TAYLOR-CALIBRATE: PRINCIPLED INITIALIZATION FOR HYBRID LINEAR ATTENTION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid linear attention reduces the quadratic cost and KV-cache growth of softmax attention, but converting a pretrained Transformer remains brittle. Projection transfer copies $Q/K/V/O$ weights yet leaves the student's new recurrent dynamics—memory decay, write strength, value scale, and output gating—unspecified. We propose Taylor-Calibrate, a lightweight teacher-conditioned initialization for these unmatched dynamics. An exact online form of softmax attention exposes a delta update; a Taylor and online-regression approximation then identifies the corresponding recurrent quantities. We estimate their operating point from teacher attention distance, concentration, and output scale, followed by brief layer-local alignment. Across four teachers and three retained-layer policies, Taylor-Calibrate produces stronger zero-shot students. On five single-run matched-PPL recovery comparisons spanning three teachers, conservative lower-bound Stage-2 token savings range from 4.9× to 9.2× relative to projection-copy initialization. Held-out representation and KL measurements confirm that the gain reflects closer teacher behavior, not perplexity alone. Matched three-seed experiments further transfer the framework to GLA and KDA. Separately, five-layer KDA conversion with a 64-GPU-sharded 975B Inkling teacher reaches PPL 6.498 versus 6.454 and 90.2% teacher top-1 agreement, validating scale integration. The complete 1.5B calibration costs 0.024 H100-hours.

1 INTRODUCTION

Transformers remain the default for large language models, but long-context decoding incurs quadratic attention and a growing KV cache (Vaswani et al., 2017). Recurrent and linear mixers instead compress history into fixed-size state, from early kernelized and fast-weight views to selective state spaces and gated linear attention (Katharopoulos et al., 2020; Schlag et al., 2021; Peng et al., 2022; Gu and Dao, 2024; Dao and Gu, 2024; Yang et al., 2024b; Zhang et al., 2024; Liu et al., 2024). Production hybrids including Jamba, Jamba-1.5, Samba, Hymba, B'Mojo, MiniMax-01, Falcon-H1, Kimi Linear, and Qwen3.5 show the resulting quality–efficiency trade-off at scale (Lieber et al., 2024; Jamba Team, 2024; Ren et al., 2024; Dong et al., 2024; Zancato et al., 2024; MiniMax Team, 2025; Falcon LLM Team, 2025; Kimi Team, 2025; Qwen Team, 2026). Because pure recurrence can still struggle with retrieval, practical hybrids retain selected softmax layers (Arora et al., 2024; Wen et al., 2025; Irie et al., 2025; Wang et al., 2025; Yang et al., 2025).

Conversion avoids training a hybrid from scratch through swap-then-finetune, attention transfer, hidden-state matching, low-rank adaptation, or staged distillation (Kasai et al., 2021; Mercat et al., 2024; Wang et al., 2024; Zhang et al., 2025; Lan et al., 2025; Nguyen et al., 2025; Goldstein et al., 2025; Bick et al., 2025; 2026); separate methods select which layers remain softmax (Hoshino et al., 2025; Li et al., 2025). Yet newly introduced recurrent dynamics often remain teacher-independent: Mamba in the Llama transfers identifiable projections but retains default state-space dynamics, while Attention to Mamba uses identity-like retention and gating (Wang et al., 2024; Moudgil et al., 2026). We ask instead: how should teacher behavior initialize recurrent parameters with no Transformer counterpart?

Memory horizon, write/erase strength, and output amplitude determine the student's initial dynamical regime. In Gated DeltaNet (GDN), copying $Q/K/V/O$ sets none of its decay, delta-rule write, or output gates (Yang et al., 2024a). The mismatch differs for Gated Linear Attention (GLA), which

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.

has no write gate, and Kimi Delta Attention (KDA), which combines channel-wise decay with delta updates (Yang et al., 2024b; Kimi Team, 2025). Defaults can therefore produce a poor student even when inherited projections are reasonable.

We propose **Taylor-Calibrate**, a two-stage initialization framework for these unmatched dynamics. First, the exact online recursion for softmax attention exposes a prediction-error update; a first-order Taylor and online-regression approximation yields the structure of a recurrent delta rule. Teacher attention distance, concentration, and output amplitude then set the operator's memory, update, and scale parameters. Second, a short layer-local objective aligns the residual approximation error before full-model distillation. The derivation determines *which* dynamical quantities require transfer; each operator requires its own native parameter mapping.

Figure 1 shows both sides of this claim. Under identical projection transfer, GDN initialization changes zero-shot PPL by orders of magnitude. On held-out inputs, Taylor-Calibrate also lowers layer-output MSE and teacher–student KL while increasing cosine similarity, directly showing closer teacher behavior rather than only a PPL effect.

We evaluate GDN, GLA, and KDA across four teacher settings and three retained-layer policies, with three-seed fidelity, sensitivity, and operator controls. A separate five-layer KDA study asks whether operator-native initialization remains stable with a 64-GPU-sharded 975B MoE teacher; a component profile measures calibration cost.

Our main contributions are:

- We formulate conversion as teacher-conditioned initialization of recurrent dynamics that projection transfer leaves unidentified.
- We connect exact softmax recursion to a recurrent delta update, state every approximation, and combine closed-form OLS/RMS calibration with brief local alignment.
- We report recovery across four teachers, direct fidelity and sensitivity controls, matched GDN/GLA/KDA evidence, five-layer KDA scale validation on a 975B MoE, and measured calibration overhead.

2 Preliminaries

2.1 Softmax Attention

Consider a decoder-only Transformer (Vaswani et al., 2017) with hidden states $x_1, \dots, x_T \in \mathbb{R}^d$. For one attention head with head dimension d_k, softmax attention computes

$$q_t = W_Q x_t, \qquad k_t = W_K x_t, \qquad v_t = W_V x_t, \tag{1}$$

and outputs

$$o_t = \sum_{s \le t} \alpha_{t,s} v_s, \qquad \alpha_{t,s} = \frac{\exp(q_t^\top k_s / \sqrt{d_k})}{\sum_{j \le t} \exp(q_t^\top k_j / \sqrt{d_k})}. \tag{2}$$

During autoregressive decoding, the hidden states of all previous keys and values must be retained, leading to a KV cache that grows linearly with sequence length. This cache is a central systems bottleneck for long-context inference (Wang et al., 2024; Kimi Team, 2025).

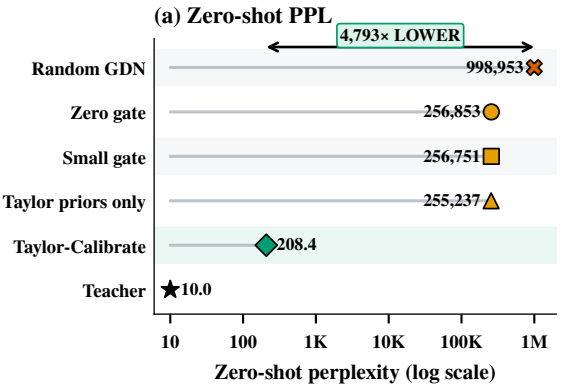
2.2 From Softmax Recursion to a Delta Update

Fix a query q and absorb the temperature $1/\sqrt{d_k}$ into the keys in this subsection. Define $Z_t = \sum_{i \le t} \exp(q^\top k_i)$ and $a_t = \exp(q^\top k_t)/Z_t$. Softmax attention has the exact online form

$$o_t = o_{t-1} + a_t\big(v_t - o_{t-1}\big). \tag{3}$$

Thus the exact increment is already "new value minus current prediction." Around uniform logits, a first-order expansion gives $a_t(q) \approx t^{-1}[1 + q^\top(k_t - \bar{k}_t)]$. Representing the output by finite statistics, $o_t \approx S_t q + \bar{v}_t$, and linearizing the remaining quadratic term yields

$$S_t \approx \left(1 - \frac{1}{t}\right) S_{t-1} - \frac{1}{t} S_{t-1} \tilde{k}_t \tilde{k}_t^\top + \frac{1}{t} \tilde{v}_t \tilde{k}_t^\top, \tag{4}$$



(b) Held-out teacher fidelity

Metric	Align-only	Taylor-Cal.	Gain
MSE ↓	.279±.009	.224±.006	20%
Cos. ↑	.323±.005	.454±.006	41%
KL ↓	.707±.075	.525±.011	26%
PPL ↓	23.15±.21	18.39±.20	21%

Mean ± s.d. · 3 seeds · same local-alignment budget

Figure 1: Initialization changes both external quality and teacher fidelity under matched $Q/K/V/O$ transfer. **Left:** on a log scale, zero-shot PPL varies by orders of magnitude with the new GDN dynamics. **Right:** after the same local alignment budget, Taylor-Calibrate improves converted-layer MSE/cosine, teacher–student KL, and 32K-token test PPL.

where $\tilde{k}_t = k_t - \bar{k}_t$ and $\tilde{v}_t = v_t - \bar{v}_{t-1}$. The resulting decay, erase, and write structure is approximate; Appendix A.2 gives the coefficient match, assumptions, and implementation gaps. It returns a uniform-history schedule whose per-head operating point Taylor-Calibrate sets from teacher statistics.

2.3 Gated DeltaNet (GDN) Linear Attention

Gated DeltaNet replaces softmax attention with a recurrent state update (Yang et al., 2024a). In the variant used here, queries and keys are projected from the same teacher-inherited hidden states and then normalized per head. Let $S_t \in \mathbb{R}^{d_k \times d_k}$ denote the recurrent memory state. For each step,

$$g_t = -\exp(A_{\log})\ \mathrm{softplus}(a_{\mathrm{proj}}(x_t) + dt_{\mathrm{bias}}), \qquad \beta_t = \sigma(b_{\mathrm{proj}}(x_t)), \tag{5}$$

$$S_t = \exp(g_t)\left(S_{t-1} - \beta_t S_{t-1} k_t k_t^\top\right) + \beta_t v_t k_t^\top, \qquad o_t = S_t q_t. \tag{6}$$

The final output is further modulated by an output gate,

$$y_t = W_O(\mathrm{RMSNorm}(o_t) \odot \mathrm{SiLU}(g_{\mathrm{proj}}(x_t)))\,. \tag{7}$$

Relative to softmax attention, GDN removes the explicit KV cache and instead compresses context into a fixed-size recurrent state (Yang et al., 2024a). This provides constant-memory inference with respect to context length, but places substantial burden on the decay and gating parameters. If these parameters are poorly initialized, the recurrent block may either erase useful history too aggressively or inject unstructured noise into the residual stream.

3 Method: Taylor-Calibrate

Taylor-Calibrate follows one principle: *teacher information should initialize recurrent dynamics that projection copying cannot inherit*. Phase 1 identifies those dynamics from the softmax-to-delta bridge in Equations 3–4 and sets their operating point from teacher statistics. Phase 2 aligns the residual layer-output error on the same calibration inputs before global distillation. We first instantiate the framework for GDN (Yang et al., 2024a); Section 3.4 gives the operator-native GLA and KDA mappings.

3.1 Inherited Parameters and Unmatched Dynamics

Let M_T be a pretrained Transformer teacher (Vaswani et al., 2017) and M_S a hybrid student that replaces a chosen layer set $\mathcal{S}_{\mathrm{rec}}$ with a recurrent operator. Projection copying transfers W_Q, W_K, W_V, W_O when shapes agree, but cannot transfer decay, write/erase, or output-gate parameters because the teacher has no such tensors. For each converted layer, we therefore minimize the local mismatch

$$\mathcal{L}_{\mathrm{init}} = \sum_{\ell \in \mathcal{S}_{\mathrm{rec}}} \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{cal}}} \left\| z_S^{(\ell)}(x; \theta_\ell) - z_T^{(\ell)}(x) \right\|_2^2, \tag{8}$$

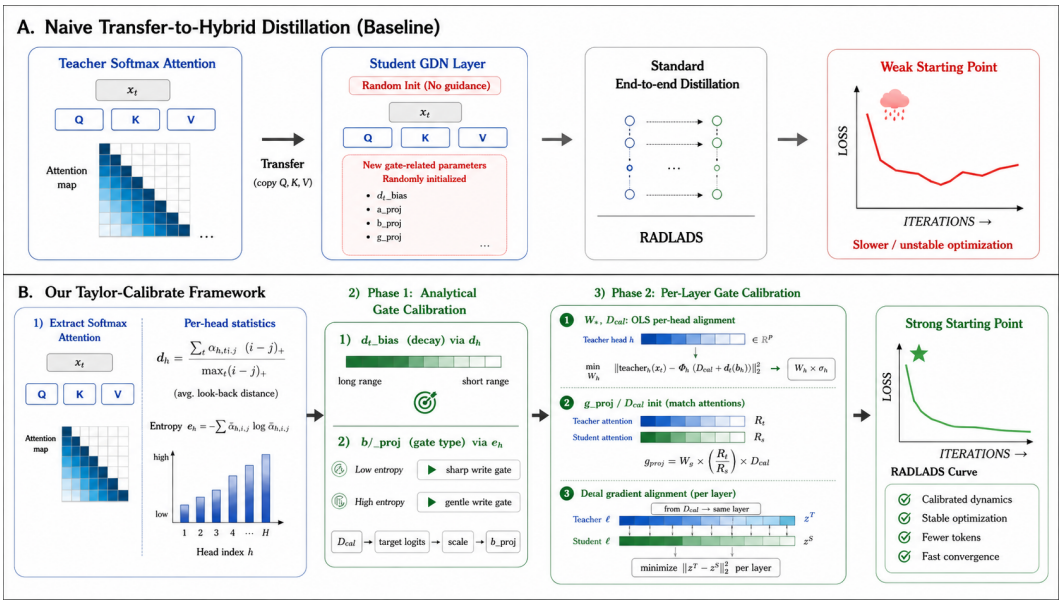


Figure 2: Taylor-Calibrate copies teacher projections, initializes unmatched GDN dynamics from teacher statistics, and locally aligns each converted layer before global distillation; standard transfer leaves those dynamics at defaults.

Table 1: Analytical Phase-1 transformations for GDN conversion. Phase 2 subsequently updates both transferred and recurrent parameters.

Parameter/statistic	Standard GDN transfer	Taylor-Calibrate initialization
$d_h, e_h, c_h, \sigma_h^{\text{impl}}$	Not used.	Teacher statistics and a clipped pre-normalization scalar fit.
W_Q, W_K, W_O	Copy from teacher.	Copy from teacher.
W_V / value path	Copy from teacher.	$W_V \leftarrow \sigma_h^{\text{impl}} W_V$ (pre-RMSNorm OLS proxy).
$A_{\log}$	Default decay scale.	Set to zero so dt_{bias} controls the initial horizon.
dt_{bias}	Default time bias.	$dt_{\text{bias},h} = \text{softplus}^{-1}(\ln 2 / d_h)$.
a_{proj}	Default decay projection.	Default decay projection.
b_{proj}	Default write projection.	Bounded concentration-conditioned row temperature/orientation.
g_{proj}	Default or small-gate heuristic.	Small/RMS-matched gate.

on a small calibration corpus $\mathcal{D}_{\text{cal}}$. This is an initialization objective, not end-to-end distillation.

For GDN, $\mathcal{S}_{\text{rec}} = \mathcal{S}_{\text{gdn}}$ and the unmatched quantities are its memory timescale, delta-rule write behavior, pre-normalization value/output scale, and output gate. Figure 1 shows that these choices change both external quality and teacher fidelity under identical projection transfer.

3.2 Phase 1: Taylor-Derived Calibration Adapted to GDN

The derivation in Appendix A.2 determines the recurrence structure but leaves its non-uniform operating point unidentified. We choose four Phase-1 proxies per converted layer: attention distance sets a memory half-life, concentration sets a bounded write-logit temperature/orientation, the teacher context defines a pre-normalization scalar fit, and teacher output scale initializes the output gate.

For teacher attention probability $a_{t,s}^{(h)}$, we compute per-head average look-back $d_h = \mathbb{E}[\sum_{s \le t} a_{t,s}^{(h)} \max(t-s, 0.5)]$, floored at 0.5, and entropy $e_h = \mathbb{E}[-\sum_{s \le t} a_{t,s}^{(h)} \log a_{t,s}^{(h)}]$. Distance summarizes memory horizon; entropy summarizes concentration.

We map average attention distance to an exponential-memory half-life by design. Requiring a head to retain half its state after d_h tokens gives $|g_h| = \ln 2/d_h$ (Appendix A.3). At initialization we set $A_{\log} = 0$ and calibrate the bias-only idle rate; the unchanged $a_{\text{proj}}(x_t)$ subsequently makes realized decay input-dependent:

$$dt_{\text{bias},h} := \text{softplus}^{-1}\left(\frac{\ln 2}{d_h}\right). \tag{9}$$

For write-logit scaling, let $\Delta_e = \max(e_{\max} - e_{\min}, 10^{-3})$, $c_h = 1 - (e_h - e_{\min})/\Delta_e$, and $\ell_h = \text{logit}(0.3 + 0.4c_h)$. For default row $\tilde{b}_h$, $m_h = \sqrt{d_{\text{in}}}\,\text{mean}\,|\tilde{b}_h|$ approximates preactivation RMS, giving

$$(b_{\text{proj}})_{h,:} \leftarrow \frac{\ell_h}{m_h}\,\tilde{b}_h. \tag{10}$$

This controls temperature/orientation, not mean gate value: symmetric zero-centered preactivations give $\mathbb{E}[\sigma(b_{\text{proj}}x)] = 1/2$ at every scale (Appendix A.5).

