038, hence $R = +0.330$ and $R_{\text{lin}} = 1.257 \geq 1/2 \rightarrow$ context-independent: quadrupling the training window moves the learned knots by $\times1.26$, nowhere near the $\sim\times1/8$ boundary-escape scenario; the far slots are held without in-window supervision (a conservative allocation well, as if the surplus mass simply has nowhere else to go). Descriptive axes (no thresholds): WT-103 160.70, LAMBADA 304.69, PG19 $53.32/52.21/48.94$ — the $4\times$ window at one quarter the optimizer steps does not rescue the long-form axes, and we record that as descriptive, not as a claim.

Post-hoc value-quantization isolation. A stress test asks whether knot-sharing or hard quantization carries the signal. A zero-training answer: take the *evaluated* GDNRM checkpoint, keep its trained router logits, and at evaluation snap each (t, h) halflife to the nearest dyadic knot $\{16, 64, 256, 1024\}$ in log-space — hard-quant values, no mixture, no hull training. Under the pre-specified rule (same s formula and thresholds), the unwrapped reproduction first reproduces 128.567 (tolerance 0.2), then the quantized pass returns WT-103 109.389, hence $s = +2.01$ — quant-values-suffice: value commitment alone, from the same router information, out-performs every arm on the short axis and trails only the attention anchor (103.6). The long axis says why the sentence cannot stop there: LAMBADA improves (284.0→271.9) but PG19 collapses hard (43.46→62.30 @16k; $68.75/67.85/62.30$ across windows), with knot occupancy $\{59.7, 7.9, 8.3, 24.1\}\%$ — 24% of (t, h) mass pinned through the slowest knot at evaluation time is what produces the copy-token gain and destroys the smooth long tail. Reading: the short-context gain does *not* require the knot-shared hull — hard per-token rate commitment is enough; the learnable hull plus routing is what keeps both axes alive at once (measured probe `evals/h2_lm_quant_gdnrm_probe.json`; single draw). The multi-seed extension (`evals/h2_lm_quant_gdnrm_seeds.json`, one evaluation) reproduces *both* outcomes at seeds 1–2: bit-stable unwrap reproductions ($|\text{err}| \leq 0.01$) and $s_q = +0.90/+1.87$ — quant-3seed-suffice; and the PG19 collapse replicates at all three seeds ($64.83/65.25/62.30$ at 16k against the trained family band 43–48), with occupancy spectra matching across seeds ($\simeq61\%$/7–8%/$\simeq24\%$; LAMBADA mixed-descriptive $279.69/258.82$). The trained hard-commitment arm (next paragraph) qualifies the second part of this reading: keeping *both* axes does not require the soft mixture either — it requires that the evaluated commitment be the one the checkpoint was trained with. An alternative reading of the same data is recorded as well: the soft hull may act purely at training time, protecting the tail strata from short-rate drift during optimization while the evaluation geometry itself is indifferent — on this reading the post-hoc snap fails because the optimizer never learned the tail content at the snapped points, not because evaluation is intolerant of the commitment. Without the refit below, the evidence does not separate the two readings; the training-time masking results of Appendix I point towards the training-time one. *Head-only refit* (single evaluation): the pre-specified instrument unties the weight-tied readout row (bit-identity at init), freezes the entire body, and refits that row alone for 1,000 steps on an identical batch stream under two geometries. The protocol is live — on the unwrap control it holds WT-103 (128.567→129.044) and PG19@16k (43.455→44.065) while adapting descriptively on LAMBADA (284.0→275.8); under the snapped geometry the same protocol moves PG19@16k by $+0.016$ — recovered fraction $f = -0.001$ of the 18.84 ppl gap — while the $4k$/$8k$ windows improve 1.8/1.1 ppl (descriptive). Readout adaptation alone recovers the local content and cannot reach the lost tail content: the evidence narrows the two readings to the training-side pair — either the optimizer never wrote that content at the snapped points, or it is bound to the trained value geometry itself. The evaluated-commitment reading stands with this qualification (rules fixed before the run; six gates passed; source `evals/h2_quant_refit_agg.json`).

Trained hard commitment (MHLDNH; seed-fragile at n=3, confirmed at n=5). The isolation, trained: the same MHLDN skeleton — same router, same learnable dyadic grid, byte-identical recipe and evaluator — but the gate is the *hard* top-1 knot per token with straight-through router gradients (same parameter count, 134.057M). Under the pre-specified joint two-axis rule (s thresholds on WT-103; PG19@16k hold band $\leq$46.0 set against the family 42.95–43.97/G4 45.40 penalty band and the 49.10/QUANT 62.30 collapse band), the joint rule was evaluated on seeds 0–2 and the WT-103 axis on five seeds. On the WT-103 axis, seeds 0 and 1 reach $115.309/113.367$, hence $s = +1.14/+1.43$ — beyond the anchor ratio and ahead of the soft combined arm's own seed-0 cell — and seed 2 reaches 118.977 ($s = +0.60$, a partial cell at 80% of the rule's $+0.75$ threshold), so the pre-specified three-seed rule reads seed-fragile (two strong cells, one partial cell). Seeds 3–4 at the anchor recipe reach WT-103 $119.314/115.108$, and the five seeds give $s_a = +1.14/+1.43/+0.60/+0.55/+1.17$ — all five positive, mean $+0.980$ — so the pre-specified five-seed rule confirms the WT-103 gain, with the two partial cells inside the spectrum (`evals/h2_lm_req3_n5_agg.json`; paired own-seed mean $s_p = +1.77$). The long axis holds at all three seeds of the joint rule: PG19 $48.61/46.93/43.72$, $49.67/48.37/44.54$, and $48.19/46.91/43.76$ at seeds 0–2 (in-band at 16k, with a mild short-window penalty), with LAMBADA $248.97/271.30/265.81$ (the G4 s2 cell, 245.60, is lower than the seed-0 cell; descriptive). The arm is the strongest hard-assignment simplification candidate: it changes none of the negative-control attribution — the arm still routes over the learnable knot grid — and it moves the two-axis boundary onto *trained-with*, not *soft*. Descriptively the long axis is a spectrum at n=5 as well: seed 3 reaches PG19@16k 46.531 (above the $\leq$46.0 hold band that the rule set for the first three seeds) with LAMBADA 295.684, and seed 4 reaches $43.259/265.078$; the hold-band reading above covers seeds 0–2.

Per-head static-rate control (STATICH; partial then two-seed strong at n=2). A further control: a per-head *learned constant-rate* vector on the same delta-rule skeleton — no router, no mixture (per-head τ_h initialised on a dyadic ladder 16→1024 spread across the eight value heads, then free to learn; param stamp 133.761M, -0.22% vs. the family, conservative side, the delta being the removed router). Under the same-rule single-seed reading: WT-103 119.214, hence $s = +0.57$ — **PARTIAL**: a static per-head rate vector recovers part of the gain, not all of it; we record the measured difference and infer neither that routing is unnecessary nor that time-varying rates are individually necessary. Its own drift is the mechanism hint (`evals/h2_statich_tau_drift.json`): the longest head (init τ=1024) ends at 0.21–0.46× of initialisation in all twelve layers (218–468 tokens) — the short-rate attractor fully saturates when constant rates are left free without routing, the long slots are abandoned, and the initially monotone ladder survives in only $7/12$ layers. Read alongside the account above: continuous or free rates shorten themselves; it is the routed allocation that keeps long slots occupied (LAMBADA 288.10, PG19 $50.72/49.07/45.33$, both descriptive). The multi-seed extension (STATICH seed 1, byte-identical recipe) reads WT-103 116.227, hence $s = +1.01$ — two-seed strong under the pre-specified three-way rule ($\{+0.57, +1.01\}$: one partial, one above threshold — a strength spectrum, not a flip). The static arm's short-context ability therefore stands at both seeds; its long-side gap persists (LAMBADA 291.86 vs. the combined arm's 264.89; PG19 $49.28/47.38/44.67$, i.e. $+1.06/+0.60/+1.28$ over the same-window family band — routed allocation still uniquely owns the long axes; this single-arm spectrum asserts nothing about what the antecedent could do at other operating points).

Clamp-basin + routed-free hybrid control (HYBCS; seed-fragile at n=3). A direct combination control: decide the clamp basin with learned router-width continuous rates on the remaining channels — isolating whether routing over the discrete hull beats clamped-slow + free-fast under matched input-dependence. Design, fixed before training (pre-specified): on the same delta-rule skeleton, value heads 0–3 are hard-clamped at τ=1024 (content-independent constant; the short-rate attractor is removed by construction on this half), heads 4–7 carry input-dependent continuous rates from a narrow router (768→4, sigmoid-bounded into the same hull 16–1024 tokens, bias-init ladder $\{27, 76, 215, 610\}$); param stamp 133.798M (-0.19% vs. the family, conservative side; gate block 3,084 vs. 24,640 params/layer), byte-identical 16k recipe, seed 0. Under the same-rule single-seed reading: WT-103 114.323, hence $s = +1.288 \geq 0.75$ — hybrid-carries: at matched input-dependence a basin-stabilized slow half plus routed continuous rates on the rest *suffices* at this scale. We record the narrowing, not a refutation: the family result, the nine negative-control cells and the QUANT/MHLDNH rules are untouched, and this arm becomes the strongest simplification candidate — what the gain selects is a slow basin that cannot drift, jointly with routed adaptivity elsewhere, not the hull's discreteness itself. The long axes hold where the static control paid: PG19

48.22/46.78/43.39 (family band inside @16k; STATICH 45.33), LAMBADA 264.89 (STATICH 288.10; descriptive). Its placement read (`evals/h2_hybcs_tau_drift.json`) closes the loop against STATICH: free-side bias-position taus end within {23–29, 46–89, 214–228, 601–613} of the init ladder across *all* twelve layers — no attractor collapse on the routed free half — while router weight norms grew $\sim$100$\times$ from the 10^{-3} init (0.3–2.9; live input-dependence, not a frozen allocation). The dynamic occupancy of the same checkpoint was not measured. At seeds 1 and 2 under the pre-specified rule ($s1$: WT-103 120.190, $s = +0.425$; $s2$: 121.337, $s = +0.256$): the three-seed joint reads hybrid-seed-fragile — positive at all three seeds ($+1.29/+0.43/+0.26$), the full carry is the strongest-seed event, and the 12/12-layer placement retention evidences at n=1 only. The mechanism-positioning conditional (full multi-seed carries) does not trigger; what survives is the directional claim — stable slow basin plus input-dependent fast routing yields a strictly positive, seed-fragile share — recorded, not upgraded. Descriptive axes at $s1/s2$: LAMBADA 267.61/289.40, PG19 49.48/47.79/44.46 and 48.52/47.18/44.74. At n=5 (seeds 3–4 same recipe; WT-103 120.592/119.137): s_a $+1.29/+0.43/+0.26/+0.37/+0.58$ — every cell positive, mean $+0.583$ below the carry threshold (`evals/h2_lm_req3_n5_agg.json`). This all-positive, below-threshold spectrum is the arm's five-seed result; the paired own-seed reference mean $s_p = +1.37$ is reported alongside it, following the anchor-seed sensitivity analysis (Appendix G).

Router-free long-knot floor control (RFLONG; seed-fragile at n=3, confirmed at n=5). The cheapest decisive probe of routing's worth, trained: a *constant* mass floor ε=0.5 per token pinned on the slowest knot (mass frozen, τ still learnable on the shared grid), the router distributing only the remaining 0.5 over the three fast knots (same skeleton and router initialization; parameter count 133.983M, -0.055% vs. the family, the delta being the narrowed router width). Three seeds under the same pre-specified rule: WT-103 111.96/120.95/114.32, hence $s = +1.64/+0.31/+1.29$ — rflong-3seed-fragile: positive at all three, two above threshold, the partial cell the mid-seed event. The mirror verdict narrows the attribution once more, at token granularity: the slow side of the gain needs a *guaranteed* mass share on a slow knot that cannot drift, *not* a routed allocation over it; routing's residual worth sits on the fast side (still input-dependent — the trained gate logs anchor the effective-rate band at $p10/p90 \approx 213/265$, not at the attractor region). The long-side spectrum reads hybrid-like (LAMBADA 261.60/270.50/284.26; PG19 48.57/48.12/48.41, 47.17/46.86/47.04, 44.09/43.10/44.15 at 4/8/16k): no arm in this family escapes seed fragility on the descriptive axes; the five-seed joint thresholds directly below give the family-level reading. With two further seeds (3–4 at the anchor recipe; WT-103 110.01/112.76), the five seeds give $s_a = +1.64/+0.31/+1.29/+1.92/+1.52$, *all five positive*, mean $+1.336$ — confirms-n5 (`evals/h2_lm_req3_n5_agg.json`); the paired own-seed reference reads stronger at mean $s_p = +2.13$. At n=5 this arm is therefore classified as a seed spectrum rather than seed-fragile: one below-threshold cell (s_2), four above, no negative cell. *Five-seed joint thresholds* (zero-compute from existing data, same-seed GDNCM controls; `evals/h2_routeuniqueness_n5.json`): WT-103 axis gains (own-seed ΔWT-103) $+11.46/+9.33/+14.45$ means for routed/hybrid/floor, anchor-s means $+0.896/+0.583/+1.336$; LAMBADA gains $+19.70/+21.33/+23.14$ means with the routed arm carrying both the single deepest cell ($+46.5$ at seed 4) and the recorded seed-3 outlier pair; PG19@16k mean penalties $+0.21/+0.57/+0.36$ ppl; joint all-axis holds (in the same seed, a WT-103 gain, a LAMBADA gain, and a capped PG19@16k penalty of at most 1 ppl) at 3/5 seeds apiece ($\{1,2,4\}/\{0,1,4\}/\{1,3,4\}$). Routing's uniqueness on the WT-103 axis therefore dissolves at n=5 — the constant floor reaches a *larger* mean gain with matched joint holds; the routed hull's measured remaining advantages are the smallest long-context penalty, the deepest single held-out cell, and the readable budget itself. Practically: the constant floor is the cheaper way to the gain; the full gate earns its $\approx$12.5% extra step time on the delta backbone (ratio 1.125 below; the $\approx$17% figure is the GLA-side measurement) only where the long axes or the budget itself is the deliverable. *Cost-normalized gains* (partial; zero-compute from existing logs: training-log tok/s medians at steps $\geq$2000 against the five-seed deltas; `evals/h2_costletter_n5.json`): pricing each arm's WT-103 axis gain against its measured step-time ratio (1.125/1.065/1.086 routed/hybrid/floor) yields 10.2/8.8/13.3 WT-103 ppl per unit cost with paired five-seed bootstrap CIs $[7.8, 12.4]/[7.6, 10.0]/[9.6, 16.5]$ — the constant floor stays the cheapest route to the gain on this normalized account as well, the three CIs overlap pairwise, and the hull's remaining advantages are unchanged: smallest long-context penalty ($+0.21$ vs. $+0.57/+0.36$), deepest held-out cell, readable budget.

Learned-knot drift of the floor arm at $n{=}5$ (zero-compute). Does the floor's larger gain come from a more quasi-frozen store? Reading on CPU the five trained floors' learnable grids (12 layers $\times$ 8 heads per seed): the slowest knot drifts *farther* than the routed arm's, not less — family slow-knot mean 776.6 vs. 467.5 tokens (median 515.6 vs. 389.0), share beyond the 2048-token training context 6.5% vs. 1.3%, per-seed deepest head reaching 6,012–10,153 vs. 2,298–4,202; the span ratio widens accordingly (median 53 vs. 44). Yet the frozen tail stays a minority (93.5% of slowest knots remain inside context; per-layer means 310–2,027), and the fastest knots hold the same sub-token band (mean 9.4 vs. 8.2). The quasi-frozen slow knots of the MQAR solving runs (half-lives $1.5{\times}10^5$–$6.4{\times}10^5$ tokens; budget-semantics paragraph above, summarized in Section 4.4) therefore have no counterpart in this family: at LM scale the pinned-mass floor pulls its slow coordinate outward but stays far short of that regime (the guarantee lies in the mass, not the placement; compare the occupation record of Table 3), and the floor's win is *not* attributable to a more frozen store than the hull's. This record is descriptive and changes no pre-specified outcome (source `evals/h2_rflong_knotdrift.json`).

LAMBADA stratum anatomy. Per-token paired NLL on `lambada_test` (five seeds, zero training; octiles of the control surprisal stream with the complementary arm-side bucketing recorded per seed; subset domain = the standard evaluation protocol, identity gates $\leq 7.2{\times}10^{-5}$ relative; full-corpus read within 10^{-2} of the subset pattern). The -19.70 LAMBADA result decomposes into a strict monotone crossover: the combined arm *pays* $+0.153$ nats on the low-surprisal band (octiles 0–2) and *gains* -0.393 nats on the high-surprisal tail (octiles 6–7), whose top two octiles carry $\approx$80% of the net delta; the anatomy reproduces the result itself (-19.71 vs. recorded -19.70 — an identity, not an approximation). All four winning seeds replicate the crossover, and the inverted seed-3 cell is a *tail collapse* (tail -0.09 against band payment $+0.18$, with octiles 0–5 all positive — the widest payment band of the five seeds), not a copy-surface failure. The same redistribution is the object behind the in-domain copy-surface cost: band-heavy surfaces (WT validation) read it as a penalty, tail-heavy surfaces (LAMBADA last-word prediction) read it as the gain, and PG19 long windows net to parity through stratum mixing (source `evals/h2_lamb_octile.json`).

Standard-unit columns. LAMBADA last-word argmax accuracy under the recorded family stream convention restates the result in accuracy units: span-accuracy 6.66% vs 6.12% (mean $+0.54$ pp, $4/5$ positive; token-level $+0.78$ pp; target-token NLL -0.049 nats, $4/5$) — the gain appears on the last-word surface as the octile anatomy's tail concentration reads it, with the -19.7 ppl result corresponding to a small absolute accuracy shift ($+0.5$ pp): tail strata carry most of the ppl mass but a modest share of argmax wins. Under the *standard full-shard* PG19 sliding protocol (full 10.57M-token shard, stride $W/2$, last-stride counted — the capped protocol with its 5M-token / 4-window caps removed), the long axis reads harder than under the capped protocol: 67.10 vs 64.23 at $W{=}16384$ ($+2.88$ mean, $5/5$; $+7.1$ at seed 0, window-flat over $4k/8k/16k$) against the capped-protocol $+0.05/+0.21$ (3/five seeds) — later shard material is harder and amplifies the penalty. Family comparability stays on the capped protocol (all family results are priced on it). The full-shard column of Table 2b, computed under its pre-specified rules (`evals/h2_rflonghyb_std.json`, one evaluation), keeps the routed hull's penalty smallest there too: $+2.88$ hull vs. $+3.35$ floor and $+5.42$ hybrid mean penalties on the full shard at $W{=}16384$ ($n{=}5$; per-seed pairs in the result file). The span-accuracy and full-shard PG19 columns come from `evals/h2_stdunit.json`.

Convex gate-form probe (A1B). Does the convex form, whose per-token gap to the geometric form is the swing bounded in Lemma 3, help when trained? We replace the geometric (log-of-mean) gate by the convex arithmetic-in-decay form $g = \log \sum_c w_c\, 2^{-1/\tau_c}$ (GLA-side code path, `gate_form=arithmetic_decay`, same router parameters, identical recipe and evaluator, seed 0; parameter count 134.853M, same convention). Under the same three-outcome rule, a single evaluation returns WT-103 124.211, hence $s = -0.166$ — a1b-insufficient: the convex form recovers nothing on the WT-103 axis. On WT-103 it nonetheless undercuts all four GDN-family insufficient rows (124.2 vs. their 126–129 band; the init-corrected GDNCAP re-run supplies the 126 end), yet its held-out long axes are not uniformly paid down: LAMBADA 282.27 is the *lowest* of the table's below-threshold rows, while PG19 51.95/49.91/47.80 sits at the weak end (only the RETNET pair and the quantized snapshot row sit higher) — the geometric gate's concavity is therefore not a forgone free channel at LM scale: it obtains the gain at a measurable cost the convex form pays on the PG19 side of these cells (still seed 0 only). At seed 1 the arm is again insufficient: 126.316,

$s = -0.48$ — both seeds fail the 0.25 threshold, and the second draw's long axes split further (LAMBADA worsens to 303.43 while PG19 improves mildly to $50.92/49.49/46.64$). Its outcome is anchor-sensitive (partial at the seed-1 anchor, carries at the seed-2 anchor; Appendix G); the pre-specified anchor convention stands. *Implementation note*: the two evaluated cells ran on the GLA-side MHL lineage (K=8 default knots, G=1, `init_long=0`, the delta-side head width ineffective) and were scored against the delta-side anchored threshold — the geometric-vs-arithmetic read is skeleton-confounded and stands as this arm's measured record only. *Same-backbone arithmetic arm* (ARITHDN). The same gate form on the MHLG skeleton (`gate_form=arithmetic_decay` on MHLDN, K=4; tested pointwise against Jensen's inequality) at seed 0 returns WT-103 119.298, rule $s = +0.556$ — PARTIAL (pre-specified three-outcome): the convex form recovers over half the gain on the delta-side skeleton, so the swing channel is not forgone and form alone does not produce the full gain; the geometric-vs-arithmetic read stays operating-point-level. Descriptive (no thresholds): LAMBADA 266.82, PG19 $49.33/47.49/44.08$. The additional seeds give the *spectrum*: $s = +0.556, +1.239, +0.890$ — $3/3$ positive, $2/3$ above the threshold, mean $+0.895$ — the convex form carries a seed-spectrum gain (positive everywhere, majority strong, one partial cell); the geometric-vs-arithmetic question is therefore a measured spectrum contrast, neither insufficiency nor subsumption. Descriptive: LAMBADA $266.82/273.62/265.61$, PG19@16k $44.08/42.99/43.77$. Anchor-seed status: positive at every re-anchor; the seed-0 read's band label itself moves (partial $+0.556 \rightarrow$ carries at the seed-1 anchor, App. G), so the spectrum contrast stands as positive-everywhere with anchor-relative class boundaries.

Functional new-domain cell (MSRS, mixed-gap retrieval). The mixed-scale recall suite emits 16 key–value pairs per sequence at gaps sampled from $\{32, 128, 512\}$ (weights $\{0.5, 0.3, 0.2\}$, fixed random seed, T=1024, 40 batches, 20,480 supervised responses per arm, `evals/h2_re1_msrs_probe_c.json`); an attention anchor must read near−far $\geq +0.10$ before any verdict is admissible (null-check gate, pre-specified; obtained $0.177 - 0.037 = +0.139$). The pre-specified verdict map: far-bucket $\{512\}$ paired bootstrap (10^4) CI of combined−control above zero with the near bucket undegraded ⇒ msrs-carries (a first functional-domain positive), below zero ⇒ msrs-losing, otherwise msrs-parity. Result: at the far bucket *all four* state-space arms read exactly $0/4049$ — paired CI $[0.0, 0.0]$ on both seed pairs — while the near-bucket read is split mildly against the combined arm (combined−control $+0.27$ pt on s_0 with CI $[-0.09, +0.63]$ vs. -0.96 pt on s_1 with CI $[-1.33, -0.60]$): msrs-parity, same-sign $2/2$ at `evals/h2_re1_msrs_agg.json`. Two caveats accompany the verdict. (i) *Floor-pin*: the far bucket pins every non-attention arm at $0/4049$ and the anchor itself degrades to 3.7% there — the parity cell is an instrument-sensitivity-limited parity at far, not a clean tie at working signal. The pre-specified resolution routes were training-time exposure or a coarser value space; the zero-training branch was executed: a coarse-candidate reading (256-way candidate sets on the bitwise-identical task stream, within-set true-value rank; pre-specified anchor gates and blind threshold, `evals/h2_msrs_coarse_agg.json`) passes its instrument checks (attention degrades $+1.94$ bits near-to-far, stays resolvable at far with median within-256 rank 38 and acc 12.6% against 0.39% chance, and replicates the near-bucket sign split) yet yields *blind-at-far-coarse* on both seed pairs — far-bucket median within-256 ranks 107–113, at or above the pre-specified 0.8×128.5 chance band — so neither state-space arm holds far-bucket comparative order even at 256-way. The training-time alternative was therefore not run at this scale. (ii) *No support for routed bandwidth*: since the combined arm does not exceed the fixed-rate control on the far bucket, the hypothesis that multi-horizon bandwidth must be token-routed receives no support at this instrument (operating-point level) — the LM results license mechanism attribution with a readable gate, and this single instrument does not settle the functional new-domain question; the rank-axis reading below adds a further scoring axis. The consequence chain pre-specified under msrs-carries (train-time mixed-gap upgrade) correspondingly is not triggered. The rank-axis second reading (same task stream, true-value rank scoring, pre-specified thresholds and floor gate, `evals/h2_re1_msrs_rank_agg.json`) yields blind-at-far on both seed pairs: the far-bucket median ranks sit at 8054/8609 and 8091/8101 — at or above the chance-rank band — where the anchor degrades monotonically but resolvably (median ranks $99/327/1531$ across gaps, $+3.35$ bits near-to-far). The @1 floor is therefore not a threshold artifact but the arms' state at this operating point, and with the coarse-candidate third reading of caveat (i) this null result holds on three scoring axes of the identical task stream: at this scale, no state-space arm shows a far-bucket advantage on any of them.

Value-grouped variant (G4). Does the MQAR value-grouping gain carry over to language modeling? The combined arm with value-grouped routing (MHLDN with `gate_groups=4`; the pre-specified pseudo-head expansion, parameter count 134.942M) returns, under the same rule, WT-103 116.966, hence $s = +0.899$ — **G4-LM-BRIDGES**: the MQAR g=128 value-grouping gain ($0.719 \rightarrow 0.837$) carries to the LM WT-103 axis — the grouped variant stands as a strong positive cell against the 123.080 anchor, at a $+0.69$-ppl same-seed trade against the G1 combined arm (116.281). The descriptive trade is LAMBADA-real and PG19-mixed: LAMBADA 298.51 (vs. 262.62 G1), while PG19 reads $48.66/47.87/45.40$ against G1's $49.08/47.57/44.13$ (ahead $@4k$ by 0.42, behind $@8k/16k$ by $0.31/1.27$). We record G4 as an axis-trade variant — a standing WT-103 gain with a LAMBADA penalty — listed on the consolidated grid; single seed, no upgrade claim beyond the measured cells.

Wall-clock-matched control. The gate costs $\approx 17.2\%$ extra step time ($92.8k$ vs. $108.8k$ tok/s, a $\sim 15\%$ throughput reduction; this is the GLA-side last-1000-step mean, while the cost comparison's $1.125/1.065/1.086$ ratios are the delta-side n=5 family under the $\geq$2,000-step log-median convention (control readings 108.8k vs. 104.4k — same wall-clock quantity, different family and statistic; neither is a serving benchmark, and the tail-mean is the closer deployed-latency correlate)); the cleanest matched-time comparison re-trains the anchor control for 18400 steps ($+15\%$ — an approximate match: exact time parity would grant the control $\approx 17\%$ more steps, so this cell leaves the control $\sim 2\%$ of budget short, as noted) byte-identical otherwise (parameter count 133.835M measured from the trained checkpoint). Under the pre-specified three readings ($\geq$ 122.580 wallclock-null / $<$ 122.080 wallclock-report / band border), a single evaluation returns WT-103 123.359 — wallclock-null: 15% more steps do not improve the control (-0.28 ppl), the combined-arm comparison is conservative as run. Descriptive axes (no thresholds, and outside the rule axis): the longer control improves LAMBADA markedly ($293.49 \rightarrow 257.00$, above the combined arm's single-seed 262.62) and all three PG19 windows ($45.69/43.66/41.34$) — a genuine training-length account that does not touch the anchored-recipe results.

Value-grouped variant, second seed. At seed 1 the same recipe returns WT-103 119.465, hence $s = +0.532$ — **G4-BORDER** against the pre-specified $\{0.75, 0.25\}$ thresholds. Family read: the bridge is positive at both seeds ($+0.90/+0.53$) and magnitude-fragile, not all-or-nothing like the grid-size result, whose best K reverses between seeds (Appendix Q.12); descriptive axes at s1 (LAMBADA 287.81, PG19 $48.02/45.82/42.48$) show the same axis trade with a lighter toll.

Value-grouped variant, third seed. A third seed, with the rule fixed before training: the same-recipe s2 cell returns WT-103 112.456, hence $s = +1.563 \geq 0.75$ — the three-seed rule reads positive at all three seeds: the value-grouping bridge holds at $+0.90/+0.53/+1.56$ with two of three above the threshold, an amplitude spectrum ($\sim 2.9\times$ spread) rather than a reversal like that of the grid-size result. The third cell's descriptive axes are family standouts: LAMBADA 245.60 — the best we have measured across the same-recipe state-space arms (descriptive) — and PG19 $48.72/46.84/43.51$ (in-band at 16k; cross-seed spectrum $45.40/42.48/43.51$), against the s0/s1 penalty pattern.

Same-recipe full-attention reference. A causal multi-head self-attention arm under the byte-identical recipe (123.551M parameters, 7.8% fewer; rotary position embeddings, causal PyTorch scaled-dot-product attention (SDPA), everything else unchanged) records on seed 0: WT-103 103.599, LAMBADA 226.955, PG19 $143.784/337.585/448.643$ ($w4096/8192/16384$) — ahead of both family arms on the short-window axes (WT-103 -12.68 vs. the combined arm, LAMBADA -35.7) and collapsed on the extrapolation windows (every arm trains at context 2048; PG19 windows are outside that length class for all; the anchor is recorded axis-wise with no composite label). The short-window gap to the anchor is an open cost at this budget; the anchor's long-window collapse on the same recipe externalizes what the family's measured PG19 results protect against. The 12.68-ppl gap's own anatomy (`evals/h2_lm_anchor_octile.json`, full-corpus teacher-forced streams, anchor-surprisal octile buckets, single s_0 pair, descriptive): paired MHLDN $-$ anchor nats read octile-monotone $\{+0.334, +0.329, +0.265, +0.206, +0.105, -0.017, -0.099, -0.153\}$ — 95.7% of the net gap lives on the three copy-surface octiles ($+0.309$ band mean), while on the three hardest octiles the combined arm *leads* the anchor ($-0.017/-0.099/-0.153$). The anchor's advantage is exactly the copy-surface stratum; the transfer reading and this anatomy share

one sign structure (the full-corpus stream perplexities differ from the head-window numbers because the two evaluations cover different positions, but the paired difference agrees: 0.121 nats vs. the head-window log-ratio 0.115).

F FULL-TEST-SET PAIRED EVALUATION

Convention: the full test set of Section 4 is scored from stored per-token streams over the full corpus (273,882 positions, chunked 2048, bf16 autocast, float32 per-position NLL; ppl $=$ $\exp$(mean per-position nats)); instrument identity re-derived per seed from the same banks (the nominal-32 head-window re-computed on-bank differs by <0.05 ppl for seeds $0/1/3/4$; seed 2 verified to 4×10^{-4} in the anatomy probe). The combined arm's seed $\{0, 1, 3, 4\}$ banks are a forward-only pass over its frozen checkpoints (one forward pass); the controls' banks are the full-corpus and seed-extension archives (Appendix P). Aggregator thresholds (CI below/above zero) were fixed in code before the job was run. ***Result (five seeds, paired per-seed).*** Versus the capacity control: deltas $\{-5.524, -13.052, -12.781, -10.487, -17.390\}$, mean -11.847, bootstrap 95% CI $[-15.088, -8.427]$, $5/5$ negative — fullcorp-confirmed. Versus the GLA-side mixture: $\{-5.805, -8.760, -5.692, -10.376, -12.465\}$, mean -8.620, CI $[-10.888, -6.351]$, $5/5$ negative — fullcorp-confirmed. Three readings. (i) The full-corpus result is at least as strong as the head-window result (-11.85 vs. -11.46): the LM-scale write-side advantage is not head-concentrated. (ii) The combined arm flips no seed: head-window and full-corpus agree in sign in all five (contrast the GLA arm's own archive reconciliation, which carries one recorded flipping seed, $+0.280$). (iii) The seed-2 stream behind the two-stratum probes is representative: its deltas (-12.78 control, -5.69 GLA-side) sit inside the five-seed spread (-5.5 to -17.4) — the per-seed answer to the representativeness question. Source: `evals/mhldn_lm_fullcorpus5.json`.

Exact tests against the bootstrap calibration. Bootstrap CIs at n=5 risk optimistic narrowness; we re-read every CI-closed result distribution-free — the two-sided sign test and the full-enumeration sign-flip permutation on the declared paired deltas (2^n flips, p = share with $|\bar{d}| \geq |\bar{d}_{\text{obs}}|$; no new measurement, a single zero-compute read). WT-103 five-seed result: $5/5$ negative, sign $=$ perm $=$ 0.0625 (the minimum at n=5); WT under the pre-specified seven-seed rule: $6/6$, sign $=$ perm $=$ 0.03125, the conventional threshold reached exactly. LAMBADA seven-seed: $6/6$, $0.03125/0.03125$. The GLA-side five-seed result: $5/5$, $0.0625/0.0625$. The $+1.89$ validation penalty reads $4/5$ positive (sign 0.3750, perm 0.1875) — direction-consistent but not a rejection at 5%, so the paper reports a small cost here rather than a third result. The exact tests change no result: they support the strong results and calibrate the weak one.

G SENSITIVITY TO THE TRAINING RECIPE

Could the combined arm's result be recipe luck — the capacity control merely undertuned at the reference learning rate ? A three-point learning-rate envelope ($\{10^{-4}$, 3×10^{-4} anchor, $10^{-3}\}$) is trained for *both* arms at seed 0 under the byte-identical LM recipe (Table 9; stamp-checked cells, the same literal parameter assertions 134.057M/133.835M), rules fixed before any training ran: recipe-robust iff the tuned-minimum separator $\min(\text{GDN_CM}) - \min(\text{MHLDN}) \geq +6.0$ (half of the claimed result at the anchor recipe); recipe-partial inside $(0, 6)$; fragile at ≤ 0.

Table 9: Per-arm learning-rate envelope, seed 0: WikiText-103 head-window test perplexity at each learning rate under the recipe of Table 1. Bold: each arm's tuned minimum. The 2×10^{-3} column is the pre-specified high-side bracket probe.

	Learning rate			
Arm (seed 0)	10^{-4}	3×10^{-4}	10^{-3}	2×10^{-3}
MHLDN (MHLG)	205.243	116.281	**107.560**	108.829
GDN_CM (GDN control)	218.318	123.080	**113.533**	113.624
Δ (MHLDN $-$ GDN_CM)	-13.08	-6.80	-5.97	-4.80

The rule resolves recipe-partial: the tuned-minimum separator reads $+5.973$ (both arms' minima sit at 10^{-3}), 87.9% of the anchor seed-0 gain and 52% of the five-seed headline gain — the separator never drops to zero at any matched recipe (worst matched-rate difference -5.97), and tuning moves the control slightly farther than the combined arm (-9.55 vs. -8.72 from the anchor). The result is not a recipe-luck artifact at this envelope; the envelope covers one seed and three rates, so this is a bound, not a proof. Descriptively at the tuned pair (10^{-3}, outside the rule): LAMBADA -26.2 persists; PG19 stays a mild positive on this seed ($+1.2$ tuned, $+2.2$ anchor), inside the seed-side spread already recorded; the validation-minimum penalty reproduces ($+2.51$ tuned vs. $+2.88$ anchor, seed 0). The pre-specified fourth point (2×10^{-3}, both arms) degrades both sides of every minimum ($+1.27$ / $+0.09$): the peak is bracketed at 10^{-3} on both arms, the four-rate reread leaves the rule unchanged (min-vs-min $+5.973$), and the worst matched-rate difference stays negative throughout (-4.80 at the bracket). Source: `evals/h2_lm_lrtune_agg.json` and `evals/h2_lm_lrtune_bracket_agg.json`. A second seed at the bracket-peak rate (10^{-3}) preserves the envelope: separator $+10.075$ (GDNCM 119.010 vs. MHLDN 108.935, seed 1; seed-0 point $+5.973$), n=2 with both seeds raw (`evals/h2_lm_lrtune_s1_agg.json`).

With the tuned control cell of the third seed, at the peaked rate, per-seed separators run $+5.973/+10.075/+3.904$ (GDNCM@1e−3 112.922 vs. MHLDN@1e-3 109.018, seed 2), mean $+6.651$, pooled paired bootstrap CI_{95} $[+3.904, +10.075]$ entirely above zero — at n=3 the interval is the seed range itself, recorded as the bound — rule outcome tuned-n3-below-zero, i.e., MHLG minus control below zero at every seed (`evals/h2_lm_req2_tunedn3_agg.json`): the tuned-recipe separator is positive on every seed, at 97.8% of the anchor reference gain and 58% of the five-seed head-window gain. Descriptively (no rule): the tuned LAMBADA gain persists at seed 2 (236.787 vs. 245.061, $+8.274$) while PG19@16k crosses to the method side at $+0.233$ — inside the recorded cross-seed spread ($-1.106/-0.213/+0.233$, tightening across seeds).

Re-based recovery reads (descriptive). A stress test asks which headline conclusions survive if every anchor is re-based at the per-arm tuned minimum. Algebra first: re-basing the recovery statistic from the anchor control (123.080) to its tuned minimum (113.533 @ 10^{-3}) shifts every s-cell by the same constant $-9.547/6.799 = -1.404$ and changes no ordering. Under that re-based read every negative control stays at or below the 0.25 threshold (GDNRM -2.21, GDNCAP -2.07, GDNBSN -2.14, GDNLAD -2.37, A1B $-1.57/-1.88$, wall-clock -1.45, RETNETM $-2.75/-3.03$), and every positive probe cell crosses down out of its band (STATICH -0.84; G4 $-0.51/-0.87$; RETNET s1 -0.18; MHLDNH $-0.26/+0.02$). We record the flips verbatim, and the caveat that defines their meaning: the probe arms were trained at the *anchored* recipe and never given the tuned one, so a tuned-base read compares unequal recipes — the cell verdicts are pre-specified at the anchor convention and do not move. What answers the question at the tuned recipe is the directly *trained* tuned pair: the separator is $+5.973$ at seed 0 and $+10.075$ at seed 1, i.e. 87.9%/148% of the anchor seed-0 gain, with the LAMBADA gain and the validation penalty persisting. The main results do not use the anchor at all: the five-seed paired head-window and full-test-set differences (-11.46 and -11.85, both with CIs below zero) and the other paired seed comparisons are computed within one recipe and are unaffected by any re-basing.

Anchor-seed sensitivity (zero-compute re-read). The recovery statistic's anchor is the seed-0 control cell (123.080); re-anchoring at the seed-1 / seed-2 control cells ($128.548/132.468$) shifts every s-cell by a constant $+0.804/+1.381$ — the within-family seed spread of the reference arm itself. We list every change of outcome (`evals/h2_v820r1_q2q5.json`): RETNET ($+0.096\rightarrow+0.900/+1.476$) and A1B ($-0.166\rightarrow+0.638/+1.215$) cross bands at *both* alternate anchors — their anchored insufficiency outcomes are anchor-sensitive; GDNCAP v2 (the re-trained cell with corrected initialization) crosses already at seed-1 ($+0.253$); the core negative controls (GDNRM, GDNBSN, GDNLAD, GDNCAP v1 (the original cell), RETNETM) stay at or below $+0.138$ at the seed-1 anchor and crosses only at seed-2 ($+0.036$–$+0.715$); every strong positive cell stays positive under both re-anchors. What this means operationally: single-anchor score thresholds carry ±0.8–1.4 units of anchor-seed uncertainty, the family results rest on multi-seed paired CIs throughout rather than on the anchor, and we report the two anchor-sensitive outcomes (RETNET, A1B) alongside the anchor. Re-anchoring at the five-seed control *mean* (128.450; constant shift $+0.790$ of every cell) reproduces the seed-1 pattern of outcome changes with one boundary-level exception — GDNCAP v2 stays at $+0.239$, just below the partial threshold: the two named anchor-sensitive outcomes cross as recorded, and on this anchor STATICH and the same-backbone arithmetic arm

(ARITHDN) cross into the carries class as well, so the anchor-sensitive arms we name are four (RETNET, A1B, STATICH, ARITHDN); every strong positive cell and every core negative-control cell stays in band (`evals/h2_v824r1_meananchor.json`).

Equal-recipe tuned probe. The key positive probe arm is also trained *at* the tuned recipe, to test which readings hold there. The hard-commitment arm at the peaked rate (`mhldnh`, lr 10^{-3}, seed 0, byte-identical otherwise) returns WT-103 108.730, hence $\text{sep}_h = 113.533 - 108.730 = +4.803$ — below the pre-specified half-anchor threshold 6.0, at 80.4% of the soft tuned gain ($+5.973$): **PARTIAL**, as pre-specified, and the unequal-recipe re-base complication is gone (both sides tuned). Descriptively, its long axes lead the whole ladder: LAMBADA 233.79 (the best measured on the hard-assignment side, 15.2 below that side's previous best) and PG19 $43.79/42.01/39.15$ ($16k$ the best measured on that side; the soft tuned comparator reads lower on both axes, and inside the family band even the control sits at 43.45), while the short-axis gap to the soft tuned arm is $+1.17$ ppl (108.73 vs. 107.56). The three-level reading we record for the sensitivity question: anchor-convention rules ($s = +1.14/+1.43/+0.60$ at n=3, seed-fragile) $\succ$ equal-recipe tuned separator ($+4.80$, 80%, the tuned partial) $\succ$ unequal-recipe re-base (algebraic only, flagged with its caveat) — none of which moves the CI-closed main results. At the fully tuned grid (both forms at the peaked rate against per-seed tuned controls; `evals/h2_v824r1_meananchor.json`) the control-relative separator is available at n=3: $\{+4.803, +9.185, +1.331\}$ for the hard form against $\{+5.973, +10.075, +3.904\}$ for the soft form — the trained commitment keeps three quarters of the tuned gain on average (76.8%; per-seed $80\%/91\%/34\%$, distributed seed-fragile), while the soft form stays ahead on all three head-to-head cells ($+1.170/+0.890/+2.573$, in line with soft-lr-peak-holds), and both forms' weakness co-locates at seed 2.

Matched-seed tuned grid. Does the hard arm's tuned result hold across seeds at the peaked recipe? The hard arm was trained at seeds 1 and 2 (byte-identical recipe, lr 10^{-3}): hard seed 1 $109.825/241.17/38.325$, hard seed 2 $111.591/233.47/41.436$ (WT/LAMBADA/PG19@16k). The soft comparators are the tuned soft cells of the same seeds (seed 2: $109.018/236.79/38.266$). The matched nine-cell comparison (hard $-$ soft, $+$ = soft ahead) reads: seed 0 $+1.170/+7.640/+0.328$ (soft ahead on all axes); seed 1 $+0.890/-8.668/-1.035$ (split: WT soft-led, LAMBADA/PG19 hard-led); seed 2 $+2.573/-3.32/+3.170$ (WT and PG19 soft-led, LAMBADA hard-led). The hard arm leads in $3/9$ cells (LAMBADA at seeds 1 and 2, PG19 at seed 1), below the pre-specified 4–5 borderline band, so the rule reads soft-lr-peak-holds: at the peaked learning rate the soft form stays ahead, and on WT-103 it leads at every seed. The seed-2 soft cell at the anchor recipe ($118.986/267.04/42.696$) is much weaker than its tuned counterpart; scored against that untuned cell, seed 2 would read $-7.395/-33.571/-1.260$ and the hard-advantage count $5/9$ (with $2/6$ over the matched seeds 0–1), so the seed-2 soft collapse (118.986) belongs to the untuned recipe, not to the seed. The tuned partial stands as the single-seed separator statement it was pre-specified as, and nothing here moves the CI-closed main results or the family reading.

H SECOND-CORPUS ROBUSTNESS UNDER THE ANCHORED RECIPE (OPENWEBTEXT)

Same recipe, no per-corpus tuning (per-corpus tuning is answered by the rate envelope, App. G); Table 10 gives the seed-0 results. The corpus-2 banks are OpenWebText under the GPT-2 tokenizer (470M/40M train/val uint16; aggregation `evals/h2_lm_corpus2_agg.py`).

The n=2 extension sharpens the transfer to an axis-typed one (seed-1 pair, byte-identical recipe; one evaluation under the pre-specified rule, with a sign correction to the rule recorded before evaluation): on the MHLG-minus-control convention of Table 10, WT-103 $\Delta = -9.99$ persists in direction but dilutes (41% of the seed-0 value, below the $1.5\times$ threshold by 0.21), LAMBADA strengthens to $\Delta = -44.59$ (vs. -36.53), while both PG19 long windows flip control-side at seed 1 ($\Delta = +19.19/+26.07/+20.63$ at $4k/8k/16k$; corpus2-dilute-n2, `evals/h2_lm_corpus2_n2.json`). The cross-corpus transfer at n=2 is thus axis-typed and seed-heterogeneous: the WT/LAMBADA gains persist (diluting/widening), the PG19 gain does not replicate, and the -24.15 row of Table 10 stands as the seed 0 end of that spectrum rather than a uniform carry.

Table 10: Anchored-recipe second corpus (OpenWebText): MHLDN − GDN_CM separators, seed 0. Same held-out evaluators as the family results; the OpenWebText (OWT) validation-minimum row is descriptive (copy-shaped val penalty replicates cross-corpus). Rule: corpus2-carries at separator $\geq +3.0$ (anchor-corpus point $+6.799$).

Evaluation	GDN_CM	MHLDN	Δ (MHLDN − GDN_CM)
WT-103 test	178.319	154.168	−24.15
LAMBADA	286.433	249.903	−36.53
PG19 4k	174.904	161.285	−13.62
PG19 16k	155.229	138.971	−16.26
OWT validation minimum (descriptive)	43.70	45.29	+1.59

I INITIALIZATION-MATCHED CONTINUOUS-RATE CONTROLS

To separate the multi-rate initialization *prior* from the grid *constraint/router* channel, the capacity control is re-trained with its native continuous gate initialized on the very ladder the mixture grid starts from: $A_{\log} \to 0$ and per-value-head $\mathrm{dt_bias} \leftarrow \mathrm{softplus}^{-1}(\ln 2/\tau_{h \bmod 4})$ with $\tau \in \{16, 64, 256, 1024\}$ — parameter count and every other byte of the recipe unchanged (133.835M parameters, counts checked; three seeds; decision rules fixed before training). *Verdict: the constraint carries, not the prior.* WT-103 (nominal-32, per-seed): the init-matched control reads $\{128.05, 129.44, 131.46\}$ — on the result that decides the attribution it is worse than MHLDN by $+12.10$ ppl (paired 95% CI $[+11.77, +12.47]$, entirely above zero) and indistinguishable from the capacity control ($+1.62$, CI $[-1.01, +4.97]$, covers zero; per-seed $+5.0/+0.9/-1.0$). The structured ladder prior alone therefore recovers *none* of the combined-arm gain; the difference is attributable in whole to the constraint/router channel itself. Descriptive secondary axes agree in kind, not in sign: on LAMBADA the init-matched control *improves* over the capacity control (-13.05, CI $[-18.93, -5.34]$) while losing to MHLDN ($+14.85$ CI $[+6.46, +22.07]$), and PG19 reads parity in both windows (all CIs span zero). Source: `evals/h2_ladder_init_agg.json`. *Zero-compute complement (direct read of the trained native gate).* The unconstrained continuous gate does not lack for latent rate diversity: per (layer, value-head) realized half-lives over all five GDN_CM checkpoints span $[0.6, 10145]$ tokens with per-seed medians 8.8–15.9, p75 26.6–49.2, p95 179.0–259.3; only 4.2–5.2% of heads sit above 256 and 0–3.1% above 1024, while 50–61.5% sit below 16. The mixture's trained allocation instead puts 0.236–0.562 of mass on its long knot per layer on the GLA-side grid, 0.330–0.848 on MHLG (App. L). So the continuous gate optimizes itself *toward shorter* timescales than the routed mixture *allocates*; the ladder prior delivers the span at initialization, yet even then the gain does not follow (above). The causal source of the readable gate's gain is the per-token constrained routing over a shared, maintained budget — not the prior, not latent diversity. Source: `evals/gdncm_nativegate_hl_dist.json`.

The optimization-side account: why the constraint helps. The ladder-init control above removes the *prior* explanation; what remains is the constraint/router acting on the optimization process itself. The two half-life distributions put the two arms on opposite sides of the sharp divide the instrument study already mapped: after training, the free continuous gate's channels collapse onto short memory (median half-life 8.8–15.9 tokens across layers, 50–61.5% of channels below 16, only 4.2–5.2% past 256) — short-rate attractors dominate under next-token reward, because fast features pay immediately while slow features pay rarely and late at this budget. Under full activation conditioning the tail thickens sixfold while the median stays short (`evals/h2_gradattractor_pop.json`; gdncm, 3 seeds, 64 fixed 2048-token windows, token-population shares): 45.2% of (token, head) pairs past 16, 28.5% past 256 — head-bimodal (half the heads ≈0% long; the decile head ≈93.6% long; concentrated at L0/L7/L10/L11), 17.0% past 1024, with token median ≈10. The per-head summary above (4.2–5.2% of heads past 256; `evals/gdncm_nativegate_hl_dist.json`) assigns one half-life to each (layer, head), with no token dimension, and its values are consistent with the gate's input-independent (bias-only) rest rate rather than with token-level rates; the supplementary material contains this result file but not the script that computed it. The attractor's *median dominance persists*; its *tail-rarity does not* — a spontaneous slow-head minority exists even in the free control, refining the mechanism: the gain's separation cannot be slow occupation alone but the occupation's maintenance guarantee *jointly* with the grid hull, consis-

tently with the $n{=}5$ floor result. The naive rare-and-noisy gloss also fails at the converged point (`evals/h2_gradattractor.json`): the gradient signal-to-noise ratio (SNR) on the rate scale *rises* with realized longevity (quartiles $0.33{\to}1.35$, monotone; corr $+0.247$), signal concentration does not rise, and two thirds of heads are pushed *longer* — the attractor persists through gradient incoherence at short scales, not an active shortening drag; the training-time collapse dynamic stays with the STATICH drift record. *Budget-invariance* of the equilibrium is measured at a $3.8\times$ span (`evals/h2_gradattractor_bq.json`; the $\sim$1B pilot finals, same windows): token-population shares 262M$\to$1B read >16 $0.452{\to}0.463$, >256 $0.285{\to}0.285$ (identical), >1024 $0.170{\to}0.196$, $>10^5$ $0.011{\to}0.021$, token median $10.2{\to}11.0$, and the SNR-rises-with-longevity pattern persists ($0.44{\to}1.65$ quartiles) — the attractor/heavy-tail equilibrium is budget-invariant at this span, not an artifact of the 262M budget. The token-routed grid resists the attractor structurally: every channel's effective rate stays bounded between the trained knots, and the router keeps a persistent mass share on the long knot (basin mass 0.236–0.562 across layers and seeds; allocations concentrate near-knot under near-one-hot routing rather than quantizing each channel to one rate — geometric mixing produces continuous in-between rates) — slow-rate occupation is held open by construction and populated by routing. That maintained slow-rate share is the load-bearing slice of the architecture (all-layer DROPLONG costs $+26.3$–$+29.4$ ppl GLA-side and $+39.7$–$+45.1$ ppl MHLG-side, deep-layer localized on both; 91–94% GLA-side and 94.5–96.4% MHLG-side of ablation damage is first-order rate damage; App. M), it acts at non-query positions (the clamp replay collapses 15.46pt on the collapsed instance when applied whole-net, while a query-only clamp moves accuracy ≤ 0.04pt), and the routing that feeds it is itself load-bearing (frozen-router ablation -0.16 accuracy). The chain from the distribution shift to the octile-7 tail is therefore causal at every measured link: maintained slow state $\to$ long-range context available on tail tokens $\to$ the $-0.565/-0.426$ novel-tail gain that persists at 75% amplitude across write sides. The single link we have *not* intervened on — a training-time (rather than eval-time) manipulation of the long-knot share producing a corresponding training-time tail change — is recorded here as the remaining open step of the causal ladder. Its first instantiation, at the one learnable gap, is reported here, and it splits (trainlong-mixed; pre-specified rule, two pilots, one read): the mask at every training step moves the pair $\{0.995, 0.459\}$ to $\{0.763, 0.831\}$, and the pre-specified third seed yields the third instance $0.702{\to}0.356$: two collapse carrier-side (-0.23, -0.35) while the floor instance gains $+0.37$ (top retrieval bins move with the same signs); mean -0.07, spread $0.54 \to 0.48$ (over the two pilots alone, the spread is 0.07). Training-time mediation of the tail is thus *instance-dependent* at this scale — the long-knot share is load-bearing on two instances yet destabilizing on the weakest, implicating free long-knot routing in the per-instance variance without removing it. Aggregates `evals/h2_mqar_trainlong_agg.json` and `evals/h2_mqar_trainlong_agg_s2.json`, each single-use. Holding the mask off until 75% of training splits from both limit rules (ct-spectrum; pre-specified gate, three seeds, one read): the pair $\{0.995, 0.459\}$ and the fourth-seed instance 0.588 move to $\{0.113, 0.459, 0.588\}$ — the rescue that every-step masking confers on the floor instance vanishes without early exposure ($+0.372{\to}+0.000$), while the top instance is damaged nearly four times harder than by the every-step mask (-0.882 vs. -0.232, its long-range bins flatlined). The timing structure is acquisition-ordered descriptively: the instance whose supervised formation window overlaps the masked stretch (seed 0, forming over steps 9000–12000) collapses far past its every-step damage, while both instances already acquiring before the gate (seeds 1 and 4) move by ≤ 0.001. Training-time mediation is thus timing-dependent along the formation schedule, not resolved by either limit rule. A fourth instance (seed 2) completes the decomposition: its every-step collapse ($0.702{\to}0.356$) vanishes exactly under the late gate ($0.7024{\to}0.7024$), so every effect carried by an early-acquiring instance — the seed-1 rescue and the seed-2 collapse alike, their supervised trajectories crossing 20% accuracy at steps 8750 and 6750 — is assembled during the first three quarters of training, and only the instance still forming inside the masked stretch is damaged. Aggregate `evals/h2_trainlong_cadence.json`; the seed-2 cell is a one-read descriptive classification on the same pre-specified gate.

J FIVE-SEED RECALL AT GAP 128 AND GROUPED RATES

An extension (fixed before training) takes $g{=}128$ to seeds $\{3, 4\}$ in both families and adds a grouped-rate variant ($G{=}4$), executed as ten cells, each passing configuration and parameter-count checks ($6.924864/6.949632/7.048320$M). *Semantic axis*: at $G{>}1$ the combined arm's rates group over *value* channels under a shared key-addressed erasure (pseudo-head kernel expansion), whereas

the GLA-side arm groups over key/input channels — the two are not the same subspace, and we do not compare them as such. *(A) Residue at n=5: null.* Combined minus control at g=128, five seeds: $\{+0.058, -0.300, -0.068, -0.148, +0.588\}$, mean $+0.026$, paired bootstrap 95% CI $[-0.198, +0.326]$ (rng(48), 10^4): the within-regime strength residue is seed-dependent in both families, not an arm property. The control's own extension carries the full span (0.9984 vs. 0.0002 — a true full-train self-collapse of seed 4, verified from logs), the combined arm's $\{0.8509, 0.588\}$. *Checkpoint-selection probe at the crossing cadence: mixed.* Checkpoints stored at the training cadence answer the pre-specified 75% runaway-crossing question with zero retraining: at step 9000 (the 75% cadence) both losing seeds already sit at or below their terminal values, with seed 1 fully beneath the winners' 9k band (s_1: $0.249 \to 0.459$; s_4: $0.536 \to 0.588$; winners $\{0.991, 0.647, 0.743\}$), the losing trajectory is already visible at the 50% cadence (s_1: 0.003 where the same-seed control reads 0.511), the null itself is cadence-invariant (paired means: -0.018, CI $[-0.294, +0.289]$ @ 9000 vs. $+0.026$, CI $[-0.198, +0.326]$ @ 12000, the latter reproducing outcome (A) exactly; ten-cell sanity $|\text{err}| \leq 0.0001$; aggregate `evals/h2_g128_ckptsel_probe.json`). The runaway crossing is therefore a *marker* of these seeds, not the event that decides them; whether knot-side regularization *at that cadence* changes the fate remains untested. The training-time axis is addressed, independently of cadence, by the long-knot masking experiment of Appendix I (stored as TRAINLONG): masking the long knot at every training step moves the first two instances $\{0.995, 0.459\}$ to $\{0.763, 0.831\}$ and the third from 0.702 to 0.356 (trainlong-mixed) — the mediation is instance-dependent, implicating free long-knot routing in the per-instance variance without removing it. *(B) Grouping at g=32: g-closure-insensitive.* Per-seed closure of the mixture-to-control gap: G=1 $\{1.032, 0.986, 1.001\}$ (mean 1.006) vs. G=4 $\{1.051, 0.999, 0.992\}$ (mean 1.014); the $0.90\times$-mean threshold never binds — value-space grouping does not perturb the write-side confirmation. *(C) Grouping at g=128: g-residue-moves.* Within-regime mean residue moves $-0.103 \to +0.015$ (shift 0.118 past the 0.05 threshold): G=4 cells $\{0.7538, 0.9515, 0.8051\}$, mean 0.837 ± 0.103 — the grouped read pays at the seed-dependent instance-level gap without costing the closed one (B). One run stopped because of a transient GPU error and was re-run once with an identical configuration, as the pre-specified policy allows; aggregate `evals/h2_recall_g128_agg.json`, read once. *Independent-rate sweep: robust across learning rates.* Is the separator dissolution confined to the 3×10^{-3} anchor? Four further cells (both arms, seed 0, lr $\{10^{-3}, 10^{-2}\}$, byte-identical recipe): the combined arm reads *instance-level* at 10^{-3} (0.121) where the control floors (0.0009), both floored at 10^{-2} — the dissolution holds at an independent rate. Aggregate `evals/h2_mqar_lrsweep_agg.json`.

Status of the recall question. The motivating recall venue stays at the accuracy floor at this scale, and the g=128 dissolution above is the most adverse five-seed answer we have. The one long-context benchmark beyond PG19-window perplexity we evaluate — LAMBADA at the document level — moves with the combined arm (-19.70 ppl, $4/5$ seeds negative; paired CI $[-38.84, +7.32]$ spanning zero on the recorded control-side outlier), a descriptive axis, not a CI-closed result; the MQAR budget configuration (12,000 steps, sequence gap $+64$, num-kv 16, batch 128, learnability threshold > 0.03) and the five-seed learnability bracket $(192, 320]$ on the $d \geq N$ exact-recall law (Arora et al., 2024a) are recorded here. A measurable retrieval advantage at a gap regime, task, or modestly larger scale where the readable budget operates is therefore not claimed here; venues that would close it (a second corpus family, a retrieval-shaped benchmark, or the training-time intervention) were candidate follow-ups — the retrieval-shaped one is executed floor-class-wide in App. K, leaving the two remaining candidates open. Both were not run at this scale, for two reasons: the training-time intervention asks gap mixing to rescue optimization that already floors at every learning rate on the easier unmixed point inside the admissible region (its method-favorable premise is falsified in advance, `evals/h2_mqar_lrsweep_agg.json`), and a second corpus family would re-measure a floor already measured floor-class-wide.

Probe-by-probe sweep numbers. Final train accuracies over gaps $\{32, 128, 512, 1024\}$ (three-seed mean$\pm$s.d. at g=32, seed-0 pilots elsewhere): mixture $\{0.522 \pm 0.060,\ 7\times10^{-4},\ 1\times10^{-4},\ 1\times10^{-4}\}$; GDN_CM $\{0.990 \pm 0.017,\ 9\times10^{-4},\ 1\times10^{-4},\ 1\times10^{-4}\}$.

Level-versus-acquisition decomposition at g=32 (zero-compute). Which axis carries the 0.522-vs-0.990 separation at matched steps? Both, separably from the recorded training curves (250-step cadence). (i) *Acquisition rate*: liftoff arrives $\{6250, 5250, 5250\}$ vs. the control's

Table 11: NIAH descriptive buckets (seed 0; n in binding-level draws $= 6\times300$; acc@1 and median rank per bucket; the $\leq$200 bucket carries no draws). Chance acc@1 2×10^{-5}, chance median rank 25,128. Attention anchor descriptive: acc@1 0.0089 overall.

Distance (tokens)	n	GDNCM acc@1	MHLDN acc@1	GDNCM median rank	MHLDN median rank
201–500	259	0	0.0039	5858	5800
501–900	423	0	0	6681	6515
901–1300	392	0	0	5922	6367
$\geq$ 1301	726	0	0	6958	6998

$\{6250, 3500, 3250\}$ (aligned on one seed, $+1{,}750$ and $+2{,}000$ steps late on the other two), the liftoff-to-half passage costs 500–1,000 steps against 3,250–4,000 on the two mixture cells that cross at all (the third never reaches 0.5 within 12k), and peak-window gains are 4–5$\times$ smaller. (ii) *Endpoint level*: at the matched 12k endpoint the mixture reads $0.448/0.571/0.518$ (final losses $4.21/3.22/3.38$ vs. $0.236/0.0005/0.006$) with late-2k slopes $\{0.010, +0.035, -0.003\}$ — $0/3$ above the 0.05 rise threshold, a structural plateau — and held-skill retention $w_8/w_7 = \{1.57, 1.15, 1.43\}$ ($3/3\geq0.85$) reads acquisition-side dilution, not forgetting. The mixture's lower g=32 accuracy therefore reflects both a slower acquisition rate *and* a lower endpoint level; an acquisition-time-matched read still separates at this budget, and no forgetting channel is implicated either way (sources `evals/h2_g32_curves.json` and `evals/h2_g32_windows.json`).

K Token-ID needle-in-a-haystack (NIAH): the retrieval axis at this scale

Protocol (`evals/h2_niah_eval.py`; aggregation `evals/h2_niah_agg.py`): 2042-token real slices from the family held-out stream, K=6 adjacent [key][value] bindings planted at corpus frequency rank 500–5000, six query slots 2042–2047, N=300 fixed trials RNG(2347) replayed identically across arms, key–query distances $[242, 1983]$ (straddling the measured MQAR floor 256 on both sides; the $\leq$200 bucket is empty by construction). The pre-specified rule asks for the five-seed per-trial paired acc@1 delta of the combined arm against the capacity control, with the synthetic instrument validation (planted $+0.08$ $\rightarrow$ detected, archived at `evals/h2_niah_sweep_synthetic_validate.json`) kept separate from all reads.

Rule verdict: niah-delta-parity. Per-seed paired deltas $\{+0.0006, +0.0006, 0, 0, +0.0006\}$, mean $+0.0003$ (95% bootstrap CI $[+0.0001, +0.0006]$, $3/5$ positive) — far below the $+0.05$ threshold. The task itself is floor-class-wide at this scale: acc@1 levels read 0–0.0006 across every family arm (seed-0 descriptive buckets in Table 11), the same-recipe attention anchor reads 0.0089 descriptively, against chance $1/50257 \approx 2\times10^{-5}$. Median rank, however, reads 5,800–7,000 on both main arms against a chance median of 25,128 — an $\approx$4$\times$ ranking preference without first-hit retrieval. The LM results therefore do not rest on explicit single-hop retrieval at this scale; the boundary is recorded as future scope, with a rank-based (not acc@1-only) endpoint suggested.

L Full-corpus allocation probe: which knots carry mass

Probe: every WT-103 test position teacher-forced through each of the three GLA-side MHL checkpoints (chunking convention identical to the full-test-set evaluation of Appendix F); per layer we record corpus-mean router mass per knot, the learned grid, and effective knot count $K_{\text{eff}} = \exp(\text{entropy})$ from the recorded meters. Seed-mean table; seed standard deviations of the mass shares stay at or below 0.10 on every entry (max 0.093; median 0.028).

Functional margins (seed-consistent, weak): a surprisal-octile trend flat within ±0.02 of the layer mean on octiles 0–6 (largest deviation 0.015), with the top octile losing 6–11% of long-knot mass in layers 1–11 (L00 inverted, r=+0.143); 16k-block stationarity of long-knot mass (head/tail 1.00–1.02), the negative control against allocation-drift accounts of head-region gain.

MHLG allocation budget (MHLDN). The same corpus probe on MHLG's three checkpoints (source `evals/h2_lm_knotmass_dn_agg.json`; Table 13) reads a sharper allocation: mass

Table 12: Allocation budget of the GLA-side grid MHL (seed mean over three checkpoints, full WikiText-103 test set): learned knot half-lives τ_k, mean router weight w_k per knot, and effective number of knots K_{eff}. All four knots are exercised in every layer; the long knot ($\tau_4 \geq 256$) carries 0.236–0.562 of the weight per layer; layer 0 runs near single-knot.

	Learned half-life τ_k (tokens)				Mean router weight w_k				
Layer	k=1	k=2	k=3	k=4	k=1	k=2	k=3	k=4	K_{eff}
0	11.0	43.7	160.4	583	0.029	0.112	0.412	0.447	1.25
1	7.9	27.0	91.3	308	0.478	0.118	0.168	0.236	1.78
2	9.1	34.2	119.0	410	0.238	0.120	0.302	0.339	2.03
3	8.2	30.7	107.8	376	0.287	0.092	0.289	0.331	1.97
4	8.4	32.1	116.2	415	0.220	0.077	0.306	0.396	1.98
5	9.5	36.6	132.1	470	0.152	0.104	0.358	0.386	2.19
6	8.9	34.6	125.8	449	0.177	0.083	0.332	0.408	2.10
7	8.4	33.3	125.4	468	0.165	0.064	0.248	0.523	1.84
8	7.2	27.9	105.0	392	0.222	0.065	0.198	0.515	1.70
9	6.4	25.8	103.0	409	0.233	0.057	0.148	0.562	1.62
10	6.8	26.8	104.3	404	0.267	0.071	0.181	0.482	1.75
11	6.1	22.3	81.1	296	0.458	0.080	0.171	0.292	1.75

concentrates on the fast–slow pole pair (knots exercised above 5% number 2–4 per layer, corpus-minimum share 0.29%), the long knot carries 0.330–0.848 per layer (layer-mean long shares; the GLA-side grid's band was 0.236–0.562), L00 runs 85% single-mode, and the corpus-share effective knot count runs $\exp H(\bar{w}) \in [1.56, 3.40]$ over the 36 (seed, layer) grids (median 2.44; the GLA-side grid's band was $[2.47, 3.77]$). Its functional margins differ in sign structure: long-knot mass *anti-correlates* with per-token surprisal in all twelve layers (seed-mean $r \in [-0.16, -0.10]$; the GLA-side grid was flat within ± 0.02, largest deviation 0.015, outside its L00 inversion), while 16k-block stationarity is intact (head/tail 1.001–1.017), removing allocation drift as a head-region account on this family as well. Grid placement on MHLG keeps the geometric law with a looser tail ($\approx$3.69 median adjacency; see the quantizer paragraph below). A compact per-layer summary of Table 13 appears in the main text (Table 4, Section 4.4).

Table 13: Where the MHLG router puts its weight, per layer (seed mean over three MHLG checkpoints, full WikiText-103 test set): learned knot half-lives $\tau_1 < \cdots < \tau_4$, mean router weight w_k per knot, and effective number of knots K_{eff}. In layers 1–11 the weight concentrates on the fastest and slowest knots; the slowest knot carries 0.33–0.85 of the weight, and layer 0 is nearly single-knot. The GLA-side counterpart is Table 12.

	Learned half-life τ_k (tokens)				Mean router weight w_k				
Layer	k=1	k=2	k=3	k=4	k=1	k=2	k=3	k=4	K_{eff}
0	9.5	34.9	126.8	456	0.005	0.003	0.144	0.848	1.59
1	8.0	26.7	89.7	303	0.452	0.091	0.099	0.358	3.20
2	7.9	26.5	89.8	307	0.489	0.097	0.083	0.331	3.01
3	8.0	27.2	92.4	315	0.437	0.065	0.046	0.452	2.80
4	8.5	29.5	103.2	361	0.226	0.021	0.036	0.717	2.16
5	7.9	28.4	102.8	379	0.275	0.041	0.071	0.614	2.62
6	9.0	36.9	163.3	790	0.163	0.030	0.047	0.760	2.11
7	7.9	30.8	126.5	560	0.254	0.029	0.064	0.653	2.43
8	8.0	30.8	121.9	494	0.312	0.042	0.062	0.584	2.65
9	7.5	30.2	125.3	536	0.295	0.022	0.061	0.622	2.46
10	8.0	32.9	141.0	628	0.224	0.015	0.043	0.719	2.15
11	8.3	32.3	127.1	507	0.217	0.049	0.051	0.682	2.40

Two-point knot snapshot at 61k vs 16k (descriptive). Same instrument on the 16k seed-0 final and the 61k pilot final: no dead knots at either point (corpus-min share 0.0016/0.0008); L0 sharpens to 0.9966 winner-take-all (effective K 1.37$\rightarrow$1.01) while L2 broadens (effective K 3.27); the dominant 61k phenomenon is learned-grid migration — only 5/12 layers (L6–L10) retain a knot at

$\tau \geq 256$, running longer (up to 1096 vs $\sim$1004); the other seven slid below the fixed-256 definition (max τ 52–236), a statement of grid position, not functional loss. The router survives $\approx$1B tokens, rearranged.

Frozen-grid controls: what does the migration cost here. Same arms, same seed, byte-identical recipe, one change surface: the half-life grid frozen at its dyadic initialization (`learn_grid=false`; the router keeps learning). Single-seed cells vs. their learnable counterparts: combined arm $\Delta_{\text{MHLDN}} = 117.755 - 116.281 = +1.474$ (between the pre-specified 1.0 and 2.0 thresholds $\rightarrow$ partial); GLA-side arm $\Delta_{\text{MHL}} = 123.469 - 122.759 = +0.710$ ($\leq$1.0 $\rightarrow$ null). Family verdict grid-learning-mixed: on the WT-103 axis at this budget and seed, grid learning costs the GLA side nothing and the combined side a moderate intermediate amount ($\approx$22% of the anchor gain); LAMBADA, descriptive and not part of the rule, degrades consistently on both arms ($+5.84$/$+10.01$), suggesting the migrating grid's benefit is corpus-stratified rather than uniform. We record the cost spectrum without a necessity statement — the migration observed above is real but its LM payoff at this budget is budget-local; hybrid setups (partially frozen grids) are a possible design option, not a tested claim.

M KNOT ABLATION AND RATE-PRESERVING MERGE: IS THE BUDGET LOAD-BEARING

Eval-time interventions on the three frozen GLA-side (MHL) mixture checkpoints, applied at the computed gate *with the decay recomputed under the intervened allocation* (the mixer's deployed object; nominal 32-window WT-103 head evaluator, identical to the paired result's). Two ablations: DOM, per layer the router output replaced by one-hot at the corpus-mean dominant knot; DROPLONG, knots at $\tau \geq 256$ zeroed and the rest renormalized. L00 is the built-in negative control (near single-knot; its interventions read $\pm$0.03–0.04 ppl — which also answers the bottleneck question: were its single-knot operation a router bottleneck at the input layer, these interventions would move the perplexity there; $\pm$0.03 reads instead as the layer genuinely wanting one timescale). All-layer ablations damage the model in 3/3 seeds: DOM $+36.9$/$+48.4$/$+36.2$ ppl, DROPLONG $+29.4$/$+28.3$/$+26.3$ ppl against baselines 121.5–124.1 — the allocation is functionally load-bearing, not decorative, and the droplong damage localizes 94% of its total to L07–L11 (peak L9 $+10.4$), i.e. the long knots carry the deep layers' retention work. ***Rate-preserving merge.*** The same instrument merges each layer's least-cost adjacent pair into one knot, choosing the merged rate analytically to preserve the effective rate on the calibration trajectory ($\lambda^* = \Sigma_t \eta_t^2 \bar{a}_t / \Sigma_t \eta_t^2$ with η_t the pair mass; exact least-squares solution of the pairwise rate error). All-layer merging of K=4 $\rightarrow$ 3 cost-effective knots costs only $+1.83$/$+2.33$/$+2.27$ ppl against $+$26–29 for direct deletion of the same knots: 92.7% of the ablation damage is first-order *rate* damage, recoverable by the remaining knots — the budget is load-bearing chiefly through the rates it carries.

The MHLG side. The same two instruments, run identically on MHLG itself (MHLDN, G=1, seeds 0–2; knot selectors from MHLG's own corpus source `evals/h2_lm_knotmass_dn_s{0,1,2}.npz`, same head evaluator; sources `evals/h2_knotprune_dn_s{0,1,2}.json` and `evals/h2_lm_knotmerge_dn_s{0,1,2}.json`; rules fixed before computation). Both pre-specified rules hold on the MHLG side. Long-knot necessity holds — all-layer DROPLONG damages $+45.1$/$+39.7$/$+43.0$ ppl against baselines 116.3–119.0 in 3/3 seeds, with the damage deep-layer localized on this side too (per-layer profiles null at L0–L4, $|\Delta| \leq 0.50$ ppl (range -0.50 to $+0.24$), carried by pockets rotating across L5–L7 and L9–L10); and the rate-preserving merge preserves quality — the all-layer rate-preserving merge costs only $+1.39$/$+0.35$/$+0.69$ ppl against $+25.4$/$+7.0$/$+19.0$ for direct deletion of the same knots, recovering 94.5/95.0/96.4% ($\geq$ 90% in 3/3) of the deletion damage. One robustness asymmetry against the GLA-side grid: forcing one-hot operation (DOM) costs MHLG only $+6.2$/$+10.7$/$+10.8$ ppl where the GLA-side grid paid $+$36.9–$+48.4$ — MHLG's allocation is robust to routing collapse in a way the GLA-side grid's is not, so on this side the dominant-knot geometry is nearly sufficient while the long knots remain the load-bearing slice. Instrument note: dropping the long knots leaves 0.126% of corpus per-token rows with essentially no residual mass; the pre-specified clamp handles them identically on both sides (counted, never asserted).

Deployed $K-1$ routers (distilled; GLA side). The GLA-side analytic merge carries one structural liability, which we address by measurement rather than argument: its λ^* was calibrated on the test corpus. Shifting calibration to a held-out training shard moves the merged half-life per layer by ratios 0.893–1.082 (median log-ratio -0.019, max $|\log|$−ratio 0.113; all twelve layers inside $[0.8, 1.25]$) — the merge economics are corpus-stable (`evals/km1_merge_calib_shift.json`). Deployment: freezing the trained seed-2 backbone and the merged 3-knot grid, a $K-1$-output router distilled on training-corpus features (3,000 steps, rate-MSE + KL to the merged teacher) costs $+1.048$ ppl on the pre-specified head evaluator (teacher 121.54, student 122.59) — recovering 54% of the $+2.27$ analytic-merge cost end-to-end — and a uniform $+0.0081$ nats full-corpus with no surprisal-stratum blow-up (per-octile $\leq +0.014$). Distillation quality per layer: mean $|\Delta a|$ at most 1.5×10^{-3} (L09; layer-mean 0.93×10^{-3}), router KL at most 8.8×10^{-4} (L11; layer-mean 5.4×10^{-4}) (`evals/km1_distill_quality.json`). Speed: router parameters -25% and router latency -3.3%; end-to-end forward throughput is a wash (-0.05%) — in this regime the compression's return is readability and parameter count, not wall-clock (`evals/km1_distill_speed.json`; measured on an L20Z GPU). The survival bound $|\Delta \log \kappa| \leq \sum |\Delta a|$ holds with measured student-side slack (per-window cumulative deviation p50 1.1–2.9 against the analytic envelope). Across three seeds the head-window cost replicates: $+0.677/+0.534/+1.048$ ppl (mean $+0.753$; each new seed inside the pre-specified $(0, 4.0]$ band, mean within both thresholds $\leq$2.0 and $\leq$ naive drop-long mean$/8 = +3.506$ against $+28.05$ — rule km1-replicates, 37.3× quality-cost margin at 25% fewer router parameters, 1,181,184→885,888 identically on both fresh seeds; gate latency flat, 9.364→9.345 ms and 9.134→9.334 ms; end-to-end forward within ±1 ms on-device (L20Z GPU, bf16 4×2048). `evals/h2_km1_seed_consistency_agg.json`.

The distilled budget stays load-bearing under the same interventions. Replaying the gate-recomputed instrument on the three deployed students (identical nominal 32-window WT-103 head evaluator; `evals/km1_knotprune_probe.json`) at student baselines 123.436/124.637/122.589 (teachers 122.759/124.103/121.542; +0.5–1.0 distillation price): DOM-all $+28.7/+42.6/+32.6$ vs. teachers $+36.9/+48.4/+36.2$ — same magnitude and the same deep-layer profile, so the distilled router's allocation remains load-bearing after the merge. DROP-LONG is applied *structurally* on the students (zero the merged knot — distill pairs index 2 in all twelve layers — and renormalize), because its realized layer-mean half-lives sit at 183.8–305.9 — only 3–4 of the 12 layers per student remain at or above the old 256 threshold, so a literal $\tau\geq256$ threshold would be a near-no-op. It reads $+67.7/+66.7/+70.5$ vs. the teachers' two-long-knot $+29.4/+28.3/+26.3$: catastrophic, localized to L07–L11 as in the teacher (peaks L07–L09). The amplification is recorded rather than over-read: the merged knot is now the corpus-dominant knot in 10–11 of 12 layers and concentrates 0.63–0.77 of the routing mass at the sampled deep layer (L08; 0.85–0.87 at L00) where the teacher spreads its long basin over two knots, so structural single-point removal is not mass-matched to the teacher intervention — the supported reading is *load-bearing and non-redundant*, not "twice as dependent". The merged grid's downward migration (183.8–305.9) is the same "allocation persists, grid repositions" pattern as the two-point knot snapshot (App. L). Descriptive, no threshold.

N CONSOLIDATED ARMS × AXES GRID

Tables 14 and 15 collect every language-model arm in one grid. The first part lists the three five-seed family arms (MHLG, the GLA-side grid, and the GDN control) and the external attention anchor, with head-window, full-corpus, LAMBADA, PG19, validation, and recall columns; its main-arm entries are the numbers of Table 1 in a second arrangement. The second part lists the capacity, initialization-prior, clamp-basin, and trained-mixture negative controls, the structural and trained probes, and the axis-trade arms, per seed on WT-103, LAMBADA, and capped PG19 at 16k. The notes below each part give the evaluator convention of each column and the pre-specified recovery statistic of each arm; the arms are described in Appendices E and G.

O VARIATION IN TRANSFER ACROSS SEEDS

This appendix and the next concern the GLA-side grid (MHL, called the *mixture* here) against the GDN control on the full WikiText-103 test set. Recomputed from the recorded per-token streams

Table 14: Consolidated arms × axes grid, part one of two: the three five-seed family arms and the external attention anchor (part two, Table 15, carries the negative controls and probes). Perplexity columns: nominal-32 head-window test evaluator, seed mean over the seed count n given in the first column. Sources: `evals/mhldn_lm_family5x3.json`, `evals/h2_val_letter_agg.json`, `evals/mhldn_lm_fullcorpus5.json`, per-arm `evals/eval_main_h2_*.json` cells. Marks are explained below the table; "floor": recall accuracy at the chance floor (Appendix J); n/a: not applicable.

Arm (n)	WT-103 (ppl)	Δ WT-103 full corpus	LAMBADA (ppl)	PG19 16k (ppl)	Val.-min. Δ (ppl)	Recall g=32	Recall g=128
MHLDN combined, MHLG (5)	116.99	$-11.85^{\dagger}$	$274.12^{\ddagger}$	43.67	$+1.89^{*}$	$0.9935^{\S}$	0.719 ± 0.269
MHL GLA-side (5)	124.13	$-8.62^{\dagger}$	293.29	49.10	$+3.03^{*}$	0.522 ± 0.060	floor
GDN_CM control (5)	128.45	(ref)	293.83	43.45	(ref)	0.990 ± 0.017	floor
ATTN anchor (seed 0)	103.60	n/a	226.96	448.64	n/a	n/a	n/a

†Paired 5-seed WT-103 *full-corpus* result vs. same-gate control or shared-gate arm, CI below zero, 5/5. ‡Sensitivity analysis only (not the primary estimate): the seven-seed rule that drops the single unfavorable pair gives -32.49, CI_{95} $[-38.91, -27.16]$, 6/6; the all-pair seven-seed mean is -23.33 (App. Q.2). *5-seed training-corpus validation-minimum delta vs. control, CI $[+0.29, +3.69]$ (4/5) and CI $[+2.02, +3.87]$ (5/5): the copy-shaped penalty band, decomposed in App. P. §Closure of the g=32 gap: 98.6–103.2% of the control deficit, 0.9935 ± 0.0066 (0.990 ceiling); at g=128 the seed-dependent instance-level regime dissolves at rates beyond the published convention (App. J); attention solves retrieval natively (its LM-side long-window collapse shown instead). Combined-arm budget persistence at ≈1B tokens: sep_{bq} $+2.277/+7.149/+11.712$ (s0/s1/s2): 3/3 same-sign, 2/3 at-or-above the 16k-anchor ratio (budget-persistence class bq-persist2, seed heterogeneity $r \in [0.335, 1.723]$; App. Q.3).

(zero new compute; per-seed head/full values reproduce the pre-specified reconciliation to $<10^{-4}$). Table 16 lists the control-referenced surprisal-octile NLL differences (mixture − control, nats; negative favors the mixture) per seed on the full test corpus, with head, full, and tail aggregates. All three seeds show the same monotone pattern: the mixture pays on the low-surprisal octiles ($+0.199$ to $+0.290$ at octile 0) and gains on the high-surprisal octiles (-0.416 to -0.565 at octile 7). Seed 0, the only seed whose full-corpus difference is positive (Table 17), combines the largest octile-0 penalty with the smallest octile-7 gain, so the balance of the two surfaces tips towards the control.

Splitting the octile-7 term into its head-window and tail parts gives $-0.407/-0.419$ (head/tail) at seed 0, $-0.446/-0.413$ at seed 1, and $-0.634/-0.545$ at seed 2: the gain is larger in the head window than in the tail only in the two seeds where the mixture wins on the full corpus.

P Additional language-model training seeds

Two further seeds ($\{3, 4\}$, trained under the byte-identical recipe on the same hardware as seeds 0–2; the design was adjusted to the available hardware before training) extend the paired full-test-set WT-103 comparison of the GLA-side grid (MHL) against the GDN control to n=5 (Table 17). Decision rule fixed before computation: the extension counts as robust (stored outcome SEEDEXT-ROBUST) iff (i) mean paired $\Delta < 0$, (ii) $\geq 4/5$ seeds negative, (iii) paired bootstrap 95% CI (rng(48), 10^4 resamples) entirely below zero.

The paired streams use the exact continuation of the pre-specified evaluator convention (per-token teacher forcing, chunked 2048 non-overlapping; atomic per-seed banks); the pre-specified 3.75σ criterion of the original paired tolerance checks (Appendix D.2) is not re-evaluated in this extension.

Five-seed axis reconciliation and the validation results. The training-corpus validation minimum — the run-log series minimum behind the three-seed $+3.53$ — has five-seed results for all three family arms; seeds 0–2 are the original three-seed runs themselves, so three-seed subsets reconcile by construction. Per-seed validation minima for seeds 0–4: MHLDN $43.03/41.30/42.73/44.64/43.02$, MHL $43.76/44.43/46.64/43.07/42.50$, and GDN_CM $40.15/41.80/42.30/39.59/41.40$. Paired results (same bootstrap settings): GLA side MHL−GDN_CM $= +3.03$ (CI $[+2.02, +3.87]$, 5/5 positive; three-seed subset $+3.526$); combined side MHLDN−GDN_CM $= +1.89$ (CI $[+0.29, +3.69]$, 4/5 positive; exact sign-flip on the paired deltas p=0.1875 — direction-consistent, sub-significance, hence recorded as a small cost, the calibration the exact tests of App. F give); MHLDN−MHL $= -1.14$ (CI $[-2.96, +0.69]$). Endpoint (step-16000) validation deltas agree in sign ($+1.43$, CI $[-0.51, +3.53]$). Both gate-bearing arms thus concede a small validation minimum to the control while winning every held-

Table 15: Consolidated arms × axes grid, part two of two: the capacity / init-prior / clamp-basin / trained-mixture negative controls, the structural and trained probes, and the axis-trade arms. Test perplexity per seed (s0–s2) on the nominal-32 head-window WT-103 evaluator, the LAMBADA stream, and capped PG19 at 16k; the four remaining columns of Table 14 are not applicable for these arms. The GDNLAD row is a three-seed mean. Blank cells: seed not listed here. Row notes follow the table.

	WT-103			LAMBADA			PG19 16k		
Arm	s0	s1	s2	s0	s1	s2	s0	s1	s2
GDNLAD init-prior (mean of 3 seeds)		129.65			286.65			43.84	
GDNCAP capacity$^{\mathrm{ms}}$	127.61			296.30			43.97		
GDNBSN clamp-basin$^{\mathrm{ms}}$	128.07			289.44			42.95		
GDNRM trained mixture$^{\mathrm{ms}}$	128.56			284.00			43.46		
RETNET pinned multi-scale¶ (s0$^{\mathrm{ms}}$)	122.43	114.75		256.07	262.12		50.52	48.62	
RETNETM matched 133.4M	132.22	134.10		306.36	297.43		50.65	53.48	
MHLDNH trained hard gate	115.31	113.37	118.98	248.97	271.30	265.81	43.72	44.54	43.76
STATICH static per-head rate	119.21			288.10			45.33		
HYBCS clamp-slow + routed free	114.32	120.19	121.34	264.89	267.61	289.40	43.39	44.46	44.74
RFLONG slow-floor 0.5 + routed fast	111.96	120.95	114.32	261.60	270.50	284.26	44.09	43.10	44.15
GDNRM-QUANT post-hoc snap	109.39			271.91			62.30		
A1B convex gate form	124.21			282.27			47.80		
G4 value-grouped	116.97	119.47	112.46	298.51	287.81	245.60	45.40	42.48	43.51

The negative-control rows are decision-rule cells whose WT-103 entries carry the pre-specified recovery statistics (r=−0.67 capacity; s=−0.73 clamp; s=−0.81 trained mixture; +1.62, CI [−1.01,4.97] init-prior — all fail to recover the result). $^{\mathrm{ms}}$Three-seed replicated: 9/9 cells ≤ 0.25 threshold, bands in App. E. A1B is the convex gate-form probe (s = −0.17, below the recovery threshold; a second seed concurs) and G4 is the value-grouped variant (s = +0.90, an axis-trade arm, not an upgrade claim). The RETNET row carries both seeds (seed-1 LAMBADA 262.12 matches the combined arm while PG19 stays penalized), and the GDNRM-QUANT row is the post-hoc value-quantization isolation (quantized values suffice on WT-103 while PG19 collapses; not a trained arm). The RETNETM row is the matched-capacity antecedent (133.400M, −0.5% vs. the family; all three seeds far below the 0.25 threshold — the parameter count is not what separates the families). MHLDNH is the trained hard-commitment arm (seed-fragile at n=3: two cells beyond the anchor ratio and one partial cell at 80% of the threshold, PG19 in the family band at all three; all five seeds positive on WT-103, App. E — its seed-0 LAMBADA, 248.97, is second only to G4 s2, 245.60, in this table), and STATICH is the per-head learned constant-rate control (s = +0.57, partial); both move the two-axis boundary onto *trained-with*, not *soft*. RFLONG is a constant slow-mass floor of 0.5 plus a routed fast side (s=+1.64/+0.31/+1.29), and HYBCS is a clamp-basin slow + routed-free hybrid (s = +1.29/+0.43/+0.26); both are seed-fragile at n=3: positive at all three seeds, the full carry being the strongest-seed event. ¶Pinned multi-scale structural control: stamp 152.013M (+13.6%), below the recovery threshold at s = +0.096; its seed-0 LAMBADA (256.07) is the lowest among the negative-control and structural rows, its PG19 sits at the GLA-side penalty level — the strong and weak sides are reported separately in App. E. Frozen-grid cells (same arms/seed conventions as the body grid): Δ_{MHLDN} = +1.474 (partial), Δ_{MHL} = +0.710 (null) — mixed at the family level. The K-grid curve across two seeds: seed 0 K=4 sweet spot (s_2=0.681, s_7=0.785 both failing the robustness thresholds), seed 1 inverted (K=2 best at 113.443) — seed-mixed; the frozen-grid probe in App. L, the grid-granularity cells in App. Q.12.

Table 16: Control-referenced surprisal-octile NLL differences (mixture − control, nats; negative favors the mixture) per seed on the full WikiText-103 test corpus, with head, full, and tail aggregates. The same signed surface trend appears in all seeds, differing in endpoint coefficients; seed 0, whose full-corpus difference favors the control, carries the largest copy-surface penalty (octile 0) with the smallest tail gain (octile 7). The head/tail split of the octile amplitudes (text) shows amplification in the winning seeds only.

	Surprisal octile								Aggregate		
Seed	0	1	2	3	4	5	6	7	Head	Full	Tail
0	+0.290	+0.242	+0.148	+0.074	+0.004	−0.101	−0.225	−0.416	−0.00265	+0.00203	+0.00350
1	+0.199	+0.165	+0.132	+0.075	+0.010	−0.127	−0.275	−0.420	−0.03509	−0.03001	−0.02841
2	+0.238	+0.193	+0.105	+0.055	−0.007	−0.121	−0.292	−0.565	−0.08616	−0.04933	−0.03774

out evaluation — the penalty is copy-shaped on both sides of the write axis, so no test gain reported here is a progress-axis artifact. Remaining three-to-five-seed shifts on the three-seed descriptive row: PG19 +5.00/+4.98/+5.46 → +5.06/+5.04/+5.65 (signs intact, $|\cdot|\leq 0.2$); LAMBADA (GLA) −8.76 → −0.54 (3/5 negative, s.d. 23.0: the sole drifting axis; the combined arm's −19.70, 4/5, persists with one recorded control-side outlier pair at seed 3); head-window WT-103 results −5.23 → −4.32 (GLA) and −10.48 → −11.46 (combined). Source for the validation-minimum and three-seed rows: `evals/h2_val_letter_agg.json`; for the head-window results: `evals/h2_gla_cell_agg.json`, `evals/h2_lm_seedext_agg.json`, and `evals/mhldn_lm_family5x3.json` (rules fixed before the single aggregation read; rng(48), 10^4 resamples).

Table 17: Paired seed extension of the GLA-side grid (MHL, "mixture") against the GDN control on the full WikiText-103 test set. Mean $\Delta = -3.227$ ppl, $4/5$ seeds negative, bootstrap 95% CI $[-5.664, -0.790]$ — all three rules pass. The single positive seed, seed 0, is the one whose octile profile tips towards the control (Appendix O); the head-window secondary axis (re-measured, *not* a re-read of the pre-specified result) is negative in $5/5$. The control-referenced surprisal-octile profiles of the new seeds reproduce the octile-0 penalty and octile-7 gain of Table 16.

	seed 0	seed 1	seed 2	seed 3	seed 4
Δ full-corpus (ppl)	$+0.280$	-4.292	-7.090	-0.110	-4.925
Mixture full-corpus ppl	138.24	140.89	140.21	143.53	142.05
Control full-corpus ppl	137.96	145.18	147.30	143.64	146.97
Δ head window, 64k tokens (ppl, descriptive)	-0.321	-4.445	-10.926	-2.554	-3.356

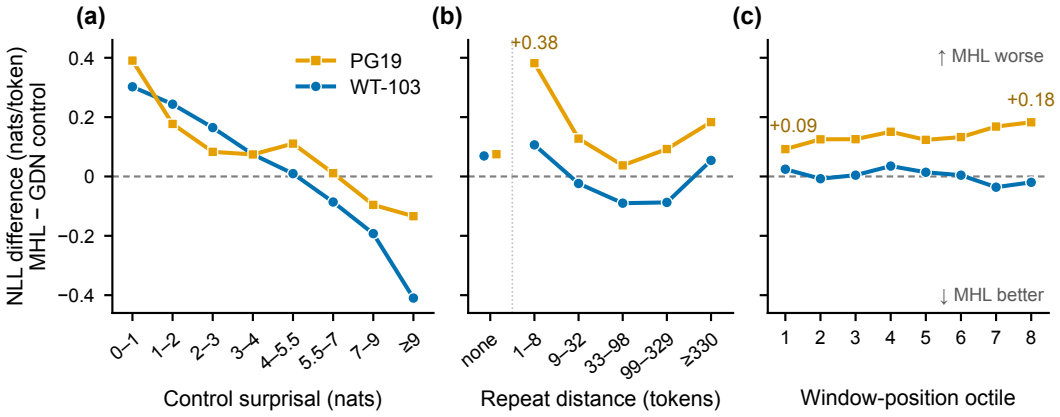


Figure 3: Token-stratified NLL difference, GLA-side grid (MHL) minus GDN control (positive: MHL worse), seed-0 checkpoints, 32,768 tokens per corpus (16 windows of 2,048 tokens). Points are token-weighted bin means for bins with $n \geq 200$; they are descriptive, with no intervals. (a) Control-model surprisal bins. (b) Distance in tokens to the previous occurrence of the target token in the window (log-spaced bins; "none" is a first occurrence, drawn unconnected). (c) Window-position octile. The PG19 penalty is largest for near repeats ($+0.38$ at 1–8 tokens) and grows with position ($+0.09$ to $+0.18$), whereas WT-103 stays near zero across positions. These associations do not isolate a copying mechanism.

Validation-stream octile anatomy. Where the $+1.89$ validation cost lives: the full held-out stream (FineWeb-Edu validation, `fw_val`, 41.9M tokens, teacher-forced 2048-token windows, both seed-0 finals) carries a paired mean delta of $+0.0194$ nats/token (MHLDN$-$GDN_CM; $\approx+0.81$ ppl at the operating point). Split into surprisal octiles by the GDN_CM-side token surprisal ($\approx$5.24M tokens each), the per-bucket deltas read $\{+0.069, +0.144, +0.127, +0.088, +0.048, +0.003, -0.067, -0.257\}$: the cost is a *band* over octiles 0–2 peaking at octile 1, not an octile-0 monopoly — the same axis family as the GLA-side copy penalty but a different local configuration — and it is partly offset by a long-tail gain in octile 7 (-0.257). Descriptive, no threshold.

Decomposition: pooling the recorded buckets, the in-domain penalty is the copy surface and only the copy surface. Octiles 0–2 (15.7M held-out tokens) carry a pooled mean of $+0.114$ nats — the whole in-domain cost — while the copy-stripped octiles 3–7 (26.2M tokens) read -0.037 nats, and the $+0.0194$ nats full-field mean is exactly the band's diluted residue ($15.7/41.9 \times 0.114 \approx 0.043$, and the tail's ≈-0.023 weighted portion, reproduces the field number within rounding). Read against the held-out results: the WT-103/LAMBADA gains are real and CI-closed but partially transfer-amplified relative to the train stream ($\sim$42 vs. 117–128 ppl); what the in-domain stream pays is paid on near-deterministic copy tokens, and the post-copy-surface remainder at seed 0 reads a small gain. The five-seed extension (`evals/h2_lm_fvval_seeds_agg.json`, pre-specified protocol — the $s0$ reference buckets, paired per-seed streams $s1$–$s4$)

puts seed-level bands on both halves: for seeds 0–4 the band means are $+0.114$, $+0.004$, $+0.009$, $+0.009$, and $+0.020$, and the tail means are -0.037, -0.012, $+0.008$, $+0.013$, and $+0.012$ (full-field $\{+0.019, -0.006, +0.008, +0.012, +0.015\}$ nats; $+0.82/-0.24/+0.34/+0.49/+0.62$ ppl). The band penalty is seed-persistent (band-tax: mean $+0.031$ nats, seed-level bootstrap CI_{95} $[+0.007, +0.073]$, entirely above zero, skewed by the seed-0 cell), the copy-stripped tail is tail-parity at five seeds (mean -0.003, CI $[-0.022, +0.012]$, spanning both signs) — the $s0$ small-gain read was the favorable end of the spectrum, recorded; the in-domain octile-7 gain does *not* replicate. Net at n=5: the in-domain cost is real, small, and copy-surfaced; everything else is parity. The positional non-head slice (positions ≥65,536; App. E) is a complementary window-position record, not in-domain evidence: it comes from the GLA-side WT-103 *test* stream — a different model arm and a held-out distribution — where the octile-7 term replicates at five seeds (-0.470 nats mean, $5/5$ negative). The in-domain record stays the FineWeb validation axis above (five-seed, span-crossing); the two strata answer different questions and neither is traded for the other.

Q Proofs of Theorems 1 and 2, and additional analyses

Q.1 Conditional approximation and identifiability results

Theorem 2 (K-term approximation of completely monotone kernels; instrument theory)**.** *Let $S(t) = \int_{[0,L]} e^{-\lambda t} d\mu(\lambda)$ with total mass $M < \infty$, and let $T > 0$. If μ has at least K support points, there exist rates $0 \le \lambda_1 < \cdots < \lambda_K \le L$ and weights $w_i > 0$ such that*

$$\sup_{0 \le t \le T} \left| S(t) - \textstyle\sum_{i=1}^{K} w_i e^{-\lambda_i t} \right| \;\le\; M \, \frac{(TL)^{2K}}{(2K)!} \;\le\; M \Big(\frac{eTL}{2K} \Big)^{2K}. \tag{7}$$

If μ has fewer than K points, the left side is 0. This is a statement about the surplus estimator's hypothesis class *— it describes what a K-term fit to a realized survival curve can certify, not what any trained scalar gate stores (Proposition 1).*

Proposition 3 (Achievable trajectories and readouts of the implemented gate)**.** *For the scalar log-convex formula with freely chosen simplex allocations over fixed knots $a_{\min}{=}a_1 < \cdots < a_K{=}a_{\max}$: (i)* within a window*, the reachable survivals are exactly the single-atom family $\{e^{-Wr} : r \in [a_{\min}, a_{\max}]\}$ (the closed-simplex relaxation includes the endpoints; finite softmax logits approach but do not attain them); (ii)* across windows*, freely specified allocations in disjoint windows can vary independently: the resulting window-endpoint survivals over M windows are the full interval product $\left[e^{-Wa_{\max}}, e^{-Wa_{\min}}\right]^M$; These are properties of the free-allocation class, not expressivity guarantees for one shared linear router. (iii)* Conditioning on identical keys, writes, and other gates*: allocation paths with the same suffix-survival profile $(\prod_{t>j} \gamma_t)_{j \le W}$ induce identical window states for any write content — dependence runs only through the realized rate sequence (γ_t) — and the linear rate map leaves a $(K{-}2)$-dimensional null family in the allocation coordinate (rate-invisible variation recorded in Appendix Q.12); (iv)* positioning*: the grid's K constrains the allocation coordinate's resolution, not the hull's coverage; what the scalar gate does not realize is the multi-atom window parallel-bank comparison in Proposition 1; the auxiliary sum in Theorem 1(ii) does not create physical bank states.*

Corollary 2 (Log-domain knot cost)**.** *Uniform* log-metric *accuracy η of a K-atom representation on $[0, T]$ is certified by $K \ge \max\big(eTL,\ (LT + \ln(2/\eta))/(2\ln 2)\big)$: relative control at the exponentially small tail costs the conversion factor e^{LT} against Theorem 2's absolute bound — the log-domain worst case grows* linearly *in LT. This certifies over pathological spectra only, and quantifies why the surplus instrument uses an empirically validated prefix-fit (at the surplus instrument's T=512 window with $L{=}\ln 2/8$ — the fastest learned knot of App. Q.12 — the first branch gives $K{\ge}\lceil eTL \rceil{=}121$; the log branch contributes <50 at every $\eta{\ge}10^{-10}$, so the max sits on the first branch).*

Corollary 3 (Knot count and identifiability)**.** *(i) Worst-case accuracy ε on $[0, T]$ is certified by the smallest K with $(eTL/2K)^{2K} \le \varepsilon/M$; the regime is super-exponential in K once $2K > eTL$. (ii) Separation of K_0 true atoms alone does not ensure that an arbitrary overparameterized K-component fit recovers them. The explicit local stability statement in Appendix Q.1 concerns a two-atom truth with the stated mass, engagement, and perturbation conditions. The empirical surplus criterion is a fit comparison, not a general certificate of identifiable memory components.*

Remark 2 (Self-averaging under stationary switching)**.** *Under a stationary ergodic allocation path, both parameterizations are long-window single-rate:* $T^{-1}\log S(T)$ *converges a.s. to* $-\mathbb{E}\big[\sum_c w_c a_c\big]$ *(log-convex) and to* $\mathbb{E}\big[\log\sum_c w_c\gamma_c\big]$ *(convex), whose gap is the stationary mean swing, bounded per Lemma 3. Hence* no *heavy-tailed long-window survival arises from a stationary router in* either *form: measured heavy tails must come from cross-sectional aggregation (Lemma 2) or non-stationarity (ergodic theorem classical; characterization ours, Appendix Q; measured dispersion in Appendix Q.11).*

Separated-atom certification. The four-point device below recovers the classical two-term case of Prony's method (Prony, 1795); see (Potts & Tasche, 2010) for the exponential-sum identification literature. What is ours is the explicit engagement threshold and the polynomial stable-certification bounds that the instrument's operating regime inherits.

Corollary 4 (Separated-atom certification)**.** *Let the realized marginal be an exact two-atom law* $S(t) = w_1e^{-ta_1} + w_2e^{-ta_2}$ *with masses* $w_1, w_2 \geq \mu > 0$*, distinct rates, and write* $\gamma_\ell = e^{-a_\ell}$*,* $\rho = |\gamma_1 - \gamma_2| > 0$*. (i)* Exact identifiability. *The pair* (w_ℓ, a_ℓ) *is determined up to permutation from* S *at any four probe points in arithmetic progression: with* $S_t = S(t)$*,* $D_t = S_tS_{t+2} - S_{t+1}^2 = w_1w_2(\gamma_1\gamma_2)^t(\gamma_1 - \gamma_2)^2 > 0$*, so the product* $p = \gamma_1\gamma_2 = D_{t+1}/D_t$*, the sum* $s = \gamma_1 + \gamma_2 = (S_{t+2} + p\,S_t)/S_{t+1}$*, and* γ_1, γ_2 *are the roots of* $z^2 - sz + p$*. (ii)* Stable certification. *If a second two-term sum* $\widehat{S}$ *matches* S *to sample error* $|\widehat{S}(t_j) - S(t_j)| \leq \varepsilon$ *at four* early *probe points* ($t_j = O(1)$*, so* $\gamma_\ell^{-t_j} = O(1)$*), there is a universal* C *with, for all* $\varepsilon \leq c\,\mu^2\rho^4$ *(*c *universal),* $|\widehat{\gamma}_\ell - \gamma_\ell| \leq C\varepsilon/(\mu^2\rho^3)$ *(in fact* $C\varepsilon/(w_\ell\rho^2)$*) and* $|\widehat{w}_\ell - w_\ell| \leq C\varepsilon/\rho^3$*: a* $K{=}2$ *representation is certified whenever the truth is two separated atoms with non-degenerate masses, and the certificate degrades* polynomially *in* $1/(\mu\rho)$*, in contrast to the linear-in-*LT *worst-case knot cost of Corollary 2. The expensive regimes (*$\mu\rho \to 0$*) are exactly the near-degenerate ones the instrument-response study of Appendix D.1 maps empirically from the other direction.*

Proof. (i) is the Cauchy identity followed by the displayed two-step inversion (exact identities verified symbolically and numerically). For (ii), the full proof — the inverse map's Jacobian in closed form, row-sum bounds, and a segment-continuation argument closing the engagement threshold — is given in Appendix Q.10; the required constants and the sufficiency of the threshold are certified by a direct adversarial scan (`evals/prony_28800_scan.json`, 28,800 perturbations, zero failures). The full proof gives both scalings: the γ bound $O(\varepsilon/(w_\ell\rho^2))$ (implying the $\mu^2\rho^3$ form of the statement), and the mass bound $O(\varepsilon/\rho^3)$ with no mass denominator; a bound of the form $C\varepsilon/(\mu\rho)$ would not be uniform in (μ, ρ). □

Q.2 Seven-seed LAMBADA evaluation and exclusion sensitivity

The five-seed LAMBADA result carried one recorded outlier pair (s3: $+31.621$, with the family-side s3 cell at 262.86 out-of-band, noted in Appendix E). The seven-seed design extends each arm by seeds $\{5, 6\}$ at the byte-identical recipe, the rule pre-specified before seeing the new numbers: over the seven paired deltas d_i (MHLDN−family), compute median m and median absolute deviation (MAD); if any $|d_i - m| > 2.5\,\text{MAD}$, drop exactly one pair (argmax-deviation), reusing the same exclusion on WT-103; CI95 = bootstrap (rng seed 48, 10k resamples) of the paired mean. Evaluation (`evals/h2_lambada7_agg.json`): deltas $\{-30.869, -25.390, -27.413, +31.621, -46.460, -37.839, -26.960\}$ give $m = -27.413$, MAD 3.456, threshold 8.640; three exceed (s3, s4, s5) and the pre-specified argmax single-drop removes pair 3 only. The two remaining exceeders both sit on the method-favoring side (-46.460 and -37.839): the pre-specified keep-up-to-one cap left s4 and s5 in place, while the sole removed pair favors the control. The exclusion therefore makes the mean comparison more favorable to the method. Post-rule $n{=}6$: mean -32.488, CI_{95} $[-38.913, -27.164]$, entirely below zero, $6/6$ negative → lambada-ci-closed. WT-103 under the same exclusion: -11.173 $[-13.454, -8.915]$, with the new seeds $-10.998/-8.865$ inside the family band. New-seed cells (WT-103/LAMBADA): family s5/s6 $125.479/126.144$ and $287.042/281.054$; MHLDN s5/s6 $114.481/117.279$ and $249.203/254.094$.

Outlier-pair diagnosis. What *is* the dropped pair? A zero-compute audit of the six trained s2–s4 training runs across GDNCM/MHLDN/MHL (`evals/lam_s3_outlier_diagnosis.json`;

per-run `log.jsonl` traces) finds no instability anywhere: final training losses sit at 3.68–3.72 for all six, gradient-norm maxima/medians at 7.6–9.9, and WT-103 cells are in-band for all three arms. The outlier is pair-coincident and off-domain: GDNCM s3's LAMBADA is 262.86 (-32 beyond its own band) while MHLDN s3's is 294.48 ($+31$ beyond its own band), with MHL s3 in-band — one shared off-domain long-form basin hit that seed on both arms at once. These diagnostics do not identify a training failure that invalidates the pair. Its exclusion is an outcome-based sensitivity rule, and the all-pair distribution remains relevant to the method comparison.

Non-drop robust reads (descriptive). A further check: whether the CI-closed result holds under a pre-specified non-drop statistic. We record the answer as a descriptive annex — the all-pair distribution remains the primary result and the pre-specified drop rule a sensitivity analysis — with the same bootstrap convention (rng 48, 10k resamples) over the seven paired deltas (`evals/h2_lambada7_robust.json`): the **median** is -27.413, CI_{95} $[-37.839, -25.390]$, entirely below zero; the one-level **winsorized mean** is -30.243, CI_{95} $[-37.559, -1.983]$, entirely below zero (upper edge thin, noted); the trimmed mean is -29.694; and the pre-specified rule's own read (-32.488, CI below zero) sits inside this neighbourhood. These robust summaries remain favorable without removing a pair, but they estimate different location statistics from the all-pair mean and do not establish that the exclusion is immaterial. For the distribution alongside the rule (`evals/h2_v820r1_q2q5.json`): the seven paired deltas sort $\{-46.460, -37.839, -30.869, -27.413, -26.960, -25.390, +31.621\}$ (6/7 negative, median -27.413, outlier-inclusive mean -23.330); the exact two-sided sign test reads p=0.1250 at n=7 (non-significant at this n — power at seven pairs is weak), p=0.03125 at the post-drop n=6 after the recorded drop, while the WT-103 seven-delta streak reads $7/7$ negative at p=0.015625 (mean -11.023). The sign axis and the CI axis tell the same layered story as the drop rule; we claim no more than the numbers carry.

Timing of the rule: the 2.5×MAD single-pair rule and its three guards (threshold, drop cap, cross-axis paired drop) were fixed *before* seeds 5–6 were trained, in the design of the seed extension (new data unseen at design, rule fixed first), while seeds 0–4 had already been run when the rule was fixed. The non-drop reads above are therefore the operative independence evidence; the order is reported for the record and does not by itself establish the result. Numerically: the post-drop n=6 value p=0.03125 is a conditional exact value on the subset the pre-specified rule selected — not a fresh confirmation on an untruncated population; the unconditional same-sign share is $6/7$ (two-sided sign p=0.125). On that basis the all-pair seven-seed distribution (mean -23.33) remains the primary result, and the post-drop and mean-free reads are sensitivity analyses.

Q.3 Budget-scaling pilot at ≈1B tokens

Pre-specified: 61,000 steps × batch 8 × ctx 2048 ≈999.4M tokens, single pass over a byte-verified 1,056,376,177-token bank (the validation-source shard identified byte-level and excluded from the bank; decontamination in App. Q.4; the data-preparation log records 858,529,925 byte-verified positions on the training side and the bank's `.fp32` extension). Triplet at the anchored recipe: family side (GDN_CM), combined arm (MHLDN), and family with ladder init $[16, 64, 256, 1024]$; parameter counts 133.835/134.057/133.835M. Rule: sep_{bq} = pilot family − pilot MHLDN on WT-103: ≥3.4 bq-carries / ≤1.0 bq-collapses-out / else bq-mid. Evaluation (`evals/h2_lm_bq_agg.json`): cells family $86.254/200.825/32.364, 31.116, 29.253$; MHLDN $83.977/192.879/34.060, 32.780, 30.741$; gdn-ladder $90.230/186.302/34.020, 32.922, 31.379$ (WT-103/LAMBADA/$\text{PG19}_{4k,8k,16k}$). sep_{bq} = 2.277 → bq-mid: positive but diluted (33.5% of the 16k anchor 6.799); LAMBADA +7.946 same-sign (family−MHLDN, positive = combined-advantage); PG19 reverses $-1.696/-1.488$ at 61k — descriptive, no threshold, reported as read.

Seed-1 paired cells (byte-identical recipe and bank): family $89.215/213.561/34.279, 32.964, 30.985$; MHLDN $82.066/185.279/34.372, 32.676, 30.734$. sep_{bq} = +7.149 and LAMBADA +28.282 — same-sign as seed 0 and widened (×3.1, ×3.6); the PG19 61k reversal does *not* replicate ($-0.093/+0.288/+0.251$: the two seeds carry no consistent long-context direction at this budget). Descriptive envelope, no threshold; the seed-0 triplet rule above stands as evaluated and is not re-read.

Seed-2 paired cells (identical recipe and bank) under the pre-specified three-point rule (r_s = $\text{sep}_{bq,s}/6.799$; ≥1.0 bq-persist2 / ≤0.5 bq-dilute2 / middle bq-middle2 / <0 bq-

flip): family $94.269/200.937/34.165, 33.074, 30.886$; MHLDN $82.557/175.380/34.089, 32.808, 30.710$. $\text{sep}_{bq,2} = +11.712$, $r_2 = 1.723$ — **bq-persist2**. The three-point WT-103 envelope $\{+2.277\,(0.335), +7.149\,(1.051), +11.712\,(1.723)\}$ reads $3/3$ same-sign and $2/3$ at-or-above the 16k-anchor ratio (median 1.051). Seed 0 alone suggested that the gain dilutes at the larger budget and seed 1 that it widens; with the third seed the gain persists, with genuine cross-seed heterogeneity ($r \in [0.335, 1.723]$, reported un-averaged). LAMBADA holds same-sign on all three ($+7.946/+28.282/+25.557$; seeds 1–2 widen $\times 3.2$–3.6 over seed 0). PG19 across seeds: seed-0 $\{-1.696, -1.664, -1.488\}$, seed-1 $\{-0.093, +0.288, +0.251\}$, seed-2 $\{+0.076, +0.266, +0.176\}$ — the seed-0 reversal is an outlying point; the cross-seed long-context comparison reads parity ($|sep| < 0.29$ at $s1/s2$), and the seed-0 descriptive result stands. The rule reads the $\approx$1B point only; nothing is asserted for larger budgets.

Q.4 Post-hoc training-corpus overlap scan

13-gram double-64-bit polynomial-hash scan (a hit requires both hashes to match; the instrument is validated by a $244/244$ positive control of validation-stream head grams against its byte-identified source shard). FineWeb-Edu (FW) bank ($1{,}056{,}376{,}177$ tokens): WT-103 bank-side $1{,}843$ hit grams (test-side coverage $\leq 0.675\%$); LAMBADA 0; PG19 $4{,}580$ ($\leq 0.044\%$). OWT line (470M tokens): WT-103 885 ($\leq 0.324\%$); LAMBADA exactly one gram ($\leq 0.00024\%$; the single-gram coincidence class is unresolved at this resolution and reported verbatim); PG19 $4{,}039$ ($\leq 0.039\%$). Sharing a training bank does not guarantee that overlap affects two models equally. These scans quantify exact 13-gram matches only; they do not establish decontaminated performance. The designated validation shard is absent from the training bank.

Q.5 Proof of Theorem 2 (K-term approximation)

Let $S(t) = \int_{[0,L]} e^{-\lambda t}\, d\mu(\lambda)$ with $M = \mu([0, L]) < \infty$, and $T > 0$. (1) *Quadrature.* Gauss–Christoffel quadrature on $[0, L]$ with respect to $d\mu$ yields nodes $\lambda_1 < \cdots < \lambda_K$ in $[0, L]$ and positive weights w_i with $\int p\, d\mu = \sum_i w_i p(\lambda_i)$ for every polynomial p of degree $\leq 2K - 1$ (Szegő, 1939, Thm. 3.4.1); the construction needs μ to have $\geq K$ support points (strict interiority of all nodes precisely when μ carries $\geq K$ *interior* points of increase; an endpoint atom reduces it to the lower-order boundary rule), and if it has fewer, S is already a $(<K)$-term sum and the error is 0. (2) *Taylor envelope.* For $x \geq 0$, the $2K$-th derivative of e^{-x} is e^{-x}, which is nonnegative: the Lagrange remainder of the even-order truncation is $e^{-\xi} x^{2K}/(2K)! \geq 0$ for some $\xi \in [0, x]$, i.e. the error $R(x) := e^{-x} - \sum_{n=0}^{2K-1} (-x)^n/n!$ is pointwise nonnegative and $0 \leq R(x) \leq x^{2K}/(2K)!$. (3) *Assembly.* With $x = \lambda t$, $t \leq T$: the polynomial part integrates exactly by (1), so $S(t) - \sum_i w_i e^{-\lambda_i t} = \int R(\lambda t)\, d\mu - \sum_i w_i R(\lambda_i t)$. Both terms lie in $[0, M\,(TL)^{2K}/(2K)!]$ — the second because $\sum_i w_i = \mu([0, L]) = M$ (take $p \equiv 1$) — hence their difference does not need a triangle inequality, and $|S(t) - \sum_i w_i e^{-\lambda_i t}| \leq M\,(TL)^{2K}/(2K)!$. Finally $r! \geq (r/e)^r$ turns $M(TL)^{2K}/(2K)!$ into $M(eTL/2K)^{2K}$. □

Q.6 Proof of Theorem 1(i) (log-convex window class)

$\log S(T) = \sum_t \sum_c w_c(t) \log \gamma_c = -T \sum_c \bar{w}_c a_c$ with $\bar{w} = T^{-1} \sum_t w(t) \in \Delta^{K-1}$, so $S(T) = e^{-T\bar{r}}$ with $\bar{r} = \sum_c \bar{w}_c a_c \in [a_{\min}, a_{\max}]$. Conversely, for any $r \in [a_{\min}, a_{\max}]$ the constant two-knot path $w_1 = (r - a_{\min})/(a_{\max} - a_{\min})$, $w_K = 1 - w_1$ realizes $\bar{r} = r$. Hence the class is exactly $\{e^{-Tr} : r \in [a_{\min}, a_{\max}]\}$: one atom per (head, channel-group), placed anywhere in the knot-rate hull, not pinned to the grid. □

Q.7 Proof of Theorem 1(ii) (convex window class)

For the convex form, $\prod_{t=1}^{T} \big(\sum_c w_c(t) e^{-a_c}\big) = \mathbb{E}\big[\prod_t e^{-A_t}\big] = \mathbb{E}[e^{-X_T}]$, where $A_1, \ldots, A_T$ are independent with $\Pr(A_t = a_c) = w_c(t)$ and $X_T = \sum_t A_t$: independence is a representation device (the product of means is the mean of the product), not an assumption on the allocation process. Thus $S(T) = \bar{\kappa}_T(1)$ for $\bar{\kappa}_T(s) = \mathbb{E}[e^{-sX_T}]$, the Laplace–Stieltjes transform of the positive finite measure $\text{law}(X_T)$, completely monotone by Bernstein's sufficiency direction. Equivalently, each factor is completely monotone, products of completely monotone functions are completely mono-

tone by Leibniz — $(-1)^n(fg)^{(n)} = \sum_k \binom{n}{k}[(-1)^k f^{(k)}][(-1)^{n-k} g^{(n-k)}] \geq 0$ termwise — and the representing measure of a product is the convolution $\mathrm{law}(X_T) = *_t \mu_t$, $\mu_t = \sum_c w_c(t)\delta_{a_c}$. Support: $X_T = \sum_c n_c a_c$ with $n_c = \#\{t : c(t) = c\}$, $\sum_c n_c = T$; distinct sums number at most the count vectors $\binom{T+K-1}{K-1}$ (collisions only shrink this) and trivially at most K^T. Half-life reading: the realized window rate $X_T/T \in [a_{\min}, a_{\max}]$, i.e. the window's effective half-life stays inside $[\tau_1, \tau_K]$. □

Remark (capacity bound versus realized paths). The count bound is a family-capacity statement, not a reading of any realized path: only a *hard* per-step allocation (each μ_t a point mass) makes $\mathrm{law}(X_T)$ degenerate, at the count sum $\sum_c n_c a_c$; fixed *soft* weights keep the convolution random ($\mathrm{Var}(X_T) = T(a_1 - a_2)^2/4$ on a two-knot $(\frac{1}{2}, \frac{1}{2})$ path). What collapses a realized window's display to one effective atom is therefore near-one-hot routing, not determinism of the soft path itself. At the measured $K{=}4$ operating points this is exactly the regime — the realized effective atom count K_{eff} settles at 1.0–1.5 on the single-timescale MQAR task (App. E; the language-model allocations in App. L are broader), orders below the $\min(K^T, \binom{T+K-1}{K-1})$ envelope; the envelope certifies what the family could express, the settlement measures what the trained router actually uses.

Coordinates, not partitions (identifiability). Allocations are parametrization coordinates, not identifiable memory spectra: with knot rates $(64, 16, 4, 1)$ in units of $\ln 2/1024$, the allocations $(0, 1, 0, 0)$ and $(5/21, 0, 0, 16/21)$ produce the identical effective rate 16 while carrying slowest-knot mass 0 and $16/21$; mixing either with a common uniform vector at $90/10$ yields two strictly positive, softmax-realizable allocations with the same rate and different slow mass. Under such reparameterization, survival values, effective rates, and per-window endpoint quantities are invariant — and, where routing is near-one-hot, knot-anchored occupation as well; knot weights read as independent slow-state shares are not. The RFLONG probe's guarantee accordingly scopes to its *mass floor* (slowest-knot weight pinned at 0.5; knot position learnable) — a pinned coordinate, not a decoupled slow bank. Measured null travel makes the caveat load-bearing (reparam-null; `evals/h2_reparam_null_audit.json`; five MHLDN finals, 16,384 tokens per stream fixed across seeds, euclidean simplex tangent, descriptive): the within-seed token-cloud null share runs $0.52/0.40/0.30$ by layer thirds (bottom/middle/top; `lambada` read $0.50/0.38/0.28$, convention-invariant), cross-seed matched pairs differ ≈ 0.43 of their simplex distance along rate-invisible directions with the allocation/rate-distance correlation climbing $0.44 \to 0.9$ with depth (layer 0 itself is the outlier: ≈ 0.9 null share at correlation 0.44 — the first layer's allocations are nearly rate-free across seeds), and the trained grids hold their spread (median log-increment 1.26 vs. init $\ln 4 \approx 1.39$; no collapse, no blowout, zero heads at the collapse threshold). One third to one half of the realized coordinate is therefore rate-invisible at the operated cells.

Q.8 Proof of the log-domain conversion behind Corollary 2

Under Theorem 2's hypotheses, $S(t) \geq Me^{-Lt}$ on $[0, T]$; with $\varepsilon_K = M(eTL/2K)^{2K}$ and $\rho_K := e^{LT}(eTL/2K)^{2K}$, $|S - S_K|/S \leq \rho_K$. If $\rho_K \leq 1/2$, then $S_K \geq S/2$ and $|\log S(t) - \log S_K(t)| \leq |S - S_K|/\min(S, S_K) \leq 2\rho_K$, using $|\log a - \log b| \leq |a - b|/\min(a, b)$. Choosing $K \geq eTL$ makes the base $eTL/(2K) \leq 1/2$, so $\rho_K \leq e^{LT}2^{-2K}$; and $2K \ln 2 \geq LT + \ln(2/\eta)$ gives $\rho_K \leq \eta/2$, hence log-domain error $\leq \eta$. □

Q.9 Proof step for Remark 2 (self-averaging)

Under the stated stationarity assumption, Birkhoff's ergodic theorem applies directly to the two stationary sequences $r(t) = \sum_c w_c(t)a_c$ and $\log \sum_c w_c(t)e^{-a_c}$, yielding the two displayed almost-sure limits; the exponent gap ordering is per-token Jensen with the Lemma 3 bound taken under the stationary expectation. □

Q.10 Full proof of Corollary 4(ii)

Take probes $t = 0$ (early $t = O(1)$ rescales all displays by the $O(1)$ factor γ_ℓ^{-t}), order $\gamma_1 > \gamma_2$, write $\rho = \gamma_1 - \gamma_2 > 0$, $S_j = S(j)$, $M = S_0 = w_1 + w_2$. Part (i) is the Cauchy identity $D_t = S_t S_{t+2} - S_{t+1}^2 = w_1 w_2 (\gamma_1 \gamma_2)^t \rho^2$ followed by the exact two-step inversion $p = D_{t+1}/D_t = \gamma_1 \gamma_2$,

$s = (S_{t+2} + pS_t)/S_{t+1} = \gamma_1 + \gamma_2$ (the classical two-term Prony method); $\gamma_{1,2}$ are the roots of $z^2 - sz + p$ and $w_1 = (\gamma_2 S_0 - S_1)/(\gamma_2 - \gamma_1)$, $w_2 = S_0 - w_1$.

Step 1: the inverse map's exact Jacobian. The inversion is rational in the four observations. Differentiating the displayed identities and evaluating at the two-atom point gives the four rows in closed form:

$$\partial\gamma_1/\partial S = \frac{1}{w_1\rho^2}\big(-\gamma_1\gamma_2^2,\ \gamma_2(2\gamma_1+\gamma_2),\ -(\gamma_1+2\gamma_2),\ 1\big), \tag{8}$$

$$\partial\gamma_2/\partial S = \frac{1}{w_2\rho^2}\big(-\gamma_1^2\gamma_2,\ \gamma_1(\gamma_1+2\gamma_2),\ -(2\gamma_1+\gamma_2),\ 1\big), \tag{9}$$

$$\partial w_1/\partial S = \frac{1}{\rho^3}\big(\gamma_2^2(3\gamma_1-\gamma_2),\ -6\gamma_1\gamma_2,\ 3(\gamma_1+\gamma_2),\ -2\big), \tag{10}$$

and $\partial w_2/\partial S = (1,0,0,0) - \partial w_1/\partial S$. The mechanism is legible directly: each γ-row carries one denominator w_ℓ of its own atom and ρ^2, while the mass rows carry *no* mass denominators at all and ρ^3 — the γ-perturbation and quotient-perturbation contributions to the mass solve cancel the w factors exactly. The three displays are certified against independent numerical inversion of the four-point map (4×10^4 random points, max scaled deviation 1.3×10^{-9}; `evals/prony_jacobian_cert.json`).

Step 2: row-sum bounds. With $\gamma_\ell \in (0,1]$ and $w_\ell \geq \mu$, $\|\partial\gamma_\ell/\partial S\|_1 \leq 8/(w_\ell\rho^2) \leq 8/(\mu\rho^2)$ and $\|\partial w_\ell/\partial S\|_1 \leq 18/\rho^3$. The γ claim of the statement, $|\widehat{\gamma}_\ell - \gamma_\ell| \leq C\varepsilon/(\mu^2\rho^3)$, follows once the segment is closed below, since $18/(\mu\rho^2) \leq 18/(\mu^2\rho^3)$ whenever $\mu\rho \leq 1$. A *mass* bound of the form $C\varepsilon/(\mu\rho)$, by contrast, cannot hold with any universal C: the third display forces $C \geq \mu/\rho^2$-type right-hand sides, which are unbounded as μ grows against ρ^2. The proved sharp form, as stated, is $|\widehat{w}_\ell - w_\ell| \leq C\varepsilon/\rho^3$.

Step 3: engagement and segment continuation. Let $\varepsilon \leq \mu^2\rho^4/128$ and parametrize the segment $S(\lambda) = S + \lambda\delta$, $|\delta_j| \leq \varepsilon$. Let $A \subseteq [0,1]$ be the set where $\widehat{\rho}(\lambda) = \sqrt{\widehat{s}^2 - 4\widehat{p}} \geq 2\rho/3$ throughout $[0,\lambda]$; A is open and contains 0. On A the rational map is C^1 and the Step-2 rows, with ρ replaced by $2\rho/3$, bound $|\partial\gamma_\ell/\partial S|_1 \leq 9\cdot8/(4w_\ell\rho^2) = 18/(w_\ell\rho^2)$, $\|\partial w_\ell/\partial S\|_1 \leq (27/8)\cdot18/\rho^3 < 61/\rho^3$. Integration then gives $|\widehat{\rho}-\rho| \leq 36\varepsilon/(\mu\rho^2) \leq 36(\mu^2\rho^4/128)/(\mu\rho^2) = 0.28\mu\rho^2 \leq 0.28\rho$, so $\widehat{\rho} > 2\rho/3$ strictly: A is also closed in $[0,1]$, hence $A = [0,1]$. Therefore the inverse stays defined on the whole segment, the discriminant never closes below $(2\rho/3)^2$, and integration of the Step-2 displays with the $(3/2)$-inflated constants gives $|\widehat{\gamma}_\ell - \gamma_\ell| \leq 18\varepsilon/(\mu\rho^2) \leq 18\varepsilon/(\mu^2\rho^3)$ and $|\widehat{w}_\ell - w_\ell| \leq 61\varepsilon/\rho^3$. The row bounds divide by the true masses ($w_\ell \geq \mu$), so the continuation itself must keep the path's masses positive: the induced excursion $\leq 61\varepsilon/\rho^3$ stays below $\mu/2$ whenever $\varepsilon \leq \mu\rho^3/122$, and we therefore read the engagement threshold in the stronger joint regime $\varepsilon \leq \min(\mu^2\rho^4/128,\ \mu\rho^3/122)$, which keeps $\widehat{w}_\ell \geq \mu/2$ along the whole segment (the constants at the early probe points carry the certified rate interval $[a_{\min}, a_{\max}]$). In particular the stated form $\varepsilon \leq c\,(\mu^2\rho^4 \wedge \mu\rho^3)$ with universal c is a sufficient engagement threshold. The constants are first-order-optimal up to the segment factor; a direct adversarial scan certifies both displays with margin (28,800 uniform perturbations over (γ_1,ρ,μ) grids, 0 invertibility failures at $\varepsilon = \mu^2\rho^4/8$, worst scaled errors $6.79 \leq 8$ and $12.81 \leq 18$), archived at `evals/prony_28800_scan.json`; the certified inversion and the scan itself are short reference implementations (`evals/prony_28800_scan.py`). □

Remark (the two scalings). The mass bound is ρ^{-3}-shaped and free of mass denominators; a bound of the form $C\varepsilon/(\mu\rho)$ is not uniform, since no universal C satisfies it. The γ bound is $O(\varepsilon/(w_\ell\rho^2))$, of which the stated $\mu^2\rho^3$ form is a relaxation. Every qualitative usage of the corollary — polynomial degradation in $1/(\mu\rho)$, the contrast with the linear-in-LT worst-case knot cost of Corollary 2, and the near-degenerate regimes the instrument study maps empirically — rests on these two scalings.

Q.11 Dispersion and self-averaging diagnostics

The stationarity assumption of Remark 2 enters only through the cross-window dispersion D of window-mean allocations. By Theorem 1(i) a window exponent is $r = \sum_c \bar{w}_c a_c \in [a_{\min}, a_{\max}]$, so an allocation change of size D (the allocation dispersion of Table 7) moves r by at most $D\,(a_{\max} - a_{\min})$. Reading the three recorded artefacts (allocation meters s_0–s_2; learned hull grids; the 68 firing

cells of the survival-curve study, Appendix D.1): $D \leq 0.036$ (median 0.0157); learned hulls give $a_{\max} - a_{\min} \in [0.061, 0.081]$ (min/median over layers); hence any switching-induced exponent swing is $\leq 2.9 \times 10^{-3}$. The marginals' measured two-atom rate splits are $0.0289/0.0507/0.0689$ (P5/P25/P50 across cells) — at least $9.9\times$ larger than the largest swing switching could inject, at the least favourable cell. Falsifying the self-averaging reading would require dispersion $D \geq 0.63$ (P25 cell) or $D \geq 0.85$ (median cell) against a simplex total-variation ceiling of 1 and an observed maximum of 0.036 — a 17–$24\times$ shortfall. Together with the 16k-block stationarity negative control above, the self-averaging conclusion is forced by the measured magnitudes, not assumed.

Q.12 Realized rates and grid placement

Theorem 2 is a worst-case statement: on the realized rate hull $L = \ln 2/8 \approx 0.087$ (fastest learned knot) and window $T = 2048$, driving the bound to $M/2$ needs $K \geq \lceil eTL/2 \rceil = 242$ — the theorem's role is the representation link, not a justification of the deployed $K{=}4$. What licenses $K{=}4$ is the token-wise deployment: each window's *endpoint* surrogate is single-rate per Theorem 1(i) (an endpoint identity, not a curve class), so the placement problem is quantization of the realized rate interval for the router's reach, not approximation of a worst-case μ. The minimax K-point quantizer under log-domain (relative) error is geometric, with worst one-hop distance $\frac{1}{2}\ln(\max_i \tau_{i+1}/\tau_i)$. The learned grids realize this nearly exactly: across all 36 (seed $\times$ layer) grids the adjacent half-life ratios have median 3.68, range $[3.35, 4.14]$, coefficient of variation 5.1% (interval span 38.7–69.7 vs. the init's 64), i.e. the placement converged to the minimax-geometric quantizer of its own realized hull with worst one-hop log-distance 0.71; the corpus-share effective knot count $\exp H(\bar{w}) \in [2.47, 3.77]$ over the 36 (seed, layer) grids (Table 12 shares) confirms all four quantizer points are used. (These two 36-grid records were measured on the GLA-side grid; the *MHLG-side* counterparts — grid side at zero compute, `evals/h2_gridquant_dn.json`; corpus side from MHLG corpus record, `evals/h2_lm_knotmass_dn_agg.json` — keep the same trend with a looser placement: adjacency median 3.69, range $[3.29, 5.27]$, CV 11.2%, span 36–108, worst one-hop distance 0.83, band $\{7.2$–$9.6,\ 25.8$–$40.2,\ 85.9$–$190.5,\ 286$–$1004\}$ at seeds 0–2; $\exp H(\bar{w}) \in [1.56, 3.40]$.) The empirically realized grid compression $\{16, 64, 256, 1024\} \to \{5.3$–$11.1,\ 20.6$–$45.1,\ 75.9$–$167.1,\ 280$–$615\}$ thus reads as the router-and-grid system learning a quantized cover of the rate interval the data actually visits, at a per-window representational cost of one atom. *Independence from the detection instrument.* The $T{=}512$ survival-curve detection measures a marginal property of a *frozen* checkpoint (the instrument's sensitivity to a second timescale); grid placement is the output of LM *training* (the quantizer above). Neither object constrains the other; the overlap of the collapsed fast-knot band (6–11 tokens) with the instrument's τ_1 scale (15–21) is a coincidence of scales, recorded as such.

Grid granularity: $K \in \{2, 7\}$. The endpoint-invariant nested family $\{2{\subset}4{\subset}7\}$ ($K{=}7$ is the exact dyadic refinement of $16..1024$ under the log-linspace constructor, used in place of a $K{=}8$ configuration). Same recipe, 16k steps, seed 0: $s_2 = 0.6808$, $s_7 = 0.7848$ against $s_4{=}1$. The pre-specified k-robust double condition ($s_2{\geq}0.75$ and $|s_7 - 1|{\leq}0.15$) fails on both sides $\to$ k-sensitive: at seed 0, $K{=}4$ is a two-sided sweet spot at this scale — coarsening loses a third of the gain, refining does not add (curve $(0.681,\ 1.000,\ 0.785)$, non-monotone, reported as read).

Two-seed family reading (same pre-specified rule per seed; seed-1 cells at the identical recipe): seed 1 reads $s_2 = 1.4174$ (≥ 0.75) and $s_7 = 1.0672$ ($|{-}1| = 0.0672$, ≤ 0.15) — both conditions pass rule-side; yet the WT-103 curve itself inverts: $K{=}2$ is best (113.443), then $K{=}7$ (115.824), then $K{=}4$ (117.382), LAMBADA agreeing ($272.2 < 285.7$, PG19 all windows likewise), against seed 0's strict $K{=}4$ optimum with both flanks penalized. The recorded family verdict is k-seed-mixed: the two-sided sweet spot is a seed-fragile, single-budget artifact — seed 0 sensitive, seed 1 robust-with-inversion — and we make no seed-stable grid-resolution claim. The curve is reported as a cross-seed spectrum; whether the optimal K has a scale- or budget-stable rule remains open (a two-cell grid can only bound this parameter space, not map it).

Static-propensity scout over the obtained K budgets (zero-compute). To the question whether any correlate in the trained budgets separates the two seeds — and thus predicts which K wins — we scouted everything that is CPU-loadable without a single forward pass (the dynamic occupancy measurement follows below): the learned per-layer grid placement and a static, propensity-level stand-in for winner-take-all, per head and knot, from the learned router rows. The measured signatures are

K-determined, not seed-determined: mean top-1 propensity share $\{0.75, 0.50, 0.34\}$ and static $\mathrm{K_{eff}}$ $\{1.6, 3.0, 5.2\}$ at $K{=}\{2, 4, 7\}$ (L0 shares $\{0.70, 0.50, 0.37\}$), while the cross-seed deltas *within* each arm are ≤ 0.005 top-1 share, ≤ 0.03 $\mathrm{K_{eff}}$, and $\sim 10\%$ on placement span — more than an order of magnitude below the cross-arm structure. The seed that inverts the winner (seed 1: $K{=}2$ best) is statically indistinguishable from the seed that does not (seed 0: $K{=}4$ best; source `evals/h2_kseed_static_scout.json`). Reading: no static-propensity correlate separates the seeds at this instrument, so the K inversion remains genuinely seed-dependent as recorded (k-seed-mixed). The dynamic occupancy measurement concurs (`evals/h2_kseed_occ_probe.json`; six cells, 32,768 FineWeb-Edu validation tokens, one evaluation): input-driven top-1 knot occupancy is K-determined — the head-mean effective occupied fraction D runs $\{0.604, 0.317, 0.220\}$ at $K{=}\{2, 4, 7\}$ — and seed-insensitive within each arm (cross-seed $|\Delta D| \leq 0.014$), while the winner/loser contrasts invert sign across seeds ($\Delta_{s0}{=}{-}0.386$, $\Delta_{s1}{=}{+}0.382$): occ-seed-mixed. Neither measurement — static propensity nor dynamic occupancy — separates the inverting seed from the non-inverting one, so the K inversion is seed-dependent under every measurement we took of these checkpoints. The copy-token penalty reads the same way (zero-compute from existing data; `evals/h2_kseed_copytax_q4.json`): its PG19 deltas against the same-seed control are positive at seed 0 (ordered $k2{>}k4{>}k7$ at $w{=}16$k, reversed at 4k) and negative at seed 1 (the $k2$ cell most negative); the in-domain FineWeb-Edu validation gap is small (≤ 1.6 ppl late-mid) and tail-unstable in both seeds; and the final *fast* atom is K-invariant (median 8.4–8.9; only the slow end extends with K, medians $222/386/667$) — copy-tax-k-seed-mixed: the pre-specified fast-side quantization hypothesis is not supported at this measurement (coarsening at small K lives on the slow-atom reach, not the fast edge). For comparison, the dominant knots per layer of the seed-0 $K{=}4$ GLA-side grid in the knot-ablation probe of Appendix M are $\{3, 0, 2, 3, 3, 3, 3, 3, 3, 3, 0\}$ (`evals/h2_lm_knotprune_gate_r1_full_s0.json`).

On the μ-dependence of Theorem 2's nodes. The theorem should be read strictly as a representation link: its quadrature nodes are a function of the unknown marginal μ, and no placement theorem over data is claimed or implied — the realized placement is training's, quantizing the empirical rate hull as above. A data-dependent placement theorem is possible in principle only relative to an *estimable* surrogate for μ; this is precisely the role the instrument side of this paper executes — the per-window two-atom certification (Corollary 4, full proof in Appendix Q.10) supplies exactly the class of stable μ estimates on which a conditional placement guarantee could then be conditioned, with engagement thresholds deciding where the estimate is trustworthy. We state the combination as a roadmap, not a result: (i) estimate realized marginals per window, (ii) apply the quadrature with the estimated marginal, (iii) bound the placement error by the certification error of step (i) propagated through the quadrature's condition. Step (iii) is open; steps (i)–(ii) are the present paper's own machinery.

Linear read-out of within-window multi-atom content. Theorem 1(ii)'s atom bound is a support bound on the window mixing measure; the sharper question is whether a read-out could *resolve* those atoms inside one window. For the present architecture the answer is threshold-controlled in both directions, and the constants are already on paper. Any read-out linear in the realized gates observes, per window, a convex query on the same half-life hull; resolving two atoms at separation ρ is therefore a super-resolution instance over the windowed exponential family — exactly the conditioning quantified by Corollary 4: identification holds with mass $61\varepsilon/\rho^3$ and threshold $\varepsilon \leq \mu^2\rho^4/128$. These are *sufficient* stability conditions of the two-atom certificate, not universal bounds: below the engagement threshold what fails is the certificate's own coverage — the certifying estimator is not guaranteed to engage — an instrument floor, not a universal non-identifiability claim, and a linear read-out a fortiori inherits that instrument floor. Above the threshold the resolvable content scopes to the certificate's two-atom class; deeper mixtures are outside its delivery, not outside identifiability in principle. The practical corollary for Theorem 1(ii) is accordingly scoped: at the measured operating point the convex arm's swing amplitude is negligible, the log-convex arm is single-rate at each window's endpoint (an endpoint identity), and the trained read stack ($\mathrm{o_norm}$ then $\mathrm{o_proj}$) is window-linear in the rates — whether a nonlinear read-out operating near threshold could activate a swing advantage is not resolved by the certificate — so the support bound, not the read-out, is what remains non-binding.

R ADDITIONAL SCALE AND ARCHITECTURE EVIDENCE

Second size: the 400M paired cell. The capacity-matched pair at d=1280, L=18 — MHLDN 400.188M parameters vs. GDNCM 399.634M (+0.14%; tied embedding counted once, as for the 134M models) — trains the byte-identical 262.1M-token recipe, seeds 0/1, on the two pre-committed axes (WT-103, PG19@16k). Seed-paired WT-103 margins (combined minus control; seed cells 102.75/99.92 vs. 108.35/113.71): −5.60/−13.79 ppl (mean -9.70 ± 5.79, n=2 descriptive), against the anchor −11.46 at 134M — the margin holds at the second size, attenuated (the anchor remains the larger point). PG19@16k margins −2.12/−1.87 (mean -2.00 ± 0.17) against an anchor of capped parity (+0.21): the long-context penalty, a parity levy at the first size, reads *negative at both seeds* at the second — it closes with scale in this protocol, n=2 descriptive evidence recorded, not yet an established scaling trend. On the byte-identical training-corpus validation stream the paired margin stays a small penalty at both sizes (final-mean +1.43 over the five anchor seeds, 2/5 negative; +0.07/+1.62 at 400M, 0/2 negative; `evals/h2scale400_valmargin.json`): the measured win stays off-distribution at both sizes. Gate telemetry at the second size replicates the family's sharpening and its fast and middle horizons (router entropy inside the five-seed band on 86/88% of samples and at terminal; τ_{eff} p10 terminal inside the band, the mean a touch below it), while the realized slow-slot horizon terminates short of the anchor band (438.90/434.70 vs. [465.91, 522.14] tokens; source `evals/h2scale400_streamband.json`): the second-size win arrives with a marginally shorter trained slow tail, measured.

Modern open-architecture control (Mamba-2; stopped mid-training). The architecture-generality question — whether the measured retention-rate phenomenology is an artifact of the in-house GDN backbone or a property of the recipe — is tested on a modern open architecture: MAMBA-2 (the `fla` reference implementation of its state-space-duality (SSD) layer, head-dim 64 ⇒ 24 heads, state 128, expand 2), trained under the byte-identical anchor recipe (16,000 steps × 8×2,048 tokens, lr 3e−4, 129.022M parameters, −3.6% against GDNCM), seeds 0/1. The arm uses the pure-PyTorch SSD path at a reduced chunk size and an exact pure-PyTorch depthwise convolution in place of the fused `mamba_ssm`/`causal_conv1d` kernels; both substitutions were checked against the reference before training (chunk 64 vs. 256: $|\Delta\text{out}| \le 6.3\times10^{-6}$ and $|\Delta\text{grad}| \le 2.6\times10^{-8}$ in fp32; conv substitution vs. a hand-written reference: 2.4×10^{-7}; source `evals/h2_mb2_chunk_cert.json`). The instrument is the same attractor replay as the family line, applied to the open architecture's own rate knot: realized $\tau = \ln 2/(\Delta_t \cdot |A_h|)$ through a fixed probe batch, init-vs-final, with the drift predicate (stored as MB2-DRIFT-CLASS) defined on the $2L$-block stack (24 blocks: shortening in $\ge 5/6$ of layers, i.e. $\ge 20/24$, and layer-median ratio final/init ≤ 0.8), the slow-minority share as the P90/P50 τ ratio, and the LM result cells as descriptive vs. the anchors (instrument `evals/h2_m2_attractor.py`). The two runs stopped at 6.1k/6.6k of the 16k steps and were not resumed, so the final-step outcome is unavailable; we report the mid-training prefix only: s0 shortens 16/24 blocks at layer-median ratio 0.78 and s1 17/24 at 0.67 — below the $\ge 20/24$ leg of the pre-specified predicate on both seeds, with the ratios already sub-0.8; whether the joint condition would hold at completion is not decidable from these records (trajectories `evals/h2_m2_attract_prefix_s{0,1}.json`). No anchor-side claim changes either way. *Companion cell on a fully open checkpoint*: the same instrument replayed on the *pretrained* MAMBA2-130M (300B-token corpus training, 24 Mamba2 blocks, d=768) reads layer-median retention horizons of 0.28–406 tokens ($1.5\times10^{3}\times$ across blocks, within-block span up to 1.6×10^{6}), a slow-minority tail (P90/P50≥2×, up to 6.0×10^{2}) in 23/24 blocks, and no deep-block median ordering (0.54× top-over-bottom) — the heterogeneity-with-slow-minority observable replicates on an independently trained open model (source `evals/h2_m2hf_attractor.json`; summarized in Section 4.4).

Interpretation across controls. The soft grid, constant floor, static-rate and hard-commitment controls do not establish a unique optimum. Their quality, long-context penalties and training costs must be read jointly, with capped and full-shard evaluation kept separate. In particular, the grid's multiple coordinates are not independently stored content. The main text reports the primary all-seed comparisons; the subset analyses in this appendix are reported with their own evaluation coverage (seeds and evaluator).