SCALING LAWS FOR ATTENTION-TO-RECURRENCE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

An efficient way of reducing the computational cost of pretrained large language models is to generate hybrids by replacing some of the traditional attention layers, whose complexity scales quadratically with context length, with recurrent alternatives such as state-space models, which scale linearly with context length. Approaches such as MOHAWK allow distilling hybrids under limited token budgets compared to output-level Knowledge Distillation (KD), while retaining model capabilities. Yet, they lack the theoretical foundation needed to guide practitioners on how many full-attention layers can be replaced and at what cost.

Here, we aim to close this gap. Using arguments from spectral theory, we argue for a conditional lower bound on the Kullback–Leibler (KL) divergence loss between hybrid and non-hybrid models in a restricted componentwise setting, connecting it to the attention matrix approximation error in MOHAWK's matrix-orientation phase. We also study a scaling law for the context-dependent KL loss based on the ratio of replaced layers, their expressivity and the Phase 2 token budget. Finally, we verify experimentally that the KL divergence loss often exhibits power-law decay within a fixed conversion setup, while finding that its late-budget shape does not transfer cleanly across model sizes. Overall, our work sets the foundation for a theory that characterises the practical utility of distilling hybrids from pretrained models.

1 INTRODUCTION

Attention gives language models flexible access to their context, but its cache and compute costs grow with sequence length. Recurrent mixers such as state-space models use a fixed-size state and linear-time sequence mixing (Gu et al., 2022; Gu & Dao, 2024; Dao & Gu, 2024). Hybrid models retain some attention layers and replace the rest. This makes them attractive at inference time, although training a useful hybrid from scratch remains expensive.

Distillation offers another route: MOHAWK converts a pretrained Transformer into a recurrent or hybrid student with far less data than training a new model from scratch (Bick et al., 2024). This raises a practical question: *how much data will a chosen conversion need, and what quality will remain out of reach?*

Our first hypothesis was deliberately simple. Structured state-space duality writes the recurrent mixer as a semiseparable matrix, so a state of size d_{state} bounds its rank across temporal cuts (Dao & Gu, 2024). Approximation difficulty should then be organised by the context-to-state load L/d_{state}, and if the local matrix error of Phase 0 controls the final distillation loss, a compact joint scaling law in data, context and state size may be possible.

Parts of this picture survive. Across ten softmax attention teachers, the spectral tail rises smoothly with L/d_{state}, supporting it as a coordinate. A conditional componentwise single-head theorem states the assumptions needed to turn that tail into a KL lower bound. The downstream experiments rule out the simplest reading. A larger recurrent state does not reliably lower final KL in our OLMo-2 1B state-size intervention. Data scaling is more stable within a fixed conversion setup, but its late-budget shape does not transfer cleanly across model sizes. Some 530M and 750M trajectories flatten or reverse while all 300M trajectories continue to improve.

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.

These negative results help identify which variables can be forecast, which conditions must remain fixed and where extrapolation breaks. Our contributions are:

- We formulate conversion as a conditional scaling problem, with held-out direct KL as the response and the setup and evaluation distribution kept explicit.
- We connect a deterministic cross-cut spectral obstruction to distillation loss through a conditional theorem for a componentwise one-block, one-head setting.
- We report Phase-2 data ladders for 15 model-size and replacement-ratio cells through D64, plus an OLMo-2 1B cross-family check. Early-budget predictions are useful, but structured errors and the state intervention reject a universal joint law.
- We give a protocol for future studies: freeze the setup, predict unseen endpoints, retain context as an evaluation axis and add architectural variables only after held-out interventions.