Finally, let $\tilde{o}_S^{(h)}$ be the student's pre-RMSNorm recurrent readout and $y_T^{(h)}$ the teacher context. Per-head OLS gives $\sigma_h^\star = \langle y_T^{(h)}, \tilde{o}_S^{(h)}\rangle / \|\tilde{o}_S^{(h)}\|_2^2$, clipped in implementation to $\sigma_h^{\text{impl}} \in [0.1, 10]$ (Björck, 1996). Let r_T be mean per-token/head teacher vector RMS and r_V the global RMS of $\text{SiLU}(W_V x)$; the gate proxy is $\alpha_\ell^\star = \lambda r_T / r_V$ with $\lambda = 0.01$ (Appendices A.6–A.7):

$$W_V^{(h)} \leftarrow \sigma_h^{\text{impl}} W_V^{(h)}, \qquad g_{\text{proj}}^{(\ell)} \leftarrow \alpha_\ell^\star W_V^{(\ell)}. \tag{11}$$

The OLS coefficient is exact for a linear pre-normalization proxy and approximate with SiLU. In the linear, $\epsilon = 0$ limit, its scale is canceled both by RMSNorm and by recomputing $\alpha_\ell^\star$ after rescaling W_V; any practical effect comes from clipping, SiLU nonhomogeneity, RMSNorm ϵ, and numerical conditioning. We therefore make no direct amplitude claim; Phase 2 fits the actual nonlinear block.

3.3 Phase 2: Per-Layer Gradient Alignment

Analytical calibration sets scales and operating points, but low-order statistics cannot recover exact parameter directions. We therefore perform brief *layer-local* alignment (Hinton et al., 2015; Zhang et al., 2025; Goldstein et al., 2025). For each converted layer ℓ, we minimize

$$\mathcal{L}_{\text{align}}^{(\ell)} := \mathbb{E}_{x\sim\mathcal{D}_{\text{cal}}} \left\| \hat{z}_{\text{GDN}}^{(\ell)}(x;\theta_\ell) - z_T^{(\ell)}(x) \right\|_2^2, \tag{12}$$

where $\hat{z}_{\text{GDN}}^{(\ell)}$ is the student GDN output at the same hidden-state input. We update both transferred projections and newly introduced recurrent parameters:

$$\theta_\ell = \{W_Q, W_K, W_V, W_O, A_{\log}, dt_{\text{bias}}, a_{\text{proj}}, b_{\text{proj}}, g_{\text{proj}}\}. \tag{13}$$

The two phases address different errors: Phase 1 selects a teacher-informed dynamical regime; Phase 2 adjusts parameter directions and residual nonlinear mismatch. Gate statistics in Appendix D.10 show that alignment largely preserves the calibrated half-life while reshaping write/output projections.

3.4 Operator-Native Transfer Beyond GDN

The framework transfers, but GDN's formulas do not. GLA (Yang et al., 2024b) uses input-dependent channel decay and an output gate, with no delta-rule write/erase gate; we therefore map teacher distance into its native decay bias and apply OLS/RMS output calibration.

KDA (Kimi Team, 2025) maintains an associative state matrix with channel-wise retention α_t and an explicit delta-rule write rate β_t. We use the value-first state $S_t \in \mathbb{R}^{d_v \times d_k}$, the transpose of Kimi Linear's key-first convention:

$$S_t = S_{t-1}\,\text{Diag}(\alpha_t)\big(I - \beta_t k_t k_t^\top\big) + \beta_t v_t k_t^\top, \qquad o_t = S_t q_t. \tag{14}$$

The first term decays old state and erases the association aligned with the current key; the second writes the new value for that key. Repeated keys therefore overwrite stale associations rather than accumulating them without correction. We map teacher attention distance to α_t, concentration to KDA's native write-logit projection, and output scale to its sigmoid gate before applying the same local-alignment objective. Appendix A.8 gives the complete native map; Section 4.5 reports matched GLA/KDA controls.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate four teachers: **Qwen2.5-1.5B/3B-Instruct**, **Llama-3.2-3B-Instruct**, and **Qwen3-8B** (Qwen Team, 2024; Meta Llama Team, 2024; Qwen Team, 2025). Students share the RADLADS $Q/K/V/O$ transfer and downstream recipe (Goldstein et al., 2025); only unmatched GDN initialization changes: **Baseline** uses defaults, **Zero-Gate** sets $g_{\text{proj}} = 0$, **Small-Gate** sets it to 0.01, **Taylor-Only** applies Phase 1, and **Taylor-Calibrate** applies both phases. Retained softmax layers follow **Uniform**, **AR (LM-PPL)**, or KL-guided **GA-S2** selection (Li et al., 2025).

After zero-shot evaluation, every student receives identical hidden-state alignment and KL distillation (Hinton et al., 2015; Goldstein et al., 2025; Li et al., 2025). We report **PPL**, a twelve-task short-context macro average (**Avg**), and aggregate **RULER** (Hsieh et al., 2024); Appendices D.1 and D.4 give task-level results. Focused 7-of-28-layer controls use three seeds. "32K PPL" is a 32K-token test stream, not context length.

Data splits and zero-leakage protocol. Calibration uses 1,024 tokens; downstream NLU and RULER are disjoint, test PPL excludes that prefix, and the Inkling calibration and evaluation slices do not overlap (Appendix D.6).

Measured calibration overhead. Calibrating 21 of 28 Qwen2.5-1.5B layers takes 86.821 s (0.0241 H100-hours) and 7.81 GiB on one H100-80GB (Appendix D.3). Recovery uses 99.98M+599.92M tokens on eight GPUs; missing wall-clock telemetry precludes an end-to-end speedup claim.

4.2 ZERO-SHOT EVALUATION

Before training, Table 2 compares Uniform-policy Avg/RULER at 25%, 50%, and 75% retained attention under the same projection-transfer setting as prior conversion work (Wang et al., 2024; Goldstein et al., 2025; Li et al., 2025); Appendix D.1 gives task-level and non-uniform results.

At 25%–50%, all methods retain little RULER capability and several Baselines approach chance Avg. At 75%, output-gate stabilization alone recovers substantial Avg and often RULER, confirming the value of suppressing an uncalibrated recurrent branch. Full Taylor-Calibrate most consistently improves Avg, especially at 50%–75%; RULER gains remain model- and budget-dependent, indicating that local alignment transfers short-context behavior more readily than long-context retrieval.

Table 2: Uniform zero-shot **Avg/RULER**; task-level and other-policy results are in Appendix D.1. **Bold**: best non-teacher value per row.

Model	Budget	Teacher	Baseline	Zero-Gate	Small-Gate	Taylor-Only	Taylor-Calibrate
Qwen2.5-1.5B	25%	64.8 / 86.2	31.2 / **0.2**	31.3 / **0.2**	31.4 / **0.2**	31.9 / **0.2**	**34.8 / 0.2**
	50%	64.8 / 86.2	31.7 / **0.1**	32.7 / **0.1**	32.7 / **0.1**	32.6 / **0.1**	**37.4 / 0.1**
	75%	64.8 / 86.2	48.6 / 2.7	53.9 / 10.1	53.7 / 10.3	53.9 / 10.1	**56.2 / 14.5**
Qwen2.5-3B	25%	67.3 / 91.3	31.1 / 0.0	**31.8** / 0.0	**31.8** / 0.0	31.7 / 0.0	30.9 / **0.1**
	50%	67.3 / 91.3	31.3 / 0.0	31.3 / 0.0	31.4 / 0.0	31.3 / 0.0	**36.3 / 0.4**
	75%	67.3 / 91.3	55.1 / 19.9	61.8 / 47.4	**62.0** / 47.2	61.9 / 47.2	61.3 / **55.9**
Llama-3.2-3B	25%	65.6 / 89.6	31.2 / 0.0	33.1 / 0.0	33.2 / **0.1**	33.1 / **0.1**	**34.2 / 0.1**
	50%	65.6 / 89.6	31.7 / 0.1	36.9 / 0.2	36.7 / 0.2	36.8 / 0.2	**43.3 / 0.4**
	75%	65.6 / 89.6	33.9 / 0.2	58.3 / **31.0**	58.6 / 30.9	58.5 / 30.9	**60.0** / 24.0
Qwen3-8B	25%	70.7 / 94.0	31.4 / **0.1**	34.2 / **0.1**	34.1 / 0.0	34.2 / **0.1**	**35.6 / 0.1**
	50%	70.7 / 94.0	36.2 / 0.2	45.1 / 0.3	44.9 / 0.3	45.0 / 0.3	**49.0 / 0.7**
	75%	70.7 / 94.0	59.9 / 42.3	64.8 / 61.6	64.7 / 61.4	64.8 / 61.6	**65.5 / 65.1**

4.3 RECOVERY SPEED AND TOKEN EFFICIENCY

The zero-shot comparison establishes that Taylor-Calibrate provides a better starting point. The next question is whether this advantage translates into faster recovery during downstream training: at which token checkpoint does the converted student reach teacher-like quality, and how much does better initialization accelerate restoration?

Table 3 summarizes checkpoint recovery for all teacher settings, comparing Baseline and Taylor-Calibrate across retained-layer policies with Avg and RULER. Figure 3 then shows training dash-

Table 3: Compact recovery summary for all teacher settings. Each entry reports **Avg/RULER**; detailed task-level metrics are reported in Appendix D.1. **Bold**: better value between Baseline and Taylor-Calibrate within each checkpoint.

Model	Selection	Teacher	100M (Stage 1)		700M (Stage 2)	
			Baseline	Taylor-Calibrate	Baseline	Taylor-Calibrate
Qwen2.5-1.5B	Uniform	64.8 / 86.2	37.2 / 0.6	**62.6 / 7.7**	61.4 / 39.9	**64.4 / 59.8**
	AR (LM-PPL)	64.8 / 86.2	44.7 / 1.0	**61.1 / 5.0**	61.4 / 16.8	**63.6 / 38.0**
	GA-S2	64.8 / 86.2	43.6 / 0.5	**62.0 / 7.1**	60.3 / 27.0	**63.8 / 47.6**
Qwen2.5-3B	Uniform	67.3 / 91.3	60.5 / 26.3	**60.8 / 29.2**	65.8 / 63.8	**66.0 / 65.7**
	AR (LM-PPL)	67.3 / 91.3	**62.1** / 11.5	**62.1 / 11.7**	65.8 / 56.1	**66.1 / 57.9**
	GA-S2	67.3 / 91.3	52.6 / **6.6**	**60.4** / 4.9	65.7 / 59.8	**65.9 / 70.5**
Llama-3.2-3B	Uniform	65.6 / 89.6	59.7 / 5.7	**61.3 / 6.6**	64.3 / 60.0	**64.8 / 63.4**
	AR (LM-PPL)	65.6 / 89.6	58.5 / 4.6	**60.1 / 5.4**	63.8 / 46.4	**64.0 / 47.3**
	GA-S2	65.6 / 89.6	57.8 / **5.0**	**59.5** / 4.2	62.3 / 41.1	**63.9 / 46.5**
Qwen3-8B	Uniform	70.7 / 94.0	49.4 / 0.1	**66.5 / 9.2**	69.4 / 60.4	**70.0 / 76.7**
	AR (LM-PPL)	70.7 / 94.0	44.7 / 0.2	**67.6 / 15.4**	69.2 / 62.2	**70.1 / 69.8**
	GA-S2	70.7 / 94.0	39.8 / 0.1	**67.4 / 11.0**	68.9 / **66.1**	**70.1** / 57.2

boards for the main recovery runs, tracking PPL and optimization losses over the consumed token budget. Detailed task-level metrics are reported in Appendix D, especially Appendix D.1.

Table 3 shows that the largest benefit of Taylor-Calibrate appears early in recovery, before Stage 2 has had enough tokens to repair a poor initialization. At 100M tokens, the gains are especially large for the harder settings: Qwen2.5-1.5B improves from 37.2 to 62.6 Avg under Uniform selection, and Qwen3-8B improves from 39.8–49.4 to 66.5–67.4 Avg across the three retained-layer policies. Llama-3.2-3B has a smaller Avg gap because the baseline already recovers reasonably after Stage 1, but Taylor-Calibrate still gives consistent short-context gains.

By 700M tokens, the Avg gap narrows because distillation has largely restored short-context behavior for both initializations. Taylor-Calibrate still gives the best Avg in every 700M row and often improves RULER as well, although long-context retrieval remains below the teacher. This suggests that better initialization mainly reduces the amount of training needed to approach teacher-level short-context quality, while full long-context restoration is a slower process.

The PPL trajectories in Figure 3 decrease smoothly for both teachers, with no evidence that the gap comes from a pathological baseline run. Taylor-Calibrate starts closer to the teacher and reaches the near-teacher region with fewer tokens; together with Table 3, this indicates that initialization quality mainly changes recovery speed under the fixed objective. Matched-PPL accounting in Figure 5 converts this gap into a 4.9×–9.2× conservative lower bound on Stage-2 token savings over five single-run curves; the shared Stage-1 budget is excluded.

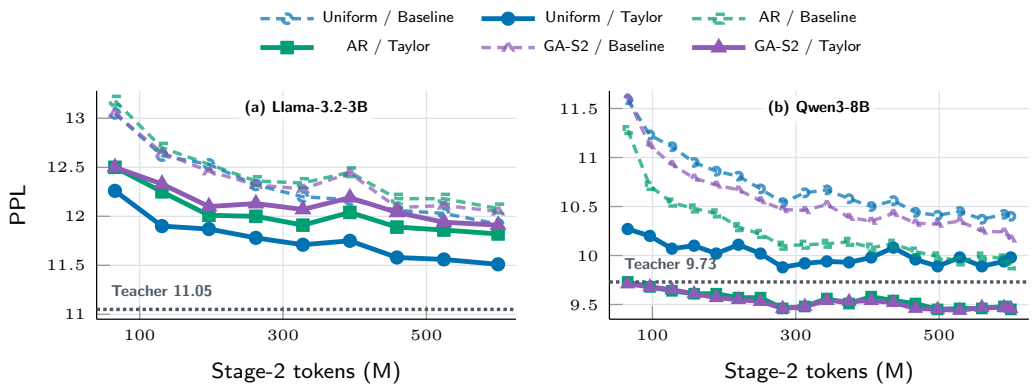


Figure 3: Stage-2 PPL recovery under the three layer-selection policies. Taylor-Calibrate starts and ends below its matched projection-copy baseline in both teacher settings; full loss dashboards are provided with the artifacts.

4.4 Long-Context Restoration

Short-context quality recovers quickly; we next ask whether long-context retrieval follows the same trajectory. RULER (Hsieh et al., 2024) recovers more slowly than PPL and short-context Avg. At 700M tokens, Llama-3.2-3B-Instruct nearly matches the teacher on Avg (64.8 vs. 65.6), but still trails on RULER (63.4 vs. 89.6). Qwen3-8B shows the same pattern: the best reported Stage-2 RULER score is 76.7, still below the teacher at 94.0. Thus 700M tokens can recover much of the short-context behavior without fully restoring long-context retrieval.

The aggregate also hides a strong interaction with layer selection. On Qwen3-8B, Uniform improves the eight NIAH/five non-NIAH means from 69.0/46.8 to 82.2/67.9, while GA-S2 changes them from 72.5/55.9 to 70.5/35.9. The discrepancy is therefore concentrated in non-NIAH tasks and depends on which softmax layers are retained; Appendix D.4 reports all three policies.

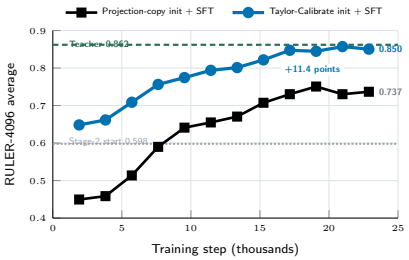


Figure 4: Extended 1.5B RULER recovery: Taylor reaches 0.850, 1.2 points below the teacher and 11.4 above baseline.

Figure 4 shows the longer-horizon behavior for the 1.5B setting. We continue training on SFTMix v0.2 for more than 20k optimization steps, rather than stopping at the short recovery schedule in Table 3. RULER reaches the teacher range in this run, showing that the Stage-2 gap can close with additional recovery training. Separately, Appendix D.9 shows that Taylor-Calibrate improves teacher KL and PPL at every tested length from 128 to 1,024 tokens; this is a local-fidelity control, not a substitute for RULER.

The conclusion is that initialization transfers local teacher behavior immediately, whereas long-range task recovery is slower and sensitive to replacement location.

4.5 Architecture and Mechanism Transfer

Table 4: Architecture transfer. (a) Three-seed Qwen2.5-1.5B operator transfer (7/28 layers). (b) Matched Qwen3-8B 3:1 hybrids (same nine retained layers and 100M-token stream).

(a) Initialization	GLA KL↓	GLA Avg↑	KDA KL↓	KDA Avg↑
Alignment only	.7595±.0506	60.83±.26	.6435±.0159	61.92±.42
+ stabilization	.6135±.0068	**62.54±.18**	**.5786±.0094**	**65.83±.09**
+ native decay/map	**.5726±.0087**	62.01±.19	.5870±.0043	65.77±.23

(b) Metric	SWA(64,4)	GDN
Parameters	8.19B	9.33B
6-task Avg↑	**69.68**	69.25
PPL↓	15.88	**11.02**
4K RULER↑	4.71	**32.80**

Having established the GDN recovery pattern, Table 4(a) tests whether initialization principles transfer across recurrent operators. We find that the effect of teacher priors reflects each architecture's inductive bias. In GLA, which lacks an associative delta-rule write/erase gate, recurrence dynamics depend entirely on channel decay; initializing decay from teacher attention distance sharply improves teacher KL from .7595 to .5726 (a 24.6% reduction). In KDA, which parameterizes both channel decay and a delta update, output stabilization delivers the primary gain, lowering KL from .6435 to .5786 and raising 3-task Avg from 61.92 to 65.83; its delta rule already guards against memory runaway, so the additional distance/concentration map is neutral (.5786 vs. .5870). This matched three-seed comparison provides the controlled evidence for KDA effectiveness; Appendix D.5 reports full statistics.

SWA comparison: matching short-task accuracy while beating SWA on PPL and RULER. Panel (b) compares 3:1 GDN and SWA(64,4) hybrids (Jolicoeur-Martineau et al., 2026) with the same nine retained layers, data, seed, optimizer, and 50M+50M budget. SWA's 0.43-point short-task advantage (69.68 vs. 69.25) is below the 0.57 standard error of difference, so there is no resolvable drop. Meanwhile, GDN lowers PPL by 30.6% (15.88→11.02) and raises 4K RULER by 6.96× (4.71→32.80). Since GDN has 1.14B more parameters and speed was not measured, this is a quality control, not an efficiency claim.

4.6 Scaling Five-Layer KDA Initialization to Inkling-975B

Inkling is a 975B-parameter, 66-layer MoE with 41B active parameters (Thinking Machines Lab, 2026). We replace five attention layers with KDA and run the operator-native pipeline against one 64-GPU FSDP/EP-sharded teacher: distance initializes retention, concentration initializes delta writing, and output statistics stabilize the value/gate path. On a disjoint sequence-128 slice, 200 local-alignment steps yield PPL 7.030 and 85.3% teacher top-1 agreement; after 600 end-to-end KL steps, these improve to 6.498 and 90.2%, versus teacher PPL 6.454.

The 0.7% final PPL gap shows that KDA-native initialization and alignment remain numerically stable at 975B scale. This run has no matched default-initialization control, so it supports scalability rather than a standalone causal claim; the matched KDA result in Table 4(a) supplies that evidence. Appendix D.6 gives the distributed protocol and split details.

4.7 Component Ablation and Mechanistic Dissection

To dissect the contribution of each initialization phase, we ablate both stages on Qwen2.5-3B-Instruct under one matched hybrid budget (Qwen Team, 2024).

Table 5: Component ablation on Qwen2.5-3B (Uniform); Avg is short-context NLU.

Variant	PPL↓	Avg	Description
Baseline	37337.3	30.9	projection copy only
Zero-Gate	22469.0	30.7	output-path stabilization
Taylor-Only	22470.6	30.9	Phase 1 only
Alignment-Only	2015.9	31.4	Phase 2 only, no analytical calibration
Taylor-Calibrate	**424.1**	**32.1**	Phase 1 + Phase 2

Phase 2 is essential: PPL falls from 37337.3 (Baseline) to 2015.9 (Alignment-Only) and 424.1 (full Taylor-Calibrate). Phase 1 provides a structured, teacher-conditioned operating point that improves subsequent local alignment rather than replacing it.

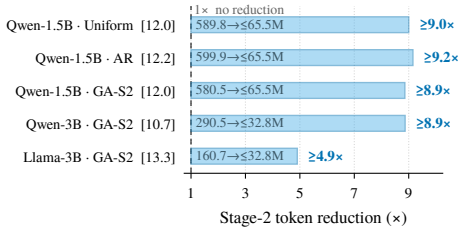


Figure 5: Matched-PPL Stage-2 savings (Appendix D.2).

For Figure 5, each projection-copy crossing is exact or linearly interpolated, while Taylor-Calibrate has already crossed by its first measured Stage-2 checkpoint. Thus the savings are lower bounds, not wall-clock claims. In one focused three-seed control, half-life variation ($0.5\times$–$2\times$) and constant write ($\beta = 0.5$) span 18.22–18.44 PPL versus 18.39 by default. This shows endpoint robustness, but not an independent gain from the distance/concentration maps (Appendices D.8 and D.10).

5 Conclusion

Taylor-Calibrate initializes unmatched recurrent dynamics from teacher statistics and local alignment. Across four teachers it improves initialization; five single-run curves spanning three teachers give conservative $4.9\times$–$9.2\times$ Stage-2 savings. Matched GLA/KDA controls validate operator transfer, while a five-layer Inkling run verifies KDA-native initialization with a 64-GPU-sharded 975B teacher.

AI USE STATEMENT

In this work, generative AI tools were used to assist with literature search, manuscript editing, figure and table formatting, research-code implementation, experiment orchestration, and consistency checks. All reported experimental measurements originate from executed runs rather than model-generated content. The authors reviewed the AI-assisted text and code and take responsibility for the final content, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The method assumptions and derivations are given in Sections 2–3 and Appendix A.2. Full protocols, task-level results, resource measurements, and figure provenance are reported in the appendices. Machine-readable plotting inputs and generation scripts accompany the supplementary source.

REFERENCES

Simran Arora, Aman Timalsina, Aaryan Singhal, Benjamin Spector, Sabri Eyuboglu, Xinyi Zhao, Ashish Rao, Atri Rudra, and Christopher Ré. Just read twice: Closing the recall gap for recurrent language models. *arXiv preprint arXiv:2407.05483*, 2024.

Aviv Bick, Tobias Katsch, Nimit Sohoni, Arjun Desai, and Albert Gu. Llamba: Scaling distilled recurrent models for efficient language processing. *arXiv preprint arXiv:2502.14458*, 2025.

Aviv Bick, Eric P. Xing, and Albert Gu. Retrieval-aware distillation for transformer-ssm hybrids. *arXiv preprint arXiv:2602.11374*, 2026.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Åke Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

Xin Dong et al. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*, 2024.

Falcon LLM Team. Falcon-h1: A family of hybrid-head language models redefining efficiency and performance. *arXiv preprint arXiv:2507.22448*, 2025.

Daniel Goldstein, Eric Alcaide, Janna Lu, and Eugene Cheah. Radlads: Rapid attention distillation to linear attention decoders at scale. *arXiv preprint arXiv:2505.03005*, 2025.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yuichiro Hoshino, Hideyuki Tachibana, Muneyoshi Inahara, and Hiroto Takegawa. Rad: Redundancy-aware distillation for hybrid models via self-speculative decoding. *arXiv preprint arXiv:2505.22135*, 2025.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Kazuki Irie, Morris Yau, and Samuel J. Gershman. Blending complementary memory systems in hybrid quadratic-linear transformers. *arXiv preprint arXiv:2506.00744*, 2025.

Jamba Team. Jamba-1.5: Hybrid transformer-mamba models at scale. *arXiv preprint arXiv:2408.12570*, 2024.

Alexia Jolicoeur-Martineau, Rhea Sanjay Sukthanker, Pashmina Cameron, and Emy Gervais. Sliding-window beats linear attention. *arXiv preprint arXiv:2608.28444*, 2026.

Jungo Kasai, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A. Smith. Finetuning pretrained transformers into RNNs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Kimi Team. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Disen Lan, Weigao Sun, Jiaxi Hu, Jusen Du, and Yu Cheng. Liger: Linearizing large language models to gated recurrent structures. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yanhong Li, Songlin Yang, Shawn Tan, Mayank Mishra, Rameswar Panda, Jiawei Zhou, and Yoon Kim. Distilling to hybrid attention models via kl-guided layer selection. *arXiv preprint arXiv:2512.20569*, 2025.

Opher Lieber et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Bo Liu, Rui Wang, Lemeng Wu, Yihao Feng, Peter Stone, and Qiang Liu. Longhorn: State space models are amortized online learners. *arXiv preprint arXiv:2407.14207*, 2024.

Jean Mercat, Igor Vasiljevic, Sedrick Keh, Kushal Arora, Achal Dave, Adrien Gaidon, and Thomas Kollar. Linearizing large language models. *arXiv preprint arXiv:2405.06640*, 2024.

Meta Llama Team. Llama-3.2-3b-instruct model card. `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

MiniMax Team. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

Abhinav Moudgil, Ningyuan Huang, Eeshan Gunesh Dhekane, Pau Rodríguez, Aswathy Balagopalan, Luca Zappella, and Federico Danieli. Attention to mamba: A recipe for cross-architecture distillation. *arXiv preprint arXiv:2604.14191*, 2026.

Chien Van Nguyen, Huy Nguyen, Ruiyi Zhang, Hanieh Deilamsalehy, Puneet Mathur, Viet Lai, Haoliang Wang, Jayakumar Subramanian, Ryan A. Rossi, Trung Bui, Nikos Vlassis, Franck Dernoncourt, and Thien Huu Nguyen. Lizard: An efficient linearization framework for large language models. *arXiv preprint arXiv:2507.09025*, 2025.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

Hao Peng, Jungo Kasai, Nikolaos Pappas, Dani Yogatama, Zhaofeng Wu, Lingpeng Kong, Roy Schwartz, and Noah A. Smith. ABC: Attention with bounded-memory control. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.5-27b. https://huggingface.co/Qwen/Qwen3.5-27B, 2026. Hugging Face model card. Highlights an efficient hybrid architecture built around Gated DeltaNet. Accessed 2026-03-25.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *arXiv preprint arXiv:2406.07522*, 2024.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 2011.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 2021.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Thinking Machines Lab. Inkling model card. https://huggingface.co/thinkingmachines/Inkling, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Dustin Wang, Rui-Jie Zhu, Steven Abreu, Yong Shan, Taylor Kergan, Yuqi Pan, Yuhong Chou, Zheng Li, Ge Zhang, Wenhao Huang, and Jason Eshraghian. A systematic analysis of hybrid linear attention. *arXiv preprint arXiv:2507.06457*, 2025.

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M. Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. *arXiv preprint arXiv:2408.15237*, 2024.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 2017.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. RNNs are not transformers (yet): The key bottleneck on in-context retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mingyu Yang, Mehdi Rezagholizadeh, Guihong Li, Vikram Appia, and Emad Barsoum. Zebra-llama: Towards extremely efficient hybrid models. *arXiv preprint arXiv:2505.17272*, 2025.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Luca Zancato et al. B'mojo: Hybrid state space realizations of foundation models with eidetic and fading memory. *arXiv preprint arXiv:2407.06324*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. *arXiv preprint arXiv:2402.04347*, 2024.

Michael Zhang et al. Lolcats: On low-rank linearizing of large language models. *arXiv preprint arXiv:2410.10254*, 2025.

Zhongzhu Zhou, Fengxiang Bie, Ziyan Chen, Zhenyu Zhang, Yibo Yang, Junxiong Wang, Ben Athiwaratkun, Xiaoxia Wu, and Shuaiwen Leon Song. CARE: Covariance-aware and rank-enhanced decomposition for enabling multi-head latent attention. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DVurf4kGag.

A Additional Preliminaries and Derivations

A.1 Hybrid Attention Architectures

We study *hybrid* models that keep a subset of layers as full softmax attention and replace the remainder with GDN (Lieber et al., 2024; Jamba Team, 2024; Yang et al., 2024a). Let

$$\mathcal{S}_{\text{soft}} \cup \mathcal{S}_{\text{gdn}} = \{1, \dots, L\}, \qquad \mathcal{S}_{\text{soft}} \cap \mathcal{S}_{\text{gdn}} = \varnothing. \tag{15}$$

Layer selection is an important problem in its own right (Li et al., 2025; Hoshino et al., 2025), but it is orthogonal to our question. We assume that a set of layers has already been chosen for replacement, and ask: *how should the newly introduced GDN layers be initialized so that distillation starts from calibrated decay, write, and output-gate scales?*

A.2 From Exact Softmax Recursion to Gated Delta Form

We sharpen the Taylor motivation by separating an exact identity from the approximations needed for finite-state compression. Fix a query q, suppress the head index, and absorb the softmax temperature into the keys by writing $k_i \leftarrow k_i/\sqrt{d_k}$. Define

$$Z_t(q) = \sum_{i \le t} \exp(q^\top k_i), \qquad a_t(q) = \frac{\exp(q^\top k_t)}{Z_t(q)}. \tag{16}$$

Because $Z_t = Z_{t-1} + \exp(q^\top k_t)$, softmax attention obeys

$$o_t(q) = \frac{Z_{t-1} o_{t-1}(q) + \exp(q^\top k_t) v_t}{Z_t} = o_{t-1}(q) + a_t(q)\big(v_t - o_{t-1}(q)\big). \tag{17}$$

Equation 17 is exact: the update is already a prediction-error (delta) correction, but its state remains query-dependent.

More generally, let $a_{t,i}(q) = \exp(q^\top k_i)/Z_t(q)$. The identity

$$a_{t,i}(q) = a_{i,i}(q) \prod_{j=i+1}^{t} (1 - a_{j,j}(q)) \tag{18}$$

follows by telescoping $\prod_{j=i+1}^{t} Z_{j-1}/Z_j = Z_i/Z_t$. It motivates expanding the diagonal update weights rather than independently linearizing every entry of the full attention matrix.

We next expand the diagonal softmax weight around equal logits. With $\bar{k}_t = t^{-1} \sum_{i \le t} k_i$ and $\tilde{k}_t = k_t - \bar{k}_t$,

$$a_t(q) = \frac{1}{t}\left(1 + q^\top \tilde{k}_t\right) + \mathcal{O}\Big(\max_i |q^\top (k_i - \bar{k}_t)|^2\Big). \tag{19}$$

Seek a fixed-size affine approximation $o_t(q) \approx S_t q + b_t$. Substitution into Equation 17 gives

$$\begin{aligned} o_t(q) \approx\ & \left(1 - \frac{1}{t}\right)(S_{t-1} q + b_{t-1}) + \frac{1}{t} v_t \\ & + \frac{1}{t}(q^\top \tilde{k}_t)\left(v_t - b_{t-1} - S_{t-1} q\right). \end{aligned} \tag{20}$$

The query-independent term therefore obeys

$$b_t = \left(1 - \frac{1}{t}\right) b_{t-1} + \frac{1}{t} v_t = \bar{v}_t. \tag{21}$$

The remaining expression is linear in q except for $-t^{-1}(q^\top \tilde{k}_t) S_{t-1} q$. Simply discarding this quadratic term yields a running linear-attention average with no state-dependent erase term. To retain that term with fixed state, replace its second occurrence of q by a representative $q^\star$. The resulting first-order replacement error is proportional to $(q^\top \tilde{k}_t) S_{t-1}(q - q^\star)$. We choose the direction where the first-order diagonal weight is largest: under a fixed query norm, $q^\top \tilde{k}_t$ is maximized when q is parallel to $\tilde{k}_t$. Matching the normalized query/key scale then gives $q^\star = \tilde{k}_t$. This gives

$$S_t \approx \left(1 - \frac{1}{t}\right) S_{t-1} - \frac{1}{t} S_{t-1} \tilde{k}_t \tilde{k}_t^\top + \frac{1}{t} \tilde{v}_t \tilde{k}_t^\top, \qquad \tilde{v}_t = v_t - \bar{v}_{t-1}. \tag{22}$$

Equation 22 matches the coefficients of the GDN-shaped update

$$S_t = \rho_t S_{t-1}\Big(I - \beta_t \tilde{k}_t \tilde{k}_t^\top\Big) + \beta_t v_t' \tilde{k}_t^\top \tag{23}$$

for $t > 1$ by setting

$$\rho_t = \frac{t-1}{t}, \qquad \beta_t = \frac{1}{t-1}, \qquad v_t' = \rho_t \tilde{v}_t. \tag{24}$$

Indeed, $\rho_t \beta_t = 1/t$ and $\beta_t v_t' = \tilde{v}_t/t$. Here β_t denotes the native gate inside the decayed update, while the effective erase coefficient is $\rho_t \beta_t = 1/t$; the two must not be conflated. This coefficient mismatch motivates the pre-normalization scalar proxy used by Taylor-Calibrate. For large t, the native $\beta_t = 1/(t-1) \approx 1/t$; the $t = 1$ state is initialized directly. Separately, the ungated delta correction $S \leftarrow S + \eta(v - Sk)k^\top$ is one gradient step on $\frac{1}{2}\|Sk - v\|_2^2$, which identifies the effective update coefficient as an online least-squares learning rate.

This is a structural coefficient match, not a literal equality to the implemented GDN block. Equation 22 uses centered keys and values and returns the affine output $S_t q + \bar{v}_t$; the implementation uses normalized inherited k_t, v_t, maintains no explicit running means, and outputs $S_t q_t$. We do not add that extra mean state. Teacher-conditioned moments choose the operating point, and layer-local alignment fits the discrepancy left by this implementation simplification.

The derivation fixes the *roles* of decay, write/erase, and value scale, but Equation 24 assumes uniform history. Teacher heads instead have finite, heterogeneous look-back and concentration. Taylor-Calibrate therefore replaces this schedule by teacher-measured per-head priors and uses local alignment for the residual error.

Approximations. The exact statement ends at Equation 17. The finite-state bridge uses six explicit approximations: (i) first-order expansion of the diagonal softmax weight; (ii) the affine ansatz $o_t(q) \approx S_t q + b_t$; (iii) replacement of the remaining query occurrence by one representative $q^\star$; (iv) the query/key normalization convention used to set $q^\star = \tilde{k}_t$; (v) reduction of a full teacher attention distribution to distance and concentration moments; and (vi) estimation from finite calibration data. The OLS coefficient is exact only for its pre-normalization scalar proxy; translating it into a W_V rescaling is exact only for a linear value path. The RMS gate match is likewise exact only for its stated one-dimensional proxy. Neither makes the full nonlinear recurrence exact.

A.3 Decay Calibration from Half-Life Matching

Our decay calibration uses a simple half-life argument to map teacher softmax-attention statistics (Vaswani et al., 2017) into the recurrent decay parameterization of GDN (Yang et al., 2024a). The implemented lag is $\max(t - s, 0.5)$, so self-attention contributes 0.5 rather than zero and the final average d_h is floored at 0.5. Let $\rho_h = \exp(g_h)$ denote the bias-only per-step multiplicative decay for head h. We choose the idle memory to decay by half after roughly d_h tokens:

$$\rho_h^{d_h} = \frac{1}{2}. \tag{25}$$

Solving for g_h gives

$$g_h = -\frac{\ln 2}{d_h}. \tag{26}$$

At initialization we set $A_{\log} = 0$. The following equation calibrates the idle value obtained when $a_{\mathrm{proj}}(x_t) = 0$:

$$g_h \approx -\,\mathrm{softplus}(dt_{\mathrm{bias},h}). \tag{27}$$

The ideal closed-form initialization is

$$dt_{\mathrm{bias},h} = \mathrm{softplus}^{-1}\left(\frac{\ln 2}{d_h}\right). \tag{28}$$

In code, the positive rate $\ln 2/d_h$ is clipped to $[10^{-3}, 5]$ before the inverse softplus to avoid numerical extremes. The default a_{proj} row is not zeroed, so token-dependent realized half-lives vary around this bias-only anchor.

A.4 Bounded Concentration-to-Logit Reference

We use normalized teacher concentration to define a bounded, nonsaturating reference logit for the bias-free GDN write projection (Yang et al., 2024a):

$$\beta_h^{\mathrm{ref}} = 0.3 + 0.4c_h, \tag{29}$$

where $\Delta_e = \max(e_{\max} - e_{\min}, 10^{-3})$ and $c_h = 1 - (e_h - e_{\min})/\Delta_e$ is the implemented concentration score. This reference parameterizes a row temperature and orientation; it is not asserted to equal the realized mean write gate.

The corresponding reference logit is

$$\ell_h := \mathrm{logit}(\beta_h^{\mathrm{ref}}) = \log \frac{0.3 + 0.4c_h}{0.7 - 0.4c_h}. \tag{30}$$

Because $c_h \in [0, 1]$, the implemented reference obeys

$$|\ell_h| \le \mathrm{logit}(0.7) \approx 0.847 \tag{31}$$

and hence the scalar reference remains nonsaturating:

$$\sigma'(\ell_h) = \sigma(\ell_h)\big(1 - \sigma(\ell_h)\big) \ge 0.21. \tag{32}$$

Median concentration gives $\ell_h = 0$. These bounds describe the scalar reference exactly, not the realized logits; the row-norm approximation is stated next.

A.5 Concentration-Conditioned Write-Logit Row Scaling

The previous subsection gives the reference logit

$$\ell_h = \mathrm{logit}(\beta_h^{\mathrm{ref}}). \tag{33}$$

In Phase 1 we do not replace row h of b_{proj} with a freshly optimized vector. Instead, we preserve the module-default direction and solve only for a scalar rescaling. Writing the default row as $\tilde{b}_h \in \mathbb{R}^{d_{\mathrm{in}}}$, we restrict the initialized row to

$$b_{\mathrm{proj}}[h, :] = s_h \tilde{b}_h. \tag{34}$$

To convert the reference logit into a row scale, we use a proxy for the row's preactivation RMS under isotropic unit-RMS inputs:

$$m_h := \sqrt{d_{\mathrm{in}}}\ \mathrm{mean}\Big(|\tilde{b}_h|\Big). \tag{35}$$

This proxy equals $\|\tilde{b}_h\|_2$ when all coordinates have the same magnitude and otherwise cheaply approximates the row's Euclidean scale. We match the absolute preactivation radius and choose the sign as an orientation convention:

$$s_h^\star := \arg \min_{\mathrm{sign}(s) = \mathrm{sign}(\ell_h)} \left(|s| m_h - |\ell_h|\right)^2. \tag{36}$$

For $\ell_h \neq 0$, the minimizer is

$$s_h^\star = \frac{\ell_h}{\sqrt{d_{\mathrm{in}}}\ \mathrm{mean}\Big(|\tilde{b}_h|\Big)}. \tag{37}$$

For $\ell_h = 0$ we set $s_h^\star = 0$. Replacing $\tilde{b}_h$ by the current module-default row gives the implemented update

$$b_{\mathrm{proj}}[h, :] \leftarrow \frac{\ell_h}{\sqrt{d_{\mathrm{in}}}\ \mathrm{mean}(|b_{\mathrm{proj}}[h, :]|)}\ b_{\mathrm{proj}}[h, :]. \tag{38}$$

Crucially, row scaling does not set the mean gate. If $Z_h = \tilde{b}_h^\top x$ is symmetric about zero, then $s_h Z_h$ is symmetric for every s_h and

$$\mathbb{E}[\sigma(s_h Z_h)] = \frac{1}{2}\mathbb{E}[\sigma(s_h Z_h) + \sigma(-s_h Z_h)] = \frac{1}{2}. \tag{39}$$

The construction controls a bounded logit temperature and chooses a signed row orientation; it does not realize $\mathbb{E}[\beta_t] = \beta_h^{\mathrm{ref}}$. Phase 2 learns the actual input-dependent gate correlations.

A.6 Closed-Form Value-Side OLS Rescaling

For a fixed head h, let y_T be the teacher pre-W_O context and y_S the student's pre-RMSNorm recurrent readout. Flatten the indices (b, t, d) into a single index i and write

$$u_i := y_{T,b,t,h,d}, \qquad v_i := y_{S,b,t,h,d}. \tag{40}$$

The value-side calibration solves

$$\sigma_h^\star = \arg\min_\sigma \sum_i (u_i - \sigma v_i)^2 . \tag{41}$$

Expanding the quadratic gives

$$J(\sigma) = \sum_i u_i^2 - 2\sigma \sum_i u_i v_i + \sigma^2 \sum_i v_i^2. \tag{42}$$

Differentiating and setting the derivative to zero yields

$$\frac{dJ}{d\sigma} = -2 \sum_i u_i v_i + 2\sigma \sum_i v_i^2 = 0, \tag{43}$$

so, provided the student context is not identically zero,

$$\sigma_h^\star = \frac{\sum_i u_i v_i}{\sum_i v_i^2} = \frac{\sum_{b,t,d} y_{T,b,t,h,d}\, y_{S,b,t,h,d}}{\sum_{b,t,d} y_{S,b,t,h,d}^2}. \tag{44}$$

This is the exact one-dimensional OLS solution for the abstract proxy σv (Björck, 1996). The implementation applies

$$\sigma_h^{\text{impl}} = \text{clip}(\sigma_h^\star, 0.1, 10) \tag{45}$$

and multiplies the corresponding rows of W_V. If the value path is linear, the recurrence is linear in W_V and this realizes $\sigma_h^{\text{impl}} v$ exactly. With a SiLU value activation it is only a proxy because $\text{SiLU}(\sigma z) \neq \sigma\, \text{SiLU}(z)$ in general.

Nor does this scalar directly determine the final output amplitude. For positive σ and per-head RMSNorm,

$$\text{RMSNorm}(\sigma o) = \frac{o}{\sqrt{d^{-1}\|o\|_2^2 + \epsilon/\sigma^2}} \odot w, \tag{46}$$

which is exactly scale-invariant when $\epsilon = 0$ and nearly so away from the numerical floor. The OLS step should therefore be read as a pre-normalization diagnostic/reparameterization, not direct amplitude calibration. Because the gate scale below is recomputed after rescaling W_V, the OLS scalar also cancels from the tied gate in the ideal linear regime. Any practical dependence arises from clipping, SiLU nonhomogeneity, RMSNorm ϵ, and numerical conditioning; Phase 2 optimizes the actual nonlinear block.

A.7 RMS-Matched Output-Gate Initialization

After the OLS step, the inherited value projection supplies the gate's feature direction. Because GDN further modulates the normalized recurrent output with an output gate (Yang et al., 2024a), we restrict the initial output-gate projection to the one-parameter family

$$W_g = \alpha_\ell W_V. \tag{47}$$

Let

$$\phi_t^{(\ell)} := \text{SiLU}(W_V x_t) \tag{48}$$

be the unscaled gate-feature prior. The implementation uses the mean per-vector teacher RMS

$$r_T^{(\ell)} := \mathbb{E}_{b,t,h} \left[\sqrt{\frac{1}{d} \sum_j \left(y_{T,b,t,h,j}^{(\ell)} \right)^2} \right] \tag{49}$$

and the global gate-feature RMS

$$r_V^{(\ell)} := \sqrt{\mathbb{E}_{b,t,h,j} \left[\left(\phi_{b,t,h,j}^{(\ell)} \right)^2 \right]}. \tag{50}$$

$$\alpha_\ell^\star := \arg\min_{\alpha \geq 0} \left(\alpha r_V^{(\ell)} - \lambda r_T^{(\ell)} \right)^2 . \tag{51}$$

Thus

$$\alpha_\ell^\star = \lambda \, \frac{r_T^{(\ell)}}{r_V^{(\ell)}}. \tag{52}$$

Because $r_V^{(\ell)}$ is recomputed after the OLS rescaling, a linear feature map would make r_V scale with σ_h^{impl} and $\alpha_\ell^\star$ inversely, canceling that scalar in $W_g = \alpha_\ell^\star W_V$. In the actual parameterization the gate is $\mathrm{SiLU}(W_g x) = \mathrm{SiLU}(\alpha_\ell W_V x)$, not $\alpha_\ell \, \mathrm{SiLU}(W_V x)$, so the objective is a proxy rather than an identity. It is accurate only when the inherited and scaled logits lie in a sufficiently local linear regime. The damping factor $\lambda = 0.01$ targets a low contribution in this surrogate but does not guarantee the actual gate RMS; Phase 2 fits the discrepancy.

A.8 Operator-Native Mappings for GLA and KDA

The softmax-to-delta bridge identifies generic dynamical roles, not architecture-independent tensors. GLA (Yang et al., 2024b) uses

$$S_t = S_{t-1} \, \mathrm{Diag}(\alpha_t) + v_t k_t^\top, \qquad o_t = S_t q_t, \tag{53}$$

followed by normalization and a Swish output gate. It has input-dependent channel decay but no delta-rule write/erase gate. We preserve the learned directions of its decay projection and shift only its final bias so that the calibration-set mean log-decay equals $-\ln 2/d_h$. Its output gate uses the same damped RMS proxy as GDN.

KDA (Kimi Team, 2025) combines channel-wise decay with a delta update. We use the value-first convention $S_t \in \mathbb{R}^{d_v \times d_k}$, the transpose of the paper's key-first state:

$$S_t = S_{t-1} \, \mathrm{Diag}(\alpha_t) \left(I - \beta_t k_t k_t^\top \right) + \beta_t v_t k_t^\top. \tag{54}$$

Here $\mathrm{Diag}(\alpha_t)$ applies channel-wise decay before the delta update. The factor $(I - \beta_t k_t k_t^\top)$ removes the retained association in the current key direction, while $\beta_t v_t k_t^\top$ writes the new value; $\beta_t = 0$ applies decay without delta overwrite, and larger β_t increasingly replaces the decayed association. This erase-then-write structure prevents repeated keys from accumulating incompatible values. We set the per-channel decay bias from d_h, zero the initial input-dependent decay projection for the controlled test, map concentration into native b_{proj} write rows, and initialize the sigmoid output-gate bias to the target damping probability. Table 6 summarizes what transfers.

Table 6: The framework shares teacher measurements, while each operator requires a native parameter map.

Quantity	GDN	GLA	KDA
Memory horizon	dt_{bias}	decay-projection bias	channel dt_{bias}
Write/erase strength	b_{proj}	absent	b_{proj}
Value/output path	pre-norm OLS + gate proxy	RMS gate proxy	sigmoid gate bias
Residual mismatch	local MSE	local MSE	local MSE

B Related Work

Efficient hybrid sequence models. Our work is motivated by the growing use of efficient sequence mixers in large language models. Production-scale hybrid and linear architectures now include Jamba, Jamba-1.5, MiniMax-01, Falcon-H1, and Kimi Linear (Lieber et al., 2024; Jamba Team, 2024; MiniMax Team, 2025; Falcon LLM Team, 2025; Kimi Team, 2025). This trend continues in the Qwen3.5 release, whose model card highlights a Gated DeltaNet-based efficient hybrid architecture (Qwen Team, 2026). A related efficiency direction compresses softmax attention itself, for example by using covariance-aware and rank-enhanced decomposition to enable multi-head latent attention (Zhou et al., 2026). These systems differ in their exact mixers, but they make the same architectural bet: many softmax attention layers can be replaced or compressed to reduce KV-cache cost and improve decoding throughput while preserving competitive quality.

Converting pretrained Transformers. Early work studied swap-then-finetune conversion (Kasai et al., 2021); recent methods add attention transfer, local matching, low-rank finetuning, or progressive distillation (Mercat et al., 2024; Wang et al., 2024; Zhang et al., 2025; Lan et al., 2025; Bick et al., 2025; Goldstein et al., 2025). RADLADS copies W_Q, W_K, W_V, W_O, then applies hidden-state alignment and KL distillation (Goldstein et al., 2025); GA-S2 uses this backbone while selecting which GDN layers remain softmax attention (Li et al., 2025). Both are held fixed in our comparisons.

Layer selection in hybrid models. For hybrid conversion, the choice of *which* layers remain softmax attention is also important. Redundancy-aware replacement and KL-guided layer selection show that layer choice can strongly affect long-context behavior and recovery quality (Hoshino et al., 2025; Li et al., 2025). This question is complementary to ours. We assume that a replacement set has already been chosen and focus on the next step: once a layer is converted, how should the new recurrent dynamics be initialized?

Initialization and local transfer. Mamba in the Llama maps teacher query/key/value projections into Mamba's $C/B/x$ paths while retaining the default A/Δ state-space dynamics (Wang et al., 2024). Attention to Mamba instead learns a Hedgehog feature map, maps it into an adapted Mamba layer, and initializes retention, convolution, and output gating to identity-like behavior before finetuning (Moudgil et al., 2026). These works establish that initialization matters. Our contribution targets a different missing quantity: teacher-conditioned initialization of the *new* recurrent dynamics themselves. The softmax-to-delta derivation identifies memory, write/erase, value-scale, and output-scale degrees of freedom; teacher statistics set them in each operator's native parameterization. This distinction also explains why Mamba mappings do not directly initialize GDN's coupled erase/write term, GLA's decay-only recurrence, or KDA's channel-wise decay.

Table 7: What prior conversion methods transfer and how they initialize newly introduced recurrent dynamics.

Method	Teacher-derived transfer	New recurrent dynamics at initialization
Mamba in the Llama (Wang et al., 2024)	Routes $Q/K/V$ into identifiable $C/B/x$ paths.	Retains default A/Δ state-space dynamics.
Attention to Mamba (Moudgil et al., 2026)	Learns a Hedgehog feature-map bridge.	Uses teacher-independent identity-like retention, convolution, and output gating.
Taylor-Calibrate	Copies $Q/K/V/O$ and measures teacher distance, concentration, and output scale.	Sets operator-native memory, write/erase, and output dynamics, then locally aligns the residual error.

C Limitations and Future Work

Taylor-Calibrate is a lightweight initialization method, not a complete solution to Transformer-to-recurrent conversion. Its role is to give the converted GDN layers a better starting point before full distillation, so several design choices remain deliberately simple. We summarize the main limitations below and describe concrete directions that could turn the initialization procedure into a more complete conversion recipe.

Scale coverage. The five-layer Inkling KDA run verifies that operator-native initialization and alignment execute with a sharded 975B MoE teacher. It is a same-domain integration test without a matched default-initialization control, so our initialization-effect claims rely on the controlled three-seed experiments; Inkling supports numerical and systems scalability.

Architecture specificity. The framework improves matched GLA and KDA students, but the exact GDN formulas are not portable. GLA has no write gate; KDA uses channel-wise decay; Mamba uses an additive diagonal state-space recurrence. Each operator must rederive how teacher distance, concentration, and output scale map into its native parameters. Our controlled non-GDN evidence covers one teacher and replacement set, and we have not tested a Mamba student.

Layer-wise rather than head-wise hybridization. We choose softmax layers at the layer level, following the layer-selection framing used by recent hybrid conversion work (Hoshino et al., 2025; Li et al., 2025). This keeps the experimental setup controlled and makes the retained-attention budget easy to compare, but it ignores an important source of structure inside a Transformer layer: different heads can serve very different roles. Some heads may perform retrieval or copying over long ranges, while others mostly aggregate local context or syntactic features. Converting all heads in a selected layer to GDN therefore may be too coarse, especially for models whose long-context behavior depends on a small number of specialized attention heads. A natural next step is head-wise conversion: keep retrieval-heavy heads as softmax, convert more local or diffuse heads to GDN, and initialize each converted head using the same distance, entropy, and value-scale statistics used by Taylor-Calibrate. This would also make the retained-attention budget more fine-grained, since a model could preserve only the heads that actually need quadratic attention.

Moment-reduction approximations. Distance-to-half-life and entropy-to-write-logit scaling reduce a full attention distribution to two moments; neither is a unique inverse of softmax behavior. Bias-free row scaling controls write-logit temperature/orientation, not the mean sigmoid gate. The OLS coefficient is exact only for a pre-normalization scalar proxy, is transferred to W_V only approximately with SiLU, and cancels from both RMSNorm and the recomputed tied gate in the ideal linear limit. Any practical effect therefore comes from clipping, nonlinearities, ϵ, and numerical conditioning. The sensitivity study in Appendix D.8 shows stable endpoints but no independent advantage for these maps over RMS+OLS in the focused control. Local alignment remains necessary for the coupled nonlinear recurrence.

Calibration and evaluation scope. Mechanism, sensitivity, GLA, and KDA controls report three-seed variation, but the expensive full recovery trajectories are single matched runs. Calibration uses a small corpus, and the multi-length fidelity control ends at 1,024 tokens. We have not exhaustively tested domain shift, much larger calibration sets, or variance after long recovery. LongBench and BFCL are complementary future evaluations, but they also mix conversion fidelity with instruction following, output formatting, and tool use. These limits separate evidence for a better initialization from stronger claims about final-model convergence.

Downstream distillation dependence. Taylor-Calibrate improves the starting point, but it does not remove the need for downstream distillation (Hinton et al., 2015; Goldstein et al., 2025; Li et al., 2025). The converted student still has a different sequence mixer from the teacher, so global training is needed to repair cross-layer interactions and adapt the residual stream to the new recurrent blocks. The final quality therefore depends on layer selection, retained-softmax budget, distillation loss, optimizer settings, and the number of recovery tokens. In this paper we keep those components mostly fixed so that the effect of initialization is easier to isolate. A more complete conversion recipe could jointly choose which layers or heads to retain, initialize the converted recurrent modules, and schedule the distillation losses based on how quickly each part of the model recovers. Another useful direction is adaptive token allocation: spend more recovery tokens on conversions whose local alignment loss or early validation PPL indicates a harder mismatch.

Broader impact. This work reduces the compute and token budget needed for hybrid conversion. That can make efficient long-context models easier to build and evaluate, especially for groups that cannot afford full pretraining of a new architecture. The same efficiency gain also inherits the usual dual-use risks of language-model research: cheaper adaptation can support beneficial applications, but it can also lower the cost of producing capable models for harmful uses. We do not release new models or datasets here, and the method should be viewed as an architecture-transfer tool rather than a deployment-ready safety mechanism. Any deployed student model would still require the same safety evaluation, data governance, and misuse analysis as other language models. Future work on efficient conversion should therefore report not only quality and compute savings, but also the evaluation procedures used before release.

D Extended Experimental Results

This appendix collects supplementary empirical material referenced in Sections 4 and 4.7. It reports additional teacher settings and the gate-statistics diagnostics used to interpret the initialization behavior.

D.1 Detailed Experimental Results

This section provides the detailed task-level results that supplement the compact main-paper summaries in Tables 2 and 3. Together the zero-shot and recovery tables contain 228 student rows and more than 2,700 task-level scores. The teacher settings use Qwen2.5 (Qwen Team, 2024), Llama-3.2 (Meta Llama Team, 2024), and Qwen3 (Qwen Team, 2025). The short-context suite includes ARC-C/E (Clark et al., 2018), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), MMLU (Hendrycks et al., 2021), OpenbookQA (Mihaylov et al., 2018), ReArc (Chollet, 2019), WinoGrande (Sakaguchi et al., 2021), BoolQ (Clark et al., 2019), LAMBADA (Paperno et al., 2016), COPA (Roemmele et al., 2011), and SciQ (Welbl et al., 2017), while RULER is reported as the long-context probe (Hsieh et al., 2024). The organization matches the main experimental flow: zero-shot evaluation for all four teacher settings, recovery dashboards for the two Qwen2.5 settings, and checkpoint recovery for all four teacher settings.

Zero-shot details. Table 8 reports the full zero-shot benchmark table for Qwen2.5-1.5B-Instruct and Qwen2.5-3B-Instruct (Qwen Team, 2024), Llama-3.2-3B-Instruct (Meta Llama Team, 2024), and Qwen3-8B (Qwen Team, 2025). RULER is reported as the long-context metric (Hsieh et al., 2024).

Table 8: Detailed zero-shot evaluation for all four teacher settings before downstream training. We compare five initialization methods under three retained-layer selection policies and three softmax budgets (25%, 50%, and 75%). Avg is computed over the available short-context metrics in each row; RULER is reported separately as a long-context probe. **Bold**: best non-teacher PPL (minimum), Avg (maximum), and RULER (maximum) within each model, softmax-budget, and selection block.

Softmax %	Selection	Init	PPL↓	ARC-C	ARC-E	Hella.	PIQA	MMLU	OBQA	RA	WG	BoolQ	LAMB.	COPA	SciQ	Avg	RULER
(a) Qwen2.5-1.5B-Instruct																	
Teacher			9.66↓	46.7	76.0	68.3	76.1	60.2	40.8	38.3	62.8	77.9	60.3	83.0	94.8	64.8	86.2
25%	Uniform	Baseline	951064↓	24.5	26.6	26.0	51.0	24.1	29.6	21.4	49.8	49.4	0.0	50.0	22.0	31.2	**0.2**
		Zero-Gate	256980↓	25.2	26.2	26.3	48.4	25.0	27.4	22.6	49.3	47.1	0.0	54.0	23.9	31.3	**0.2**
		Small-Gate	253804↓	24.3	25.3	26.3	48.5	25.1	27.4	21.1	50.2	47.4	0.0	57.0	23.7	31.4	**0.2**
		Taylor-Only	256725↓	25.3	25.7	26.5	48.7	24.9	27.6	21.6	50.7	48.0	0.0	59.0	24.9	31.9	**0.2**
		Taylor-Calibrate	**210.9**↓	24.7	31.6	28.6	53.9	22.9	29.4	22.6	51.5	54.3	0.5	59.0	38.8	**34.8**	**0.2**
	AR (LM-PPL)	Baseline	2763↓	23.5	29.8	27.2	50.8	23.2	26.0	23.0	50.3	46.4	0.1	52.0	28.1	31.7	0.0
		Zero-Gate	543.0↓	23.9	33.5	28.3	54.6	22.9	27.6	22.9	52.0	47.5	1.2	60.0	38.6	34.4	0.0
		Small-Gate	543.3↓	24.4	33.6	28.4	55.1	22.9	27.4	23.0	53.0	47.8	1.2	60.0	38.1	**34.6**	0.0
		Taylor-Only	542.6↓	24.2	33.7	28.4	55.0	22.9	27.6	23.3	51.9	47.7	1.3	59.0	38.5	34.5	0.0
		Taylor-Calibrate	**163.8**↓	23.8	34.0	28.5	54.6	23.1	26.2	22.5	51.2	43.2	1.0	58.0	46.8	34.4	**0.1**
	GA-S2	Baseline	98568↓	25.3	27.2	25.6	50.7	23.8	26.6	22.3	47.8	42.8	0.0	53.0	24.2	30.8	0.1
		Zero-Gate	1273↓	24.1	30.8	28.6	54.4	24.4	27.4	22.9	48.2	49.5	0.8	58.0	41.7	34.2	0.3
		Small-Gate	456.6↓	22.6	36.5	30.7	56.4	23.0	27.6	22.9	51.9	40.0	2.4	61.0	57.5	**36.0**	**0.5**
		Taylor-Only	526.8↓	23.9	33.3	29.1	55.1	23.1	28.0	21.5	51.1	39.7	1.5	58.0	53.2	34.8	**0.5**
		Taylor-Calibrate	**123.7**↓	24.0	33.6	28.4	55.1	23.6	24.8	25.2	49.8	40.2	1.2	58.0	47.6	34.3	0.2
50%	Uniform	Baseline	2892.9↓	24.3	27.6	27.2	48.4	23.6	26.8	21.9	51.1	44.2	0.0	54.0	31.6	31.7	**0.1**
		Zero-Gate	2377.8↓	24.8	29.8	26.9	51.6	23.4	28.6	22.3	52.3	43.8	0.0	55.0	33.5	32.7	**0.1**
		Small-Gate	2375.3↓	24.6	29.7	26.7	51.5	23.5	27.6	21.6	51.9	44.4	0.0	57.0	34.0	32.7	**0.1**
		Taylor-Only	2423.6↓	23.6	29.4	26.7	51.5	23.2	28.2	22.8	52.0	43.6	0.0	57.0	33.1	32.6	**0.1**
		Taylor-Calibrate	**75.2**↓	24.5	39.1	31.7	57.5	23.1	28.6	23.4	50.9	45.1	3.5	63.0	58.2	**37.4**	**0.1**
	AR (LM-PPL)	Baseline	133.0↓	27.0	39.7	34.9	58.4	24.1	30.2	25.2	52.1	45.8	2.9	64.0	56.3	38.4	**0.5**
		Zero-Gate	45.7↓	29.7	48.7	45.8	68.0	24.9	35.0	27.2	52.0	48.6	14.7	73.0	72.1	**45.0**	**0.5**
		Small-Gate	45.7↓	30.0	48.6	45.7	68.5	24.8	35.8	27.2	52.3	47.9	14.7	72.0	72.0	**45.0**	**0.5**
		Taylor-Only	45.6↓	29.8	48.4	45.9	68.1	24.8	35.2	27.1	52.3	47.7	14.6	72.0	72.3	44.8	**0.5**
		Taylor-Calibrate	**36.5**↓	29.0	48.4	42.2	67.5	26.7	31.6	27.8	52.3	40.0	15.3	72.0	75.4	44.0	**0.5**
	GA-S2	Baseline	2698.0↓	23.0	28.5	27.3	52.1	24.1	25.6	24.8	49.3	44.2	0.4	55.0	35.3	32.5	0.2
		Zero-Gate	100.5↓	24.2	45.5	34.0	62.8	24.3	28.0	24.5	51.2	51.9	6.2	58.0	74.1	40.4	0.6
		Small-Gate	139.4↓	27.8	45.5	35.1	65.7	24.1	29.6	27.0	51.7	46.2	9.5	68.0	68.5	**41.6**	**0.9**
		Taylor-Only	100.5↓	24.1	45.5	34.0	62.7	24.2	27.6	24.4	50.4	51.8	6.0	58.0	74.7	40.3	0.6
		Taylor-Calibrate	**45.5**↓	22.5	42.1	31.9	58.0	24.0	27.4	25.9	52.6	44.9	12.4	61.0	75.4	39.8	**0.9**
75%	Uniform	Baseline	43.9↓	34.4	55.1	49.5	69.0	29.3	34.2	30.5	56.4	42.8	27.3	79.0	76.2	48.6	2.7
		Zero-Gate	30.8↓	39.3	62.9	55.1	71.8	38.8	38.0	31.7	56.2	47.3	40.5	83.0	82.0	53.9	10.1
		Small-Gate	30.8↓	39.2	62.4	55.1	72.2	38.7	37.2	31.9	56.2	47.3	40.7	82.0	81.8	53.7	10.3
		Taylor-Only	30.8↓	39.5	62.5	55.2	72.1	38.6	37.2	32.0	56.2	47.4	40.7	83.0	82.0	53.9	10.1
		Taylor-Calibrate	**14.3**↓	42.2	66.2	59.3	72.5	42.2	38.8	34.5	57.7	48.1	46.5	80.0	86.1	**56.2**	**14.5**
	AR (LM-PPL)	Baseline	23.1↓	34.3	59.4	50.4	70.9	30.5	35.8	31.7	53.0	44.2	26.4	76.0	78.9	49.3	6.9
		Zero-Gate	15.0↓	38.4	60.5	59.5	73.2	43.1	38.6	31.9	54.5	48.1	36.0	81.0	84.4	54.1	**19.4**

Continued on next page

Table 8 (continued)

Softmax %	Selection	Init	PPL↓	ARC-C	ARC-E	Hella.	PIQA	MMLU	OBQA	RA	WG	BoolQ	LAMB.	COPA	SciQ	Avg	RULER
		Small-Gate	15.0↓	38.5	60.2	59.5	73.3	43.1	39.0	31.8	53.6	48.0	36.1	81.0	84.3	54.0	19.1
		Taylor-Only	15.0↓	38.6	60.6	59.5	73.5	43.0	38.4	31.8	54.1	48.0	36.0	81.0	84.2	54.0	19.2
		Taylor-Calibrate	**13.2**↓	40.6	68.2	59.0	74.3	44.5	39.4	32.1	53.5	53.8	42.3	81.0	87.8	**56.4**	19.2
	GA-S2	Baseline	32.9↓	28.9	49.8	42.6	66.5	24.8	32.4	27.8	52.0	43.8	17.8	59.0	70.1	43.0	4.2
		Zero-Gate	20.1↓	33.4	59.9	53.8	72.4	38.4	37.0	30.9	54.1	49.9	32.8	74.0	86.9	52.0	7.7
		Small-Gate	22.1↓	36.9	62.8	55.9	72.4	37.4	38.4	29.5	54.6	62.9	30.8	76.0	85.2	53.6	2.3
		Taylor-Only	21.4↓	36.1	60.8	54.4	71.8	36.6	37.2	31.5	53.7	50.0	29.7	79.0	88.1	52.4	2.4
		Taylor-Calibrate	**14.3**↓	44.0	71.6	57.7	72.1	40.6	37.6	35.4	57.6	68.4	46.8	82.0	93.3	**58.9**	**39.3**
(b) Qwen2.5-3B-Instruct																	
Teacher			8.56↓	48.0	72.9	75.0	78.2	66.4	41.8	40.2	69.5	80.2	65.7	85.0	94.7	67.3	91.3
25%	Uniform	Baseline	13323↓	25.5	26.9	26.0	52.2	22.9	27.2	21.4	50.1	37.8	0.0	60.0	23.1	31.1	0.0
		Zero-Gate	11503↓	24.6	26.8	26.2	52.6	22.9	28.0	21.3	51.9	37.8	0.0	66.0	23.5	**31.8**	0.0
		Small-Gate	11534↓	24.4	26.8	26.2	52.6	22.9	27.8	21.1	52.1	37.8	0.0	66.0	23.4	**31.8**	0.0
		Taylor-Only	11503↓	24.4	26.8	26.1	52.2	22.9	28.2	21.4	51.7	37.8	0.0	65.0	23.6	31.7	0.0
		Taylor-Calibrate	**241.0**↓	22.0	30.4	27.1	51.9	22.9	24.4	20.5	49.3	38.7	0.1	55.0	28.8	30.9	**0.1**
	AR (LM-PPL)	Baseline	1815↓	22.1	30.8	28.5	50.9	23.6	22.8	24.7	51.5	46.1	0.8	61.0	47.6	34.2	1.3
		Zero-Gate	333.8↓	23.3	33.5	29.4	55.6	24.6	24.2	23.4	51.8	39.1	3.2	57.0	45.9	34.3	1.4
		Small-Gate	337.1↓	23.7	33.4	29.5	55.4	24.4	24.6	23.4	50.8	39.7	3.3	56.0	44.9	34.1	**1.5**
		Taylor-Only	334.9↓	23.0	33.6	29.4	55.6	24.6	23.6	23.6	51.9	39.1	3.3	57.0	45.1	34.2	1.4
		Taylor-Calibrate	**112.8**↓	22.5	35.2	29.5	56.1	24.0	24.6	24.0	50.0	52.9	2.8	61.0	67.1	**37.5**	0.4
	GA-S2	Baseline	14342↓	26.4	25.9	26.1	50.8	22.9	29.4	21.6	50.1	37.8	0.0	53.0	21.6	30.5	0.0
		Zero-Gate	11638↓	26.2	26.3	25.7	50.8	22.9	28.4	22.2	48.1	37.8	0.0	51.0	21.9	30.1	0.0
		Small-Gate	12138↓	26.1	26.4	25.7	50.4	22.9	27.6	21.9	47.8	37.8	0.0	52.0	22.2	30.1	0.0
		Taylor-Only	11571↓	25.9	26.3	25.7	50.3	22.9	27.8	22.0	48.5	37.8	0.0	51.0	22.0	30.0	0.0
		Taylor-Calibrate	**115.9**↓	21.8	35.4	29.4	55.0	23.0	26.2	23.3	50.7	57.2	4.4	56.0	56.9	**36.6**	**0.2**
50%	Uniform	Baseline	8761↓	25.8	27.2	26.0	52.1	22.9	24.8	20.8	49.6	37.8	0.0	64.0	24.1	31.3	0.0
		Zero-Gate	8811↓	26.5	27.9	25.8	51.6	22.9	26.0	21.2	49.4	37.8	0.0	62.0	24.6	31.3	0.0
		Small-Gate	8804↓	26.1	28.2	25.9	51.4	22.9	26.2	21.0	50.4	37.8	0.0	62.0	24.4	31.4	0.0
		Taylor-Only	8807↓	26.3	28.0	25.8	51.5	22.9	26.0	21.1	49.2	37.8	0.0	62.0	24.9	31.3	0.0
		Taylor-Calibrate	**33.3**↓	24.2	37.2	32.7	57.1	23.0	24.4	25.5	48.6	46.2	7.4	58.0	51.8	**36.3**	**0.4**
	AR (LM-PPL)	Baseline	319.6↓	23.4	37.9	34.0	56.7	24.1	27.6	25.0	50.5	43.9	3.7	64.0	55.5	37.2	2.4
		Zero-Gate	55.9↓	29.6	54.5	44.8	66.4	27.0	34.4	27.9	50.4	48.0	15.7	78.0	73.1	45.8	**2.5**
		Small-Gate	56.1↓	29.4	54.3	44.8	66.9	26.9	34.0	28.0	51.5	48.5	15.9	78.0	73.4	**46.0**	**2.5**
		Taylor-Only	55.8↓	29.6	54.3	44.9	66.6	26.9	34.2	27.3	50.7	47.7	15.6	77.0	73.1	45.7	2.4
		Taylor-Calibrate	**34.2**↓	26.2	49.6	40.7	64.9	26.1	33.6	26.1	51.9	57.9	19.7	70.0	71.1	44.8	**2.5**
	GA-S2	Baseline	8950↓	25.4	26.4	26.1	51.9	23.0	28.0	22.5	50.0	37.8	0.0	56.0	21.6	30.7	0.0
		Zero-Gate	10477↓	25.9	26.5	25.6	49.9	22.9	27.6	22.6	47.9	37.8	0.0	52.0	21.8	30.0	0.0
		Small-Gate	10290↓	25.8	26.2	25.8	50.9	22.9	27.8	22.5	48.4	37.8	0.0	52.0	21.8	30.2	0.0
		Taylor-Only	10393↓	25.8	26.4	25.5	50.1	22.9	27.4	22.4	48.2	37.8	0.0	51.0	21.8	29.9	0.0
		Taylor-Calibrate	**53.3**↓	25.8	45.6	34.5	59.7	23.7	30.0	25.2	50.4	50.4	15.0	69.0	68.9	**41.5**	**0.9**
75%	Uniform	Baseline	19.7↓	41.8	63.8	63.1	74.8	41.0	40.6	31.3	59.4	62.5	21.8	78.0	83.1	55.1	19.9
		Zero-Gate	12.9↓	47.8	72.9	69.6	76.6	54.5	40.8	34.7	62.4	72.7	32.0	85.0	92.0	61.8	47.4
		Small-Gate	12.9↓	48.0	73.2	69.7	76.9	54.7	40.6	34.7	63.1	72.8	32.0	86.0	92.1	**62.0**	47.2
		Taylor-Only	12.9↓	48.0	72.9	69.5	76.7	54.7	41.0	34.8	63.0	72.9	31.9	85.0	92.0	61.9	47.2
		Taylor-Calibrate	**12.0**↓	47.7	72.4	69.2	75.2	54.0	41.8	35.8	62.0	73.4	32.9	83.0	88.5	61.3	**55.9**
	AR (LM-PPL)	Baseline	28.9↓	39.2	61.5	55.7	71.1	40.2	37.8	34.6	55.6	63.0	40.7	77.0	87.3	55.3	38.7
		Zero-Gate	13.2↓	43.9	71.3	65.4	74.1	50.7	40.0	35.5	62.2	76.2	56.1	81.0	91.3	**62.3**	56.2
		Small-Gate	13.2↓	44.2	71.1	65.3	74.2	50.8	40.2	35.5	62.2	76.3	56.3	80.0	91.4	**62.3**	56.1
		Taylor-Only	13.2↓	43.9	70.8	65.3	74.4	50.9	40.4	35.4	61.8	76.4	56.0	79.0	91.1	62.1	56.2
		Taylor-Calibrate	**12.5**↓	43.1	71.3	63.7	73.2	49.1	41.2	36.7	62.5	78.2	56.4	80.0	89.3	62.1	**63.7**
	GA-S2	Baseline	13728↓	25.6	26.1	26.0	51.8	22.9	28.6	21.7	48.4	37.8	0.0	56.0	22.2	30.6	0.0
		Zero-Gate	10588↓	25.7	26.4	25.7	50.5	22.9	28.4	22.7	49.3	37.8	0.0	51.0	21.7	30.2	0.0
		Small-Gate	10674↓	25.4	26.3	25.6	51.1	22.9	27.8	22.4	48.7	37.8	0.0	52.0	21.8	30.2	0.0
		Taylor-Only	10572↓	25.9	26.1	25.7	50.8	22.9	27.8	22.9	48.4	37.8	0.0	52.0	21.9	30.2	0.0
		Taylor-Calibrate	**12.9**↓	48.7	72.6	67.2	75.5	44.6	38.8	38.0	61.2	72.3	54.7	79.0	86.2	**61.6**	**13.8**
(c) Llama-3.2-3B-Instruct																	
Teacher			11.05↓	46.0	71.4	71.7	76.7	60.6	39.4	41.6	68.8	78.6	64.4	82.0	95.4	65.6	89.6
25%	Uniform	Baseline	15109↓	26.1	24.9	26.2	51.8	24.4	28.4	21.4	50.6	44.4	0.0	55.0	21.1	31.2	0.0
		Zero-Gate	1010↓	21.4	31.9	28.3	55.9	23.0	28.2	22.6	52.1	37.9	0.7	51.0	44.1	33.1	0.0
		Small-Gate	1036↓	21.2	32.2	28.4	56.6	23.0	27.8	22.4	52.1	37.9	0.7	52.0	44.4	33.2	**0.1**
		Taylor-Only	1013↓	21.1	32.3	28.4	55.7	23.0	27.8	22.4	51.8	37.9	0.7	52.0	44.4	33.1	**0.1**
		Taylor-Calibrate	**385.3**↓	22.9	32.1	28.7	55.2	22.9	26.2	22.0	49.6	38.3	1.4	59.0	51.6	**34.2**	**0.1**
	AR (LM-PPL)	Baseline	16207↓	24.2	25.4	25.7	51.7	23.9	26.6	21.7	50.0	42.3	0.0	63.0	19.9	31.2	**0.1**
		Zero-Gate	696.6↓	23.9	28.8	28.3	56.7	22.9	26.8	22.2	49.7	37.9	0.8	54.0	36.1	32.3	**0.1**
		Small-Gate	693.7↓	24.3	29.0	28.2	57.1	22.9	27.0	22.0	51.1	37.8	0.8	53.0	36.4	32.5	**0.1**
		Taylor-Only	696.8↓	24.3	29.0	28.3	56.7	22.9	26.4	22.1	49.6	37.9	0.9	55.0	36.0	32.4	**0.1**
		Taylor-Calibrate	**132.0**↓	24.3	34.0	30.9	58.8	24.5	27.2	23.4	51.1	42.4	0.7	63.0	47.5	**35.7**	0.0
	GA-S2	Baseline	7649↓	25.0	26.0	26.7	53.6	23.5	28.4	21.8	50.9	41.3	0.0	53.0	21.6	31.0	**0.1**
		Zero-Gate	1347↓	24.0	27.8	27.4	55.0	24.0	29.4	21.9	48.3	37.8	0.2	56.0	34.7	32.2	0.0
		Small-Gate	1355↓	24.1	27.7	27.3	55.4	23.9	29.0	21.9	49.3	37.8	0.2	55.0	34.1	32.1	0.0
		Taylor-Only	1348↓	24.0	27.8	27.4	55.5	24.1	29.2	21.9	48.1	37.8	0.2	56.0	34.8	32.2	0.0
		Taylor-Calibrate	**146.4**↓	27.3	37.0	35.0	61.5	23.4	27.6	22.9	51.4	56.8	0.2	63.0	36.8	**36.9**	**0.1**
50%	Uniform	Baseline	8843↓	24.4	26.6	25.9	51.9	23.3	27.4	22.9	47.7	41.0	0.0	64.0	25.5	31.7	0.1
		Zero-Gate	595.1↓	21.5	38.4	33.4	60.4	23.8	27.2	24.0	50.1	42.3	3.1	66.0	52.6	36.9	0.2
		Small-Gate	607.4↓	22.4	38.2	33.3	60.3	23.8	27.2	24.0	49.3	42.8	3.0	65.0	51.6	36.7	0.2
		Taylor-Only	599.1↓	21.3	38.4	33.4	60.3	23.8	27.2	24.3	49.3	42.2	3.0	66.0	52.7	36.8	0.2
		Taylor-Calibrate	**70.4**↓	25.1	47.7	40.1	64.6	25.2	30.2	27.6	52.7	56.4	10.1	71.0	68.4	**43.3**	**0.4**

Continued on next page

Table 8 (continued)

Softmax %	Selection	Init	PPL↓	ARC-C	ARC-E	Hella.	PIQA	MMLU	OBQA	RA	WG	BoolQ	LAMB.	COPA	SciQ	Avg	RULER
	AR (LM-PPL)	Baseline	2033↓	22.7	31.3	28.7	54.4	23.5	26.6	22.8	50.8	43.9	0.0	62.0	31.2	33.2	0.1
		Zero-Gate	111.1↓	25.8	41.6	41.7	64.0	29.0	27.6	24.5	52.3	58.4	13.6	68.0	65.7	42.7	10.1
		Small-Gate	111.5↓	26.4	41.8	41.7	63.7	28.8	28.6	24.6	51.5	59.0	13.5	68.0	65.0	42.7	10.3
		Taylor-Only	111.2↓	25.8	41.7	41.8	64.1	29.0	27.8	24.2	51.3	58.3	13.6	67.0	65.6	42.5	10.1
		Taylor-Calibrate	**33.0**↓	32.8	53.5	50.8	69.0	33.4	33.6	29.7	56.6	65.0	15.7	75.0	75.2	**49.2**	**14.3**
	GA-S2	Baseline	2181↓	25.7	27.1	26.6	51.0	23.5	26.6	23.0	48.2	39.5	0.1	53.0	25.2	30.8	0.2
		Zero-Gate	322.4↓	23.2	34.6	31.0	56.5	22.9	26.4	23.7	51.3	37.9	3.4	52.0	63.3	35.5	0.2
		Small-Gate	322.4↓	23.4	34.9	31.0	57.6	23.0	26.8	23.9	51.2	37.9	3.5	53.0	67.6	36.2	0.2
		Taylor-Only	322.2↓	23.1	35.2	31.0	57.7	22.9	26.2	23.6	50.9	37.9	3.4	54.0	67.5	36.1	0.2
		Taylor-Calibrate	**57.6**↓	35.2	59.3	56.9	70.5	43.3	34.8	29.7	57.0	61.7	19.1	79.0	84.4	**52.6**	**1.2**
75%	Uniform	Baseline	3081↓	23.1	30.9	28.2	52.1	23.5	27.2	23.6	52.0	48.6	0.1	57.0	40.9	33.9	0.2
		Zero-Gate	25.4↓	39.7	66.9	62.1	73.1	41.7	38.6	35.1	62.4	68.6	44.1	75.0	92.5	58.3	**31.0**
		Small-Gate	25.4↓	40.0	66.7	62.1	72.8	41.8	39.0	35.5	62.9	68.5	44.2	77.0	92.6	58.6	30.9
		Taylor-Only	25.4↓	39.8	66.8	61.9	73.3	41.8	38.8	35.5	62.8	68.5	44.1	76.0	92.5	58.5	30.9
		Taylor-Calibrate	**21.9**↓	40.4	68.3	64.7	74.3	42.7	38.2	38.3	62.6	70.8	50.7	76.0	93.3	**60.0**	24.0
	AR (LM-PPL)	Baseline	46.3↓	32.7	56.6	51.3	69.8	33.5	35.6	29.9	53.3	52.4	21.2	71.0	80.3	49.0	9.5
		Zero-Gate	14.8↓	42.2	68.8	66.6	75.1	52.3	39.2	38.3	65.4	67.7	45.8	77.0	89.5	60.7	55.4
		Small-Gate	14.8↓	42.2	68.7	66.5	75.1	52.3	39.8	38.1	65.0	67.4	46.0	77.0	89.4	60.6	55.5
		Taylor-Only	14.8↓	42.1	68.9	66.5	75.1	52.3	39.8	38.2	65.0	67.6	45.9	77.0	89.5	60.7	55.4
		Taylor-Calibrate	**13.7**↓	44.1	71.0	68.0	75.3	54.0	39.4	39.1	66.0	71.7	50.9	79.0	91.7	**62.5**	**60.5**
	GA-S2	Baseline	41.6↓	35.3	54.8	55.0	70.8	35.2	34.8	32.4	58.6	61.2	17.7	74.0	80.5	50.9	2.9
		Zero-Gate	158.4↓	22.9	42.0	34.4	62.6	23.7	27.4	24.7	52.6	47.1	11.7	62.0	79.8	40.9	0.5
		Small-Gate	157.5↓	23.2	42.3	34.4	63.1	23.7	27.2	24.6	52.2	47.2	11.7	63.0	79.7	41.0	0.5
		Taylor-Only	158.3↓	22.9	42.2	34.5	63.1	23.6	27.4	24.4	52.4	47.1	11.7	62.0	80.0	40.9	0.5
		Taylor-Calibrate	**14.6**↓	44.8	70.2	69.4	77.0	61.7	38.0	41.5	68.2	79.7	52.5	81.0	93.6	**64.8**	**66.3**
(d) Qwen3-8B																	
Teacher			9.73↓	56.7	80.9	74.9	77.9	74.9	41.8	41.4	68.0	86.6	64.1	85.0	96.6	70.7	94.0
25%	Uniform	Baseline	43609↓	25.1	28.4	26.6	49.2	24.9	29.8	21.1	50.6	41.9	0.0	55.0	23.9	31.4	**0.1**
		Zero-Gate	1106↓	21.8	32.0	27.6	53.9	24.2	26.2	22.1	49.5	54.9	0.3	56.0	42.1	34.2	**0.1**
		Small-Gate	1103↓	21.5	31.6	27.7	53.6	24.3	26.6	22.2	49.0	55.1	0.3	55.0	42.4	34.1	0.0
		Taylor-Only	1105↓	21.7	31.6	27.7	53.5	24.3	26.0	22.2	48.8	54.7	0.3	57.0	42.4	34.2	**0.1**
		Taylor-Calibrate	**166.2**↓	21.2	37.5	29.3	55.7	23.0	27.0	22.8	50.7	47.3	0.3	59.0	53.3	**35.6**	**0.1**
	AR (LM-PPL)	Baseline	3015↓	23.0	27.7	26.1	51.9	24.1	29.6	21.7	50.7	38.9	0.2	55.0	32.7	31.8	**0.1**
		Zero-Gate	482.7↓	22.2	34.6	28.2	53.3	24.0	25.4	22.9	50.5	37.9	2.5	53.0	49.5	33.6	0.0
		Small-Gate	484.6↓	22.5	34.6	28.2	54.0	23.9	25.4	23.1	50.3	37.9	2.5	50.0	49.9	33.5	0.0
		Taylor-Only	483.0↓	22.3	34.8	28.1	53.6	23.9	25.8	23.1	50.2	37.9	2.5	52.0	49.5	33.6	0.0
		Taylor-Calibrate	**131.9**↓	21.8	35.0	28.7	55.1	23.1	24.4	23.5	51.2	38.6	4.8	62.0	58.3	**35.5**	**0.1**
	GA-S2	Baseline	571090↓	22.8	26.1	25.9	51.5	24.3	27.6	21.5	51.4	46.1	0.0	49.0	22.7	30.7	0.1
		Zero-Gate	16897↓	24.7	28.2	25.9	53.2	23.2	24.4	23.0	49.8	38.0	0.0	57.0	30.9	31.5	0.0
		Small-Gate	16915↓	24.1	28.2	25.9	53.3	23.3	25.2	22.3	49.6	38.2	0.0	55.0	30.8	31.3	0.0
		Taylor-Only	16970↓	24.7	28.0	26.2	53.0	23.2	24.6	23.1	50.2	38.0	0.0	58.0	30.7	31.6	0.0
		Taylor-Calibrate	**204.8**↓	20.4	34.6	28.3	54.2	23.1	27.4	23.4	51.1	38.1	2.7	61.0	53.3	**34.8**	**0.2**
50%	Uniform	Baseline	361.1↓	27.7	33.8	35.9	56.9	23.9	29.4	25.6	50.6	45.1	1.1	63.0	41.2	36.2	0.2
		Zero-Gate	53.5↓	32.9	49.5	47.3	68.5	23.8	30.2	28.2	52.2	55.8	8.1	77.0	67.2	45.1	0.3
		Small-Gate	53.5↓	33.1	49.6	47.3	68.1	23.9	30.4	28.2	52.2	55.6	8.1	76.0	66.8	44.9	0.3
		Taylor-Only	53.5↓	33.1	49.5	47.4	68.0	23.7	30.8	28.1	51.6	55.4	8.1	77.0	66.9	45.0	0.3
		Taylor-Calibrate	**32.9**↓	35.8	57.4	48.9	71.8	24.9	33.2	28.3	54.0	62.6	13.7	80.0	77.6	**49.0**	**0.7**
	AR (LM-PPL)	Baseline	223.4↓	23.7	42.2	33.0	57.9	25.2	25.6	23.8	49.4	39.4	5.9	55.0	63.5	37.0	0.2
		Zero-Gate	129.2↓	27.3	50.8	34.0	61.9	25.1	27.6	25.6	49.3	40.2	8.9	64.0	71.6	40.5	0.3
		Small-Gate	128.9↓	27.0	50.8	34.0	62.0	24.8	27.6	25.6	49.6	40.2	8.9	64.0	72.3	40.6	**0.4**
		Taylor-Only	129.4↓	27.1	50.7	34.0	62.0	24.9	27.8	25.7	49.7	40.2	8.9	64.0	72.1	40.6	**0.4**
		Taylor-Calibrate	**52.0**↓	28.2	53.8	37.5	65.7	25.4	28.8	28.1	52.1	50.0	14.7	69.0	74.6	**44.0**	**0.4**
	GA-S2	Baseline	1046↓	27.4	39.0	34.6	59.1	24.9	32.6	23.3	50.8	43.1	2.3	57.0	48.1	36.8	0.2
		Zero-Gate	1043↓	22.2	40.3	31.1	57.2	23.0	28.6	22.3	51.1	38.1	2.0	61.0	64.4	36.8	0.1
		Small-Gate	373.4↓	20.8	37.5	30.0	57.9	24.7	27.2	22.9	49.0	38.5	2.1	57.0	61.6	35.8	0.0
		Taylor-Only	1045↓	21.7	40.1	31.0	57.0	23.1	28.6	22.4	51.8	38.1	1.9	62.0	64.1	36.8	0.1
		Taylor-Calibrate	**69.6**↓	23.4	47.4	34.4	63.9	25.4	29.6	25.4	52.6	38.4	11.3	67.0	75.9	**41.2**	**0.4**
75%	Uniform	Baseline	19.2↓	47.5	69.9	61.6	73.3	55.7	39.0	37.6	62.1	72.8	32.8	77.0	88.9	59.9	42.3
		Zero-Gate	11.8↓	52.6	76.9	67.3	76.1	60.8	38.8	39.7	63.9	80.4	49.0	79.0	93.0	64.8	61.6
		Small-Gate	11.8↓	52.8	76.8	67.2	76.3	60.9	38.8	39.8	63.7	80.4	49.1	78.0	93.2	64.7	61.4
		Taylor-Only	11.8↓	52.5	76.8	67.0	76.1	60.8	38.8	39.7	63.8	80.6	49.0	79.0	93.2	64.8	61.6
		Taylor-Calibrate	**11.4**↓	54.3	77.7	67.5	75.8	61.1	38.4	39.3	65.7	81.5	49.0	82.0	93.9	**65.5**	**65.1**
	AR (LM-PPL)	Baseline	24.2↓	37.6	64.4	53.5	72.1	27.4	34.0	29.9	54.7	51.3	25.3	80.0	83.9	51.2	16.6
		Zero-Gate	15.9↓	42.6	71.6	56.0	75.2	37.0	37.4	33.7	56.4	61.0	37.1	81.0	88.5	56.5	31.4
		Small-Gate	15.9↓	43.4	71.6	56.0	75.2	36.9	37.8	33.6	56.4	61.3	37.1	81.0	88.4	56.6	31.2
		Taylor-Only	15.9↓	43.1	71.7	56.0	75.2	36.7	38.0	33.2	56.1	60.8	37.0	81.0	88.6	56.5	31.1
		Taylor-Calibrate	**13.8**↓	44.9	73.5	56.8	76.6	41.3	36.2	33.9	58.5	66.1	40.9	80.0	90.0	**58.2**	**38.2**
	GA-S2	Baseline	**20.6**↓	39.8	64.8	58.5	72.2	43.1	34.8	33.2	57.5	62.4	35.1	78.0	88.4	**55.6**	**15.8**
		Zero-Gate	27.2↓	35.3	63.0	49.5	71.9	26.1	34.6	30.6	53.6	46.6	25.5	73.0	83.4	49.4	7.1
		Small-Gate	27.1↓	35.4	63.0	49.6	71.5	26.2	34.2	30.6	53.7	46.6	25.6	72.0	83.4	49.3	7.0
		Taylor-Only	27.2↓	34.9	63.1	49.6	71.9	26.3	34.4	30.5	53.3	46.5	25.4	72.0	83.6	49.3	7.0
		Taylor-Calibrate	23.1↓	36.5	62.6	48.8	71.8	29.3	37.0	30.9	53.5	53.3	24.6	75.0	76.8	50.0	6.3

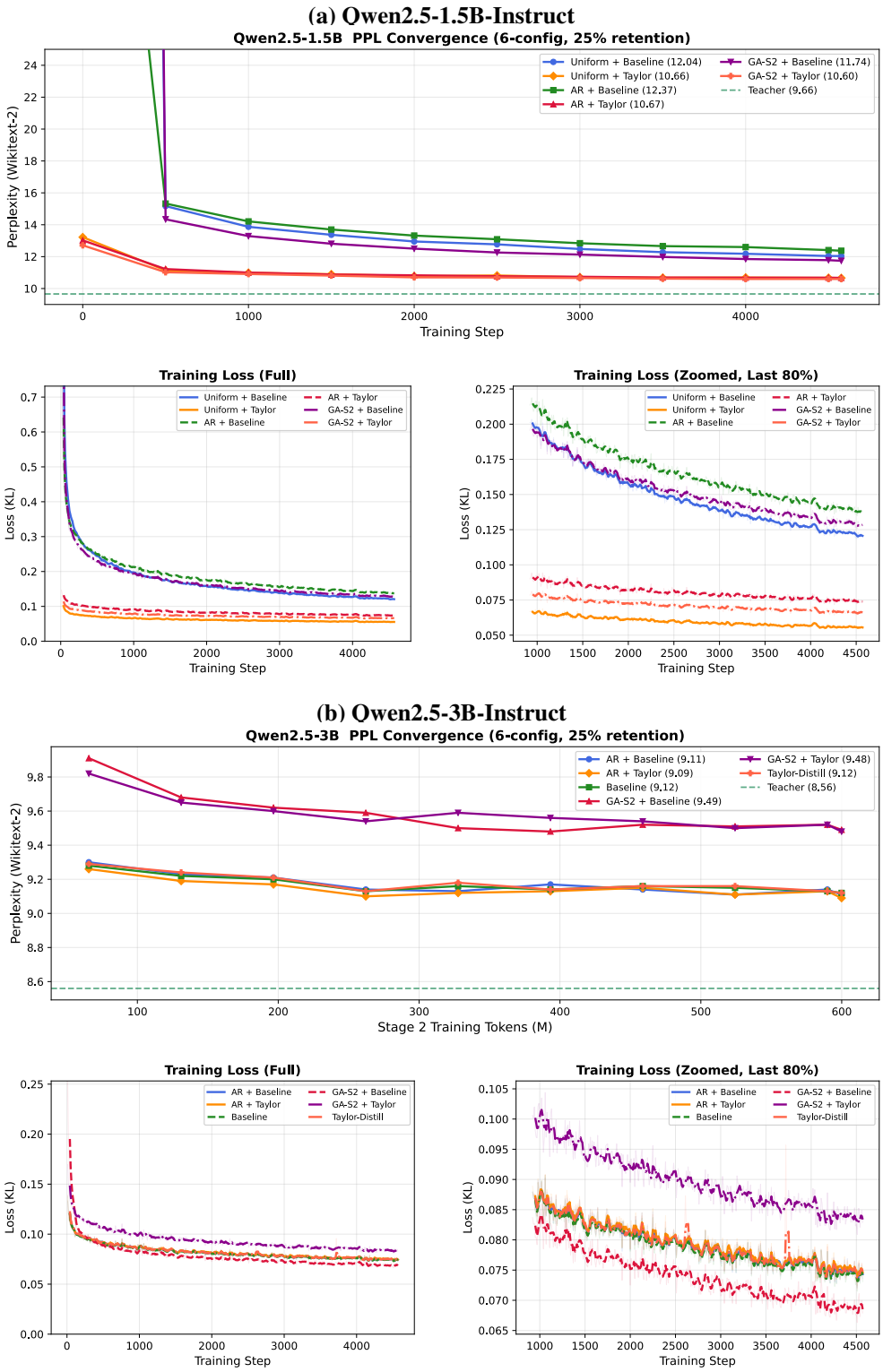


Figure 6: Full recovery dashboards for the two Qwen2.5 settings. The top panels report PPL versus Stage-2 tokens; lower panels show the complete and late-training KL trajectories.

Recovery details. Table 9 reports the full checkpoint benchmark table for Qwen2.5-1.5B-Instruct and Qwen2.5-3B-Instruct (Qwen Team, 2024), Llama-3.2-3B-Instruct (Meta Llama Team, 2024), and Qwen3-8B (Qwen Team, 2025).

Table 9: Detailed recovery results for all four teacher settings at 100M tokens after Stage 1 and 700M tokens after Stage 2. Avg is shown only when the full short-context set needed for that row is available; RULER is reported separately as a long-context probe. **Bold**: best non-teacher PPL (minimum), Avg (maximum), and RULER (maximum) within each model block.

Ckpt	Selection	Init	PPL↓	ARC-C	ARC-E	Hella.	PIQA	MMLU	OBQA	RA	WG	BoolQ	LAMB.	COPA	SciQ	Avg	RULER
(a) Qwen2.5-1.5B-Instruct																	
Teacher			9.66	46.7	76.0	68.3	76.1	60.2	40.8	38.3	62.8	77.9	60.3	83.0	94.8	64.8	86.2
100M (Stage 1)	Uniform	Baseline	210.13	23.6	29.6	33.9	55.9	25.5	29.2	25.7	50.9	46.0	3.8	62.0	59.8	37.2	0.6
		Taylor-Calibrate	13.54	45.9	75.9	65.9	75.4	50.8	40.8	35.4	63.1	75.8	46.7	83.0	92.8	62.6	7.7
	AR (LM-PPL)	Baseline	58.71	30.8	59.0	48.0	67.8	23.1	31.4	27.9	52.6	46.1	5.6	74.0	70.0	44.7	1.0
		Taylor-Calibrate	12.89	42.7	73.4	65.4	75.3	53.4	39.8	33.4	59.1	66.5	46.8	85.0	92.5	61.1	5.0
	GA-S2	Baseline	142.41	35.1	60.8	40.6	68.9	27.1	29.2	30.1	51.2	44.1	5.8	71.0	59.4	43.6	0.5
		Taylor-Calibrate	12.54	45.6	74.5	65.5	75.8	51.9	40.8	35.0	60.5	66.8	51.6	83.0	92.5	62.0	7.1
700M (Stage 2)	Uniform	Baseline	11.99	41.0	71.6	63.5	75.5	44.1	41.2	34.4	63.3	71.8	54.7	82.0	93.1	61.4	39.9
		Taylor-Calibrate	10.76	45.6	76.8	66.6	76.2	53.1	39.8	37.4	63.9	76.5	59.0	83.0	94.6	**64.4**	**59.8**
	AR (LM-PPL)	Baseline	12.23	43.4	73.6	63.3	75.4	48.5	40.4	34.8	61.3	66.9	51.8	86.0	92.0	61.4	16.8
		Taylor-Calibrate	10.65	45.0	75.2	66.4	75.8	55.7	39.0	35.3	62.9	72.0	56.8	85.0	94.0	63.6	38.0
	GA-S2	Baseline	11.96	43.9	73.9	64.1	75.5	41.6	39.0	34.8	61.6	63.5	53.7	81.0	91.1	60.3	27.0
		Taylor-Calibrate	**10.57**	46.8	76.3	66.4	75.9	54.6	40.0	34.9	63.0	70.9	57.5	85.0	93.7	63.8	47.6
(b) Qwen2.5-3B-Instruct																	
Teacher			8.56	48.0	72.9	75.0	78.2	66.4	41.8	40.2	69.5	80.2	65.7	85.0	94.7	67.3	91.3
100M (Stage 1)	Uniform	Baseline	10.78	44.3	76.0	53.6	77.7	48.5	32.2	35.7	62.5	71.6	47.2	83.0	93.5	60.5	26.3
		Taylor-Calibrate	10.83	44.2	75.3	53.6	77.8	51.5	33.0	34.9	62.0	72.2	47.1	84.0	93.9	60.8	29.2
	AR (LM-PPL)	Baseline	10.32	43.8	74.7	53.3	78.0	57.5	32.2	36.5	61.6	71.3	58.6	84.0	93.4	62.1	11.5
		Taylor-Calibrate	10.33	44.2	75.8	53.2	78.2	57.8	32.8	35.8	61.7	69.4	58.2	85.0	93.0	62.1	11.7
	GA-S2	Baseline	15.97	29.7	53.7	42.7	67.9	49.8	22.6	34.5	59.0	68.8	36.7	82.0	84.3	52.6	6.6
		Taylor-Calibrate	11.73	42.1	73.7	53.5	78.4	57.2	32.6	35.5	60.3	62.3	54.7	83.0	92.0	60.4	4.9
700M (Stage 2)	Uniform	Baseline	9.09	46.9	71.9	72.8	77.8	55.1	43.6	36.5	65.8	76.8	63.2	87.0	92.5	65.8	63.8
		Taylor-Calibrate	**8.95**	45.4	69.9	72.8	77.6	57.0	43.2	36.9	68.0	78.0	64.1	87.0	92.5	66.0	65.7
	AR (LM-PPL)	Baseline	9.08	46.2	70.2	72.6	76.8	61.5	43.0	36.8	65.5	74.8	64.0	86.0	92.3	65.8	56.1
		Taylor-Calibrate	8.98	46.9	69.3	72.9	77.1	61.8	43.2	37.8	66.4	75.3	64.3	86.0	91.7	**66.1**	57.9
	GA-S2	Baseline	9.45	46.2	71.9	72.8	77.0	55.8	42.0	36.6	67.5	77.4	63.5	86.0	92.2	65.7	59.8
		Taylor-Calibrate	9.13	47.3	71.7	72.9	76.9	61.3	42.8	37.3	65.7	73.7	64.6	85.0	91.4	65.9	**70.5**
(c) Llama-3.2-3B-Instruct																	
Teacher			11.05	46.0	71.4	71.7	76.7	60.6	39.4	41.6	68.8	78.6	64.4	82.0	95.4	65.6	89.6
100M (Stage 1)	Uniform	Baseline	20.74	44.6	70.5	66.8	75.3	38.3	37.2	35.5	61.1	73.7	40.7	80.0	93.1	59.7	5.7
		Taylor-Calibrate	17.41	44.8	70.4	67.5	75.5	39.9	38.2	37.4	61.3	74.7	50.6	82.0	93.0	61.3	6.6
	AR (LM-PPL)	Baseline	20.43	47.7	73.4	66.9	76.7	31.3	39.2	34.4	59.8	72.2	28.4	82.0	90.1	58.5	4.6
		Taylor-Calibrate	17.45	45.2	71.0	67.7	76.1	36.1	39.6	36.1	59.5	73.4	43.8	81.0	91.8	60.1	5.4
	GA-S2	Baseline	18.87	43.7	71.0	65.8	75.5	26.5	38.8	34.0	58.1	70.1	37.9	82.0	90.7	57.8	5.0
		Taylor-Calibrate	16.83	45.4	71.2	67.3	76.3	31.9	38.6	34.7	63.7	73.9	39.0	80.0	92.2	59.5	4.2
700M (Stage 2)	Uniform	Baseline	12.05	46.7	72.4	69.5	76.1	49.6	37.2	41.1	65.0	77.3	61.1	81.0	94.9	64.3	60.0
		Taylor-Calibrate	**11.62**	45.9	72.2	69.9	76.0	49.7	39.0	41.6	66.3	77.7	61.3	83.0	94.8	**64.8**	**63.4**
	AR (LM-PPL)	Baseline	12.22	47.3	73.2	69.6	76.5	41.4	39.6	39.0	67.2	74.5	58.6	84.0	94.3	63.8	46.4
		Taylor-Calibrate	11.85	44.8	71.3	69.8	76.7	45.0	40.4	40.0	67.0	75.0	60.9	82.0	95.1	64.0	47.3
	GA-S2	Baseline	12.05	45.4	70.9	68.8	76.2	34.6	39.8	37.6	66.5	73.2	58.1	82.0	94.0	62.3	41.1
		Taylor-Calibrate	11.92	46.2	71.7	69.5	77.1	42.1	39.4	39.0	67.1	75.7	60.2	84.0	94.9	63.9	46.5
(d) Qwen3-8B																	
Teacher			9.73	56.7	80.9	74.9	77.9	74.9	41.8	41.4	68.0	86.6	64.1	85.0	96.6	70.7	94.0
100M (Stage 1)	Uniform	Baseline	83.86	37.6	60.6	51.4	69.7	31.8	33.4	30.1	51.4	66.0	5.1	76.0	79.9	49.4	0.1
		Taylor-Calibrate	14.92	56.1	79.9	72.8	77.3	66.2	43.6	37.1	67.9	79.0	36.9	86.0	94.7	66.5	9.2
	AR (LM-PPL)	Baseline	172.30	32.4	51.6	49.0	68.4	27.2	32.6	29.4	51.1	57.3	3.0	69.0	65.0	44.7	0.2
		Taylor-Calibrate	11.47	55.5	79.2	72.5	77.3	66.4	43.6	37.2	68.3	78.9	53.4	84.0	95.0	67.6	15.4
	GA-S2	Baseline	297.36	31.1	53.4	35.9	68.3	24.2	27.4	26.0	50.0	45.7	4.8	53.0	57.4	39.8	0.1
		Taylor-Calibrate	11.12	56.7	79.7	72.5	77.4	63.0	41.8	36.5	68.1	78.0	53.5	87.0	95.3	67.4	11.0
700M (Stage 2)	Uniform	Baseline	10.40	56.2	80.1	73.5	78.0	67.9	42.4	36.6	69.2	82.3	61.3	89.0	96.2	69.4	60.4
		Taylor-Calibrate	9.98	56.1	80.3	74.4	78.3	70.2	42.4	38.0	69.2	84.6	63.1	87.0	96.3	70.0	**76.7**
	AR (LM-PPL)	Baseline	9.90	55.5	81.4	73.4	77.9	66.8	42.6	36.8	70.7	82.9	60.5	85.0	96.6	69.2	62.2
		Taylor-Calibrate	**9.45**	56.1	79.2	74.2	78.3	70.2	43.6	39.7	70.3	84.5	62.9	86.0	96.0	**70.1**	69.8
	GA-S2	Baseline	10.18	55.2	79.8	73.9	79.0	62.9	43.8	36.8	70.3	83.7	59.7	86.0	96.1	68.9	66.1
		Taylor-Calibrate	9.46	56.9	80.7	74.0	78.3	69.0	43.0	40.1	71.4	82.9	61.9	87.0	96.0	**70.1**	57.2

D.2 MATCHED-PPL TOKEN-BUDGET CALCULATION

Figure 5 is a descriptive matched-quality comparison over five selected model–policy trajectories spanning three teacher models; it is not an exhaustive summary of all teachers or policies. The analysis fixes the row-specific target PPL values T in Table 10. If the projection-copy curve does not contain an exact evaluation at T, we interpolate linearly in PPL versus cumulative Stage-2 tokens

between adjacent checkpoints (τ_1, p_1) and (τ_2, p_2):

$$\tau_{\mathrm{BL}}(T) = \tau_1 + \frac{p_1 - T}{p_1 - p_2}\,(\tau_2 - \tau_1)\,. \tag{55}$$

We do not extrapolate. In every reported row, Taylor-Calibrate has already passed T by its first evaluated Stage-2 checkpoint, so τ_{TC} is only known to be no larger than that checkpoint. Consequently, $\tau_{\mathrm{BL}}/\tau_{\mathrm{TC}}$ is a conservative lower bound. The common approximately 100M-token Stage-1 budget is excluded, and each trajectory is one run without a variance estimate.

Table 10: Inputs to Figure 5. Token coordinates τ are in millions. "Exact" uses an evaluated projection-copy checkpoint; other rows use Eq. 55. Figure labels round T to one decimal place.

Model	Selection	T	Projection-copy source (τ, PPL)	τ_{BL}	τ_{TC}	Saving
Qwen2.5-1.5B	Uniform	12.00	$(589.824, 12.00)$ (exact)	589.824	≤ 65.536	$\geq 9.0\times$
Qwen2.5-1.5B	AR	12.21	$(599.917, 12.21)$ (exact)	599.917	≤ 65.536	$\geq 9.2\times$
Qwen2.5-1.5B	GA-S2	12.00	$(524.288, 12.06) \rightarrow (589.824, 11.99)$	580.462	≤ 65.536	$\geq 8.9\times$
Qwen2.5-3B	GA-S2	10.70	$(262.144, 10.83) \rightarrow (294.912, 10.68)$	290.543	≤ 32.768	$\geq 8.9\times$
Llama-3.2-3B	GA-S2	13.26	$(131.072, 13.40) \rightarrow (196.608, 13.09)$	160.669	≤ 32.768	$\geq 4.9\times$

D.3 Measured Calibration Cost

We profile a 21-of-28-layer Qwen2.5-1.5B conversion on one H100-80GB with 1,024 calibration tokens and 300 local steps per converted layer. Loading is separate from the calibration total.

Table 11: Measured initialization cost.

Component	Wall time
Teacher forward and statistics	0.190 s
Analytical calibration	1.302 s
Layer-local alignment	85.328 s
Total calibration	**86.821 s (0.0241 H100-h)**
Checkpoint construction/loading (separate)	51.9 s
Peak calibration memory	7.81 GiB

D.4 RULER Subtask Decomposition

The 4K bundle contains eight NIAH variants (`niah_single_1/2/3`, `niah_multikey_1/2/3`, `niah_multiquery`, and `niah_multivalue`) and five tracking, extraction, and QA probes (`ruler_vt`, `ruler_cwe`, `ruler_fwe`, `ruler_qa_squad`, and `ruler_qa_hotpot`). Table 12 separates these groups; the GA-S2 regression is concentrated in the latter.

Table 12: Qwen3-8B RULER at 700M tokens. Entries are Baseline→Taylor-Calibrate.

Selection	Overall	8 NIAH mean	5 other-task mean
Uniform	60.4→**76.7**	69.0→**82.2**	46.8→**67.9**
AR (LM-PPL)	62.2→**69.8**	70.7→**79.4**	48.7→**54.2**
GA-S2	**66.1**→57.2	**72.5**→70.5	**55.9**→35.9

D.5 Matched GLA and KDA Transfer

All rows in Table 13 use Qwen2.5-1.5B-Instruct, replace the same 7 of 28 layers, and hold copied projections, calibration data, alignment budget, and evaluation fixed. Three-task Avg is the macro average of ARC-Easy normalized accuracy, HellaSwag normalized accuracy, and PIQA accuracy.

In a 14-of-28-layer GLA stress test, Alignment-only, output stabilization, and native distance calibration obtain PPL 181.18 ± 3.76, 129.89 ± 1.03, and 123.00 ± 2.43, respectively. The ordering survives, but aggressive replacement sharply lowers absolute quality.

Table 13: Operator-native initialization, mean±standard deviation over three seeds.

Operator	Initialization	MSE↓	Cosine↑	KL↓	32K PPL↓	3-task Avg↑
GLA	Alignment only	.3409±.0065	.2660±.0047	.7595±.0506	26.17±.18	60.83±.26
	+ output stabilization	.2889±.0027	.3022±.0004	.6135±.0068	23.08±.20	**62.54±.18**
	+ native distance calibration	**.2885±.0011**	**.3078±.0007**	**.5726±.0087**	**22.31±.06**	62.01±.19
KDA	Alignment only	.2547±.0069	**.3552±.0009**	.6435±.0159	23.17±.51	61.92±.42
	+ output stabilization	**.2281±.0027**	.3534±.0027	**.5786±.0094**	**22.65±.28**	**65.83±.09**
	+ native distance/concentration	.2343±.0043	.3379±.0005	.5870±.0043	23.15±.11	65.77±.23

D.6 INKLING-975B FIVE-LAYER KDA SCALE STUDY

Model and numerical validation. Inkling is a 975B-parameter, 66-layer MoE with 41B active parameters per token, 55 sliding-window attention layers, 11 full-attention layers, and 256 routed experts (Thinking Machines Lab, 2026). We use the official 1.9TB BF16 checkpoint rather than a dequantized NVFP4 model. Before conversion, the xorl FSDP2/EP implementation obtains sequence-128 PPL 7.177 versus 7.152 for the Transformers reference, with selected per-layer activation cosine approximately 1.000. This validates the checkpoint mapping, including the checkpoint's element-interleaved gate/up expert layout, and the distributed teacher forward.

KDA-native initialization and distributed protocol. Copied $Q/K/V/O$ projections do not determine KDA's memory horizon, delta write/erase strength, or output amplitude. We therefore map teacher attention distance to channel-wise retention α_t, concentration to write rate β_t, and output scale to the sigmoid gate, then locally align all five converted layers. All 64 GPUs form one FSDP2/EP-sharded model replica of the 1.9TB teacher. The run uses sequence length 128, 200 local-alignment steps, and a 600-step end-to-end KL continuation. Section 4.6 reports PPL 7.030 and 85.3% top-1 agreement after Taylor-Calibrate plus local alignment, improving to 6.498 and 90.2% after global KL, versus teacher PPL 6.454.

Data-split segregation and held-out guarantees. To guarantee zero data contamination across all evaluations: (i) For all 1.5B–8B teacher conversions, $\mathcal{D}_{\text{cal}}$ uses a compact 1,024-token calibration set (4 sequences of 256 tokens), while test perplexity is evaluated on the remaining held-out portion ($>$ 280K tokens) with the calibration prefix strictly excluded; all downstream NLU benchmarks and RULER are independent standardized datasets with zero overlap. (ii) For Inkling-975B, sequence-128 diagnostics use disjoint token slices ($[0, 1024)$ for calibration vs. $[1024, 17408)$ for evaluation), although both slices come from the same source domain.

This is a five-layer integration diagnostic, not a throughput or full-linearization result: 61 original attention sublayers remain, and there is no matched default-initialization control. It verifies that teacher-statistic extraction, KDA-native parameter mapping, local alignment, and global KL remain numerically stable at 975B scale; only the matched smaller-scale studies identify the causal initialization effect.

D.7 DIRECT TEACHER-FIDELITY CONTROLS

We evaluate Qwen2.5-1.5B-Instruct with the same seven converted layers, calibration split, and local alignment budget across three seeds. MSE and cosine are averaged over converted-layer outputs on held-out inputs; KL is token-level $D_{\text{KL}}(p_T \| p_S)$; PPL uses a 32K-token held-out stream.

Table 14: Direct mechanism evidence, reported as mean±standard deviation over three seeds.

Initialization	Pre-align MSE↓	Post-align MSE↓	Cosine↑	KL↓	PPL↓
Alignment only	.5558±.0088	.2793±.0085	.3231±.0055	.7068±.0753	23.15±.21
Taylor-Calibrate	**.4148±.0000**	**.2240±.0061**	**.4543±.0062**	**.5248±.0109**	**18.39±.20**

D.8 Functional-Form Sensitivity

This focused control uses Qwen2.5-1.5B-Instruct, the same 7 of 28 converted layers, and three seeds. We vary one Phase-1 choice at a time. "Output-only" retains output-gate stabilization but omits decay, write-logit, and OLS proxies; "RMS+OLS" adds the pre-normalization scalar OLS proxy. The narrow endpoint range tests robustness rather than broad model coverage.

Table 15: Calibration sensitivity, mean±standard deviation over three seeds.

Setting	Held-out MSE↓	Cosine↑	KL↓	32K PPL↓
Default Taylor-Calibrate	.2240±.0061	.4543±.0062	.5248±.0109	18.39±.20
Half-life 0.5×	.2235±.0035	.4525±.0021	.5187±.0134	18.26±.08
Half-life 2×	.2226±.0054	.4552±.0047	.5153±.0038	18.44±.05
Constant $\beta = 0.5$	**.2160±.0010**	.4584±.0017	.5161±.0071	**18.22±.03**
Output scale .005	.2236±.0059	.4543±.0060	.5226±.0038	18.37±.17
Output scale .02	.2227±.0059	.4546±.0054	.5224±.0106	18.35±.21
Output-only control	.2430±.0022	.3597±.0027	.6245±.0160	21.12±.18
RMS+OLS control	.2226±.0022	**.4606±.0080**	**.4852±.0101**	18.23±.10

At 300 local steps, output-only, RMS+OLS, and full Taylor-Calibrate obtain PPL 21.12, 18.23, and 18.39, respectively. The 0.16 PPL difference between RMS+OLS and full Taylor-Calibrate is comparable to the reported seed-to-seed standard deviations and should be read as a tie, not a reliable reversal. The KL difference is larger (.4852 vs. .5248), however, so this is not evidence that distance/write-logit priors improve every metric. Figure 7 likewise shows overlapping held-out MSE/cosine trajectories from 10–300 steps. Full Taylor-Calibrate still substantially improves over Alignment-only (PPL 18.39 vs. 23.15; KL .5248 vs. .7068), and the broad recovery study compares against projection-copy initialization. The correct attribution is therefore narrower: output stabilization, the OLS/gate proxy, and local alignment explain most of the gain in this seven-layer control; distance and concentration provide bounded teacher-conditioned operating points, not independently established accelerators.

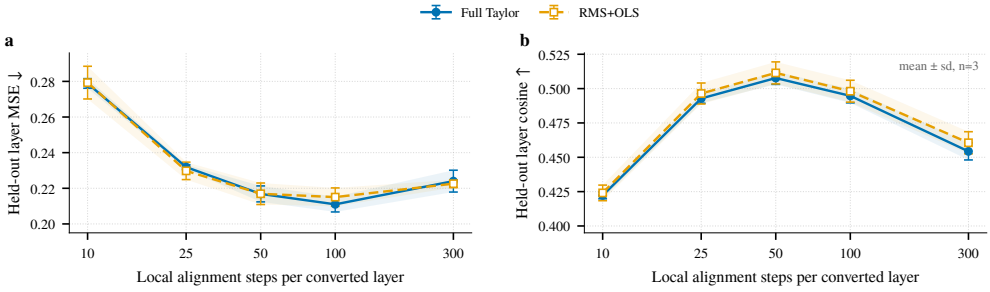


Figure 7: Alignment-budget control on Qwen2.5-1.5B-Instruct with the same seven converted GDN layers. Curves report mean±standard deviation over three seeds. Full Taylor-Calibrate and the RMS+OLS control track closely from 10 to 300 local steps, providing no evidence that decay/concentration calibration independently accelerates local alignment under this protocol.

D.9 Fidelity Across Evaluated Sequence Lengths

Table 16 evaluates one held-out sequence at each length. Taylor-Calibrate improves PPL and teacher KL at every point. The maximum length is 1,024, so this table tests whether local fidelity degrades with length; RULER remains the long-context behavioral evaluation.

Table 16: Held-out local fidelity across input lengths.

Length	PPL↓		Teacher KL↓	
	Alignment only	Taylor-Calibrate	Alignment only	Taylor-Calibrate
128	17.78	**15.91**	.734	**.533**
256	11.79	**10.30**	.678	**.546**
512	12.76	**10.90**	.765	**.596**
1,024	12.51	**10.06**	.843	**.632**

D.10 Gate Value Statistics

This section examines whether the gains from Taylor-Calibrate are reflected in the internal GDN gate parameters (Yang et al., 2024a), rather than only in downstream benchmark scores. We analyze the Qwen2.5-3B-Instruct conversion (Qwen Team, 2024) with the uniform 25% full-attention policy. The retained softmax layers are $\{0, 4, 8, 12, 16, 20, 24, 28, 32\}$, leaving 27 recurrent GDN layers and 432 recurrent heads per initialization variant. We compare four settings: Baseline random GDN initialization, Zero-Gate, Taylor-Only after analytical calibration, and the full Taylor-Calibrate initialization after the layer-local alignment step.

For each recurrent head, we compute the idle half-life implied by the decay parameters,

$$\tau_{1/2} = \frac{\ln 2}{\exp(A_{\log})\ \mathrm{softplus}(dt_{\mathrm{bias}})}. \tag{56}$$

This statistic measures the memory timescale induced by the initialized recurrence before input-dependent updates are applied. Table 17 summarizes the aggregate statistics, while Figures 8, 9, and 10 show the corresponding layer-wise distributions and phase-wise changes.

Table 17: Aggregate gate statistics for the Qwen2.5-3B-Instruct uniform-25% conversion. Half-life is reported in idle recurrent steps. Gate columns report mean absolute parameter magnitude.

Variant	HL mean	HL median	HL p90	HL max	$\lvert A_{\log}\rvert$	$\lvert b_{\mathrm{proj}}\rvert$	$\lvert g_{\mathrm{proj}}\rvert$
Baseline	131.995	10.260	88.079	3.06×10^4	1.923	0.016	0.016
Zero-Gate	73.887	11.794	115.951	2834.355	1.855	0.016	0
Taylor-Only	69.337	71.049	100.053	128.373	0	0.009	1.32×10^{-4}
Taylor-Calibrate	70.030	71.975	100.420	127.503	0.011	0.016	0.012

The main pattern is that Taylor calibration imposes a much more regular recurrent memory prior than the default GDN initialization. Baseline initialization has a heavy-tailed decay spectrum, with extreme half-life outliers up to roughly 3.06×10^4 steps. In contrast, the Taylor-calibrated variants concentrate around a median half-life of approximately 71 steps and cap the maximum half-life near 128 steps. This supports the intended role of Phase 1: it sets a controlled memory timescale using teacher attention statistics, rather than relying on the default random decay spectrum.

The layer-local alignment step changes a different part of the parameterization. The mean half-life changes only from 69.337 to 70.030 steps after alignment, a shift of 0.692 steps, indicating that Phase 2 largely preserves the analytically calibrated timescale. Its main effect is instead on the projection gates. The mean absolute magnitude of g_{proj} increases from 1.32×10^{-4} after Taylor-Only to 0.012 after full Taylor-Calibrate, an 89.8× increase, while b_{proj} increases by 1.73×. Thus Phase 1 places the recurrent state in a controlled memory regime, and Phase 2 activates and orients the write/output gates so that the converted layer produces teacher-like outputs.

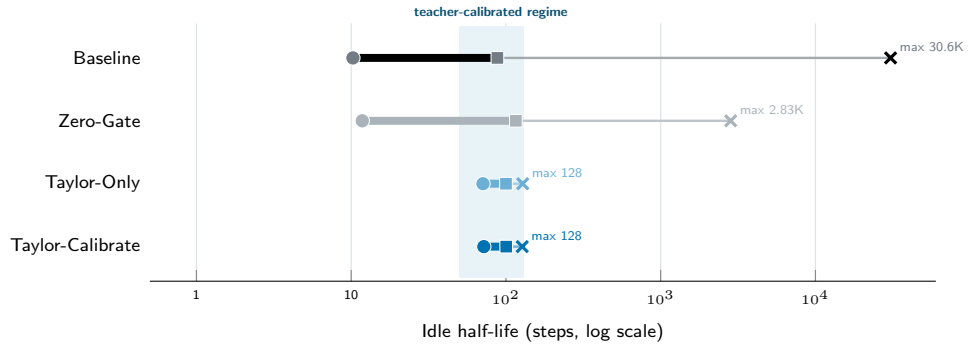


Figure 8: Idle half-life distribution over 432 recurrent heads. Circles, squares, and crosses denote the median, p90, and maximum. Taylor calibration removes extreme random-decay tails and concentrates both calibrated variants in the teacher-conditioned 50–128-step regime.

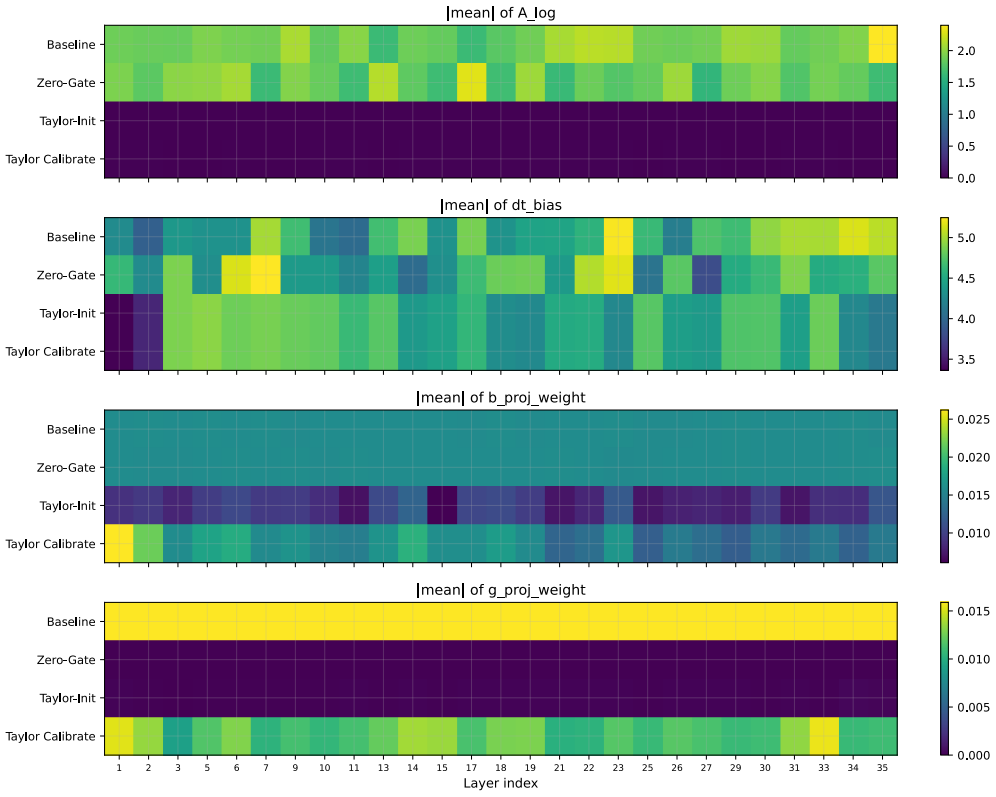


Figure 9: Layer-wise magnitudes of the main GDN gate parameters. The Taylor-Only setting suppresses unstable random gate scales, while full Taylor-Calibrate restores non-trivial projection-gate magnitude after local alignment.

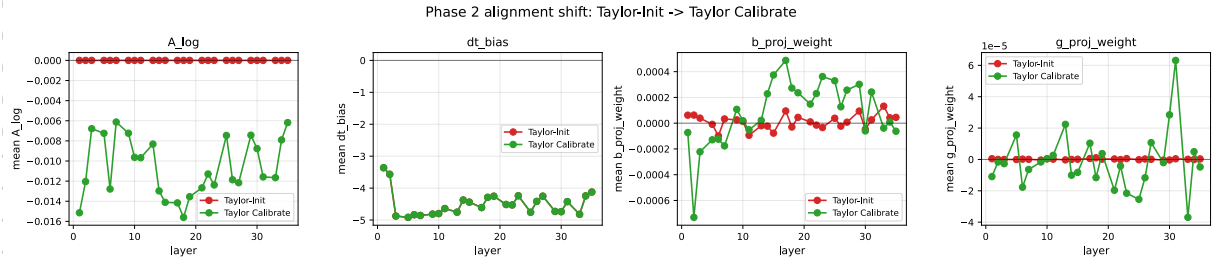


Figure 10: Parameter changes from Taylor-Only to full Taylor-Calibrate after layer-local alignment. The recurrent half-life remains nearly fixed, while the write and output projection gates receive most of the alignment-induced change.