2 Related work

From Transformers to recurrent and hybrid models. Hybrids can be pretrained with attention and recurrent layers (AI21 Labs, 2024; Ren et al., 2025; Chandiramani et al., 2026) or found by architecture search after pretraining (Gu et al., 2025). MOHAWK, Llamba, Mamba in the Llama and LoLCATs instead convert a pretrained Transformer through staged distillation, projection transfer or attention linearisation (Bick et al., 2024; 2025a; Wang et al., 2024; Zhang et al., 2025), as do related linearisation methods (Zhang et al., 2024; Mercat et al., 2024; Goldstein et al., 2025). In-context retrieval remains a recurring weakness. It has been linked to a small set of gather-and-aggregate heads, and retaining these heads as attention narrows the gap (Bick et al., 2025b; 2026). Studies of pretrained hybrids also find that recall and the preferred layer arrangement depend on the attention ratio and context length (Wang et al., 2025; Lee et al., 2025). We ask how the remaining imitation gap changes with conversion data, replacement ratio and context.

Scaling laws and their estimation. Classical scaling laws model pretraining loss from model size, data and compute (Kaplan et al., 2020; Hoffmann et al., 2022), and have been extended to linear-complexity and hybrid models trained from scratch (Shen et al., 2024; Poli et al., 2024). Distillation laws add teacher size and quality but keep the architecture fixed (Busbridge et al., 2025). Recent estimation work uses small model ladders, branched cooldowns, horizon-dependent learning rates and broken power laws (Bhagia et al., 2025; Choshen et al., 2025; Hu et al., 2024; Hägele et al., 2024; Bjorck et al., 2025; Caballero et al., 2023). We apply these ideas to the imitation gap after an architectural conversion, a response for which we are not aware of a prior scaling law.

What a fixed-size state can represent. Structured state-space duality writes Mamba-2 as a semiseparable matrix mixer (Dao & Gu, 2024). Related analyses identify limits in state tracking, copying, associative recall, in-context retrieval and state capacity over long contexts (Merrill et al., 2024; Jelassi et al., 2024; Arora et al., 2023; Wen et al., 2025; Chen et al., 2025). We instead connect a matrix-level obstruction to distillation KL and measure its response to data.

3 Attention-to-recurrence distillation

Sequence mixers as matrices. Attention and recurrent layers both mix the positions of an input $X \in \mathbb{R}^{L \times d_{\text{model}}}$ through an $L \times L$ matrix (Hwang et al., 2024),

$$H(X) = M(X)V(X), \tag{1}$$

where $V(X)$ holds one value vector per position and $M_{ij}(X)$ is the weight with which output position i reads position j. Causality makes M lower triangular, and since heads act independently we describe a single head. In typical softmax attention (Vaswani et al., 2017), with $Q = XW_Q$, $K = XW_K$ and $V = XW_V$, the *mixing matrix* is

$$M(X) = \operatorname{softmax}\big(QK^\top/\sqrt{d_k} + C\big), \tag{2}$$

where C is the usual causal mask. The logit matrix $QK^\top$ has rank at most d_k (Bhojanapalli et al., 2020), but the row-wise softmax removes this constraint, so the attention matrix is generally full rank

even though it is often approximately low rank (Wang et al., 2020). A Mamba-2 layer, like other recurrent mixers with a matrix-valued state (Katharopoulos et al., 2020; Yang et al., 2024; 2025), instead carries a state of fixed size d_{state}: for one value channel it computes $h_i = a_i h_{i-1} + b_i v_i$ and $y_i = c_i^\top h_i$, where $a_i \in \mathbb{R}$ and $b_i, c_i \in \mathbb{R}^{d_{\text{state}}}$ depend on the input, and unrolling the recursion recovers Equation (1) with $M_{ij} = c_i^\top b_j \prod_{l=j+1}^{i} a_l$ for $j \le i$ (Dao & Gu, 2024). Because information from the prefix can reach the suffix only through the state, every block below the diagonal has rank at most d_{state}, a property known as semiseparability (Vandebril et al., 2008). Replacing attention by recurrence therefore swaps an unconstrained causal matrix for a rank-limited one (Section 4).

MOHAWK as a sequence of approximations. MOHAWK exploits this shared view in three stages (Bick et al., 2024). Phase 0 (matrix orientation) fits each recurrent mixer to the frozen teacher attention matrix by minimising the Frobenius distance between the two mixing matrices on the same inputs. Before Phase 0, each recurrent mixer is initialised from the attention layer it replaces: its value and output projections are copied from the teacher, and for each head the query–key bilinear form $W_Q W_K^\top$ is replaced by its best rank-d_{state} factorisation, split evenly between the projections producing c_i and b_j so that $c_i^\top b_j$ approximates the scaled logit $q_i^\top k_j / \sqrt{d_k}$, while the remaining recurrent parameters keep their default initialisation. Phase 1 (hidden-state alignment) then trains each converted block to reproduce the output of the teacher block it replaces, in the spirit of hint-based distillation (Romero et al., 2015), and Phase 2 (knowledge distillation (Hinton et al., 2015)) assembles the student, transfers compatible teacher weights (Wang et al., 2024) and distils the whole network against the teacher distribution. Because the stages move from a local linear object to the full nonlinear model, success in Phase 0 need not determine the final Phase-2 KL: values, residual paths and later layers can either amplify or repair a local mismatch.

Response variable. For an evaluation distribution $\mathcal{E}$ over contexts and target positions, and Phase-2 budget D, our primary response at context length L is

$$K_{\mathcal{E}}(D, L) = \mathbb{E}_{(x,t)\sim\mathcal{E}}[D_{\text{KL}}(p_T(\cdot \mid x_{<t}) \,\|\, p_{S,D}(\cdot \mid x_{<t}))]\,. \tag{3}$$

We use this forward, teacher-to-student KL as our measure of imitation, rather than the reverse direction favoured by some generative distillation methods (Gu et al., 2024). It measures imitation rather than absolute language-modelling quality. We also report cross-entropy (CE) and excess CE, the student's CE minus the matched teacher's CE. These metrics answer related but different questions.

Unlike the settings of prior scaling laws (Section 2), the student here is produced by an architectural intervention. Phase 2 starts from the checkpoints produced by Phase 0 and Phase 1, not from a generic initialisation. The replacement ratio and layer placement determine where attention is replaced, while d_{state} limits the rank available across temporal cuts. These choices can affect both the initial direct KL and its response to Phase-2 data.

Context length matters for a similar reason. Transformer loss itself follows a power law in context length (Kaplan et al., 2020; Xiong et al., 2023), while a recurrent mixer must compress a longer prefix into a fixed-size state, and prior work reports related weaknesses on copying and retrieval (Section 2). We therefore keep L visible and report $K_{\mathcal{E}}(D, L)$ at each context length. This lets us ask whether more Phase-2 data reduces KL in the same way at short and long contexts, and whether a trend measured at one context carries over to another. An average over context lengths would fold these questions into one number and could hide a student that tracks the teacher at short contexts but not at long ones, as averaged perplexity is known to obscure long-context ability (Fang et al., 2025; Hsieh et al., 2024). Keeping L explicit therefore lets us test the context-to-state hypothesis developed next.

4 From matrix spectra to a conditional KL bound

Section 3 leaves a concrete question: why should a longer context be harder for a recurrent replacement with a fixed state? We approach it by stepping away from training loss and asking what each mixer can represent across a temporal cut. Attention maps all prefix positions directly to later positions. A recurrent mixer must carry the same interaction through a fixed-dimensional state, which restricts its rank. We then give conditions under which this obstruction appears in KL.

4.1 A REPRESENTATION-LEVEL OBSTRUCTION

Let $P \in \mathbb{R}^{L\times L}$ be one causal attention matrix. For a cut $k \in \{1, \dots, L-1\}$, define the cross-cut block $B = P_{k+1:L,1:k}$. It records suffix-to-prefix attention. A d-semiseparable causal matrix has cross-cut rank at most d (Dao & Gu, 2024, Def. 3.1). If $\widetilde{B}$ is the corresponding block of any such student mixer, the Eckart & Young (1936) theorem then gives the following statement.

Proposition 1 (Cross-cut spectral obstruction)**.** *Let $\sigma_1(B) \geq \sigma_2(B) \geq \cdots$ be the singular values of B, and define $\tau_d(B) = \sum_{j>d} \sigma_j(B)^2$. Then*

$$\|P - M\|_F^2 \geq \left\|B - \widetilde{B}\right\|_F^2 \geq \tau_d(B) \tag{4}$$

for every d-semiseparable mixer M.

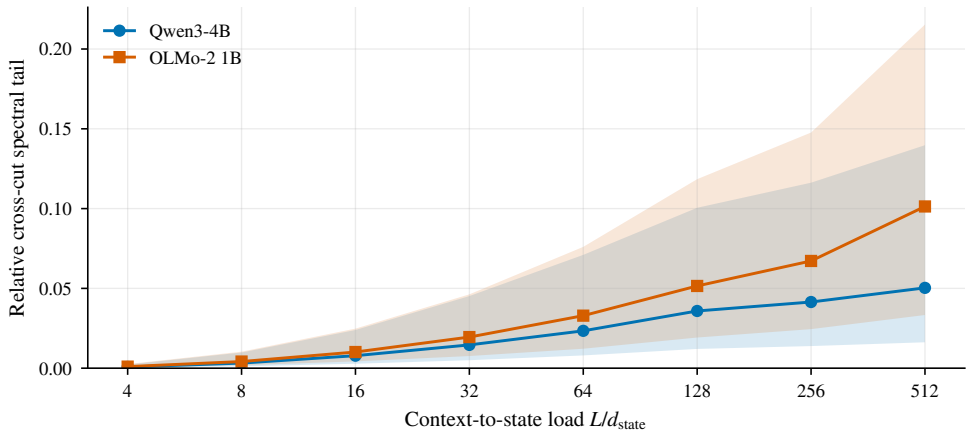

Figure 1: **Spectral evidence for a context-to-state coordinate.** For cross-cut blocks from Qwen3-4B and OLMo-2 1B attention matrices, the median relative spectral tail grows smoothly with L/d_{state}. Lines are medians across layer-head blocks after averaging the probe windows. Shading is the interquartile range, and only contexts $L \geq 512$ are shown. The broad overlap motivates L/d_{state} as a representation-level coordinate. The wide distributions and the separation at high load argue against calling it universal.

The bound applies to the attention matrix realised for a particular input. With midpoint cuts of the sequence length, increasing L enlarges the cross-cut block B, while increasing d allows a higher-rank student block. We therefore plot the relative spectral tail $\sqrt{\tau_d(B)}/\|B\|_F$ against $\rho = L/d_{\text{state}}$. The tail need not depend only on ρ, so Figure 1 tests whether this ratio is useful for organising the measurements. The two teachers trace similar median curves when plotted against ρ. The spread across heads and layers is large, so the plot supports a useful coordinate but not a one-variable law. Appendix D extends the test to ten softmax attention teacher architectures. Their effective tail exponents at $L = 2048$ range from 2.14 to 2.79. The power-law shape recurs, but the fitted exponent varies across teachers.

4.2 WHEN MATRIX ERROR IMPLIES KL ERROR

To isolate the effect of the mixer, we consider a transformer with one block and one attention head. The student replaces the teacher's attention mixer with a state-space mixer of state dimension d, drawn from the class $\mathcal{M}_d$ of d-semiseparable causal mixers. The value matrix and the computation from the head output to the logits are frozen. In this setting, the cross-cut bound gives a matrix error. We show when that error reaches the head output, changes the logits, and produces a KL loss.

Let B and $\widetilde{B}$ be the teacher and student cross-cut blocks, and let $U \in \mathbb{R}^{k\times d_v}$ contain the prefix values matrix. Their contributions to the suffix head output are $H_T = BU$ and $H_M = \widetilde{B}U$. We hold the

other contributions fixed and use the same downstream map G to obtain logits Z_T and Z_M. The loss is the teacher-to-student KL averaged over the $n = L - k$ suffix positions and the evaluation inputs.

Write $\Delta B = \widetilde{B} - B$, $\Delta H = H_M - H_T = \Delta B\, U$, and $\Delta Z = Z_M - Z_T$. We use three conditions:

1. **Residual excitation.** A change in the mixer must produce a change in the head output. For some $\nu_{\text{eff}} > 0$,
$$\mathbb{E}\big[\|\Delta H\|_F^2\big] \geq \nu_{\text{eff}}\, \mathbb{E}\big[\|\Delta B\|_F^2\big].$$
2. **Downstream observability.** On average, a change in the head output must reach the logits. Since adding the same value to every logit leaves the probabilities unchanged, Π removes that common-shift direction. For some $\mu > 0$,
$$\mathbb{E}\big[\|\Pi \Delta Z\|_F^2\big] \geq \mu^2\, \mathbb{E}\big[\|\Delta H\|_F^2\big].$$
3. **Restricted curvature.** The logit change must change the predicted probabilities. For some $\lambda_F > 0$, the categorical Fisher quadratic form along each teacher-to-student logit segment is bounded below in the observed logit-error direction.

Theorem 1 (Conditional one-head KL lower bound)**.** *Under these assumptions, every $M \in \mathcal{M}_d$ obeys*

$$\mathcal{R}_{\text{KL}}(M) \geq \frac{\lambda_F \mu^2 \nu_{\text{eff}}}{2(L-k)} \mathbb{E}\, \tau_d(P_{k+1:L,1:k})\,, \tag{5}$$

where $\mathcal{R}_{\text{KL}}(M)$ is the mean suffix KL in the controlled one-head comparison.

Appendix A gives the full derivation. The bound concerns the controlled one-head comparison, in which other pathways are held fixed. A full-model KL floor would additionally require full-logit curvature and a composite transmission coefficient η that remains positive as the selected error passes through those pathways. We do not establish that condition or a power-law rate in training data.

5 A METHODOLOGY FOR CONDITIONAL SCALING

5.1 CANDIDATE LAW AND FALSIFICATION PROTOCOL

Let N denote model size, r the replacement ratio, d the recurrent state size, L the context length and F the recurrent family. Let c denote the rest of the conversion setup: the Phase-0 and Phase-1 checkpoints, layer placement, optimiser, schedule and data order. For fixed c and evaluation distribution $\mathcal{E}$, we consider the empirical family

$$K_{\mathcal{E}}(D \mid c, N, r, d, L, F) = A_{c,\mathcal{E}} \left(\frac{D}{N}\right)^{-\alpha_{c,\mathcal{E}}}. \tag{6}$$

The arguments of A and α are suppressed. Both may depend on all the displayed conditioning variables. Equation (6) is a candidate model rather than a derived law. A natural extension adds an offset for an irreducible distillation gap. We leave it out because the observed data budgets do not constrain such a term (Section 6.2).

We evaluate this family through prediction (Bhagia et al., 2025; Choshen et al., 2025). Within one cell N is fixed, so Equation (6) reduces to $K(D) = aD^{-\alpha}$ with $a = AN^{\alpha}$. We fit this form in log space (Besiroglu et al., 2024) on early budgets, freeze the fit, and predict D64. We report error on the unseen endpoint instead of fit quality on the observed points. We also inspect the residuals by model size. A universal exponent or a joint replacement-ratio term would need to transfer to held-out sizes and ratios, which the current grid does not yet support.

5.2 EXPERIMENTAL DESIGN

The primary study converts 16-layer OLMo-1 checkpoints at 300M, 530M and 750M parameters released with DataDecide (Magnusson et al., 2025). We refer to these teachers as OLMo-1 throughout. We replace fractions $r \in \{0.25, 0.50, 0.625, 0.75, 0.875\}$ of the attention layers with Mamba-2 mixers. Replacement layers are distributed through depth rather than grouped at one end, so changing r does not introduce a single contiguous bottleneck. The primary data-scaling grid fixes $d_{\text{state}} = 64$.

Table 1: Primary experimental grid. Each sub-1B cell has a complete Phase-2 ladder on the same held-out suite. The 1B run is a cross-family check with its own fixed experimental setup. The last column gives the Phase-2 token budget of each learning-rate calibration run. At every OLMo-1 size we compared peak rates $\{3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\}$ at $r = 0.5$, and 10^{-4} was selected at all three sizes.

Teacher	Ratios r	d_{state}	Phase-2 rungs	Eval. contexts	LR calib.
OLMo-1 300M	.25,.50,.625,.75,.875	64	D1–D64	256–2048	18.9M tok.
OLMo-1 530M	.25,.50,.625,.75,.875	64	D1–D64	256–2048	33.3M tok.
OLMo-1 750M	.25,.50,.625,.75,.875	64	D1–D64	256–2048	47.1M tok.
OLMo-2 1B	.50,.75,1.0	64	D8,D16,D32	256–2048	–

Phase 2 uses sequence length 2048 and peak learning rate 10^{-4}. We call the first 18.86M-token rung D1 and each doubling D2 through D64. D64 contains about 1.21B tokens. This regular spacing in log tokens lets us test whether a power law fitted on early rungs predicts a later rung. Every rung is a fully annealed endpoint of a branching warmup–stable–decay (WSD) schedule (Hu et al., 2024; Hägele et al., 2024). For each model-size and replacement-ratio cell, one shared trunk warms up to the peak rate and then remains stable. At 90% of each rung's step count, a branch restores the full training state and decays the rate linearly over the remaining 10%. The rungs in a cell therefore share their trunk and data order. They differ only in the length of the stable phase and in their own decay (Appendix B).

We calibrated the Phase-2 peak learning rate separately at each OLMo-1 model size, using the $r = 0.5$ student and three candidate rates (Table 1). The same rate, 10^{-4}, was selected at all three sizes, and we froze it before launching the main grid. The fixed rate provides a common optimisation spine. Available compute limited the calibration runs to roughly one to two and a half D1 budgets, and since the optimal rate decreases as the training horizon grows (Bjorck et al., 2025), the selected rate may stop being optimal at the largest budgets or at other replacement ratios.

Every primary endpoint is evaluated on the same immutable suite of 256 held-out documents with 64 target tokens per document at contexts $L \in \{256, 512, 1024, 2048\}$. The suite and target positions are fixed across students. This matters because training loss and unmatched validation samples can make small late-budget differences look like changes in the scaling curve. We use the model's faithful recurrent inference path rather than a matrix surrogate at evaluation time.

The 1B study differs in teacher family, tokenizer, data mixture and training scale. We therefore treat it as a cross-family check rather than another point on the same curve. Its Phase-2 budgets are 0.50B, 1.01B and 2.01B tokens. We also ran a state-size intervention at $r = 0.5$ with $d_{\text{state}} \in \{16, 64, 256\}$.

The primary sub-1B cells use the same staged conversion setup across replacement ratios. Their Phase-1 checkpoints are ratio-specific because each replacement pattern defines a different student, but the phase budgets, optimisation recipe and checkpoint-selection rule are matched. Cross-ratio comparisons therefore share a common procedure. The non-monotonic rankings we observe are empirical evidence against assuming a simple separable law in r.

6 Results

6.1 Phase-2 trajectories differ across model sizes

Figure 2 contains the central empirical result. At $L = 2048$, direct KL falls through D64 for all five OLMo-1 300M students. The D32 to D64 reductions range from 10.8% to 20.9%. At 530M and 750M, the same additional data has a different effect. Direct KL rises by 6.4% to 17.8% in the $r = 0.25$ and $r = 0.50$ cells, while most higher replacement ratios move by only a few per cent. Excess CE follows the same broad pattern, so the late reversals are not specific to direct KL.

Lower learning-rate runs did not show the same late reversal. The calibration runs were much shorter than the D32 to D64 interval, so they could not expose this late behaviour. We retain the fixed-rate curves as the primary comparison because they follow the common optimisation setup. The lower-rate

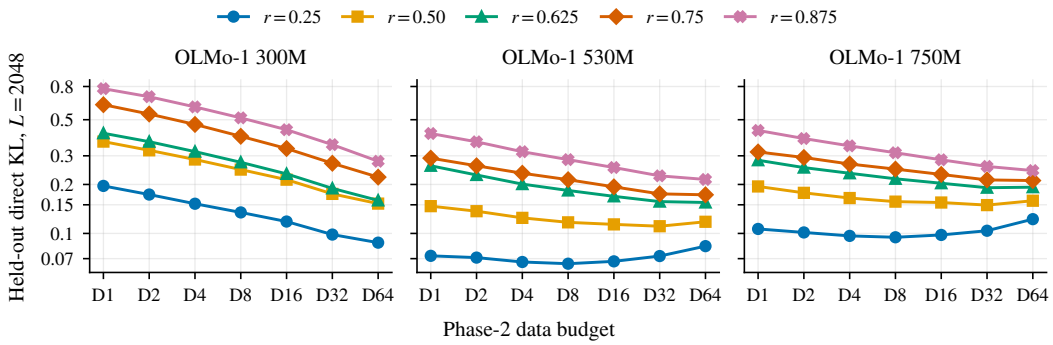

Figure 2: **Complete Phase-2 learning curves at context 2048.** Held-out direct KL is shown for 15 OLMo-1 model-size and replacement-ratio cells from D1 through D64. Direct KL falls throughout at 300M. At 530M and 750M, the low-replacement cells flatten or reverse late, while higher-replacement cells improve only modestly. All ratios use the same staged conversion setup.

runs show that the reversal depends on that setup and may reflect a budget-dependent optimal rate, consistent with the finding that longer training calls for a smaller learning rate (Bjorck et al., 2025).

This does not mean that larger models distil worse. The three sizes use different teachers, and model size was not varied as an isolated intervention. It does show that an exponent estimated at 300M cannot simply become a law that is independent of model size. The most defensible unit of prediction is the learning curve under a fixed setup.

The OLMo-2 1B cross-family check supports that narrower claim. From D8 to D32, direct KL at $L = 2048$ falls by 36.0%, 34.1% and 31.9% for replacement ratios 0.5, 0.75 and 1.0, so each ratio follows a well-behaved learning curve in a second teacher family.

6.2 Early budgets forecast a useful range

The strongest evidence for a scaling methodology is out-of-sample prediction, which we test with our data in Fig. 3. There, we fit each cell only on its early budgets. Training through D16 gives D64 RMSE 0.023 and median absolute percentage error 11.1%. Including D32 improves these values to 0.017 and 8.5%, which are small enough to help decide whether another doubling is likely to be worthwhile.

The misses also have a consistent direction: the fits under-predict D64 KL at 530M and 750M, especially at low replacement ratios, and over-predict it at 300M, so the smaller students improve more than their early trend suggests while the larger ones improve less. This points to a variable associated with model size, schedule or conversion quality that the law omits.

Fitting each complete ladder separately quantifies both the regularity and its limits (Table 7 in Appendix C.2). The 95% interval for α excludes zero in 13 of 15 cells, the exceptions being the $r = 0.25$ cells at 530M and 750M, whose KL does not decrease. The exponent rises with replacement ratio (at 300M only up to $r = 0.75$), and the $r = 0.25$ and $r = 0.875$ intervals are disjoint at every size, so students that start further from the teacher reduce KL faster in relative terms. A quadratic term in $\log D$ is significant in 13 of 15 cells, but its sign changes with model size. Every significant 300M curve steepens, whereas every significant 530M and 750M curve flattens or turns upwards. The pure power law is therefore a local description (Caballero et al., 2023), which agrees with the structured forecast errors.

We also compared pure-power, offset-power and other saturating descriptions on the observed points. Some slices prefer a positive offset descriptively, but within the budgets we selected the ladders are too short to identify a floor reliably. A positive asymptote should improve cold prediction, remain stable after removing early points and have uncertainty that excludes zero. The natural test is at 530M and 750M with $r \geq 0.75$, where the curves flatten without turning upwards: extending those ladders to D256 under a learning rate extrapolated from several budgets (Bjorck et al., 2025), with endpoints

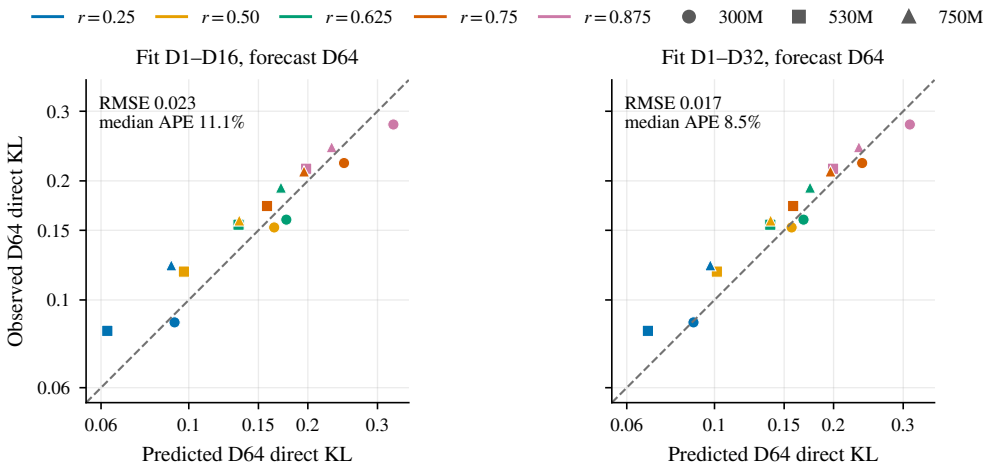

Figure 3: **Cold prediction of D64.** Each point is one sub-1B model-size and replacement-ratio cell at $L = 2048$. Pure power laws are fit only through D16 (left) or D32 (right), then used to predict the unseen D64 endpoint. The D1–D32 fits reach root mean square error (RMSE) 0.017 and median absolute percentage error 8.5%. Residuals remain structured by model size, which is evidence against a universal joint law.

declared in advance and a bootstrap over the held-out documents (Besiroglu et al., 2024), would compare offset and pure-power forecasts out of sample.

6.3 State dimension does not organise the downstream KL loss

The spectral analysis predicts a monotonic representation-level benefit from increasing d_{state} at fixed context, but the 1B end-to-end intervention does not preserve that ordering. At $r = 0.5$ and the primary D32 endpoint, the $d_{\text{state}} = 16$ student has lower held-out direct KL than the $d_{\text{state}} = 64$ and $d_{\text{state}} = 256$ students. Across the measured budgets, the three trajectories are close and do not improve monotonically with state size (Figure 4, right).

This negative result rules out using raw state size as a sufficient predictor of the final distillation gap under the current protocol, but it does not contradict Proposition 1. That proposition concerns the best cross-cut approximation of one matrix. The downstream result may reflect compensation through values, gates, residual pathways and further training. State size may also interact with optimisation. Testing the theorem's bridge therefore requires value-weighted and downstream-weighted residuals, not only a raw Frobenius tail.

6.4 Longer context reveals a larger imitation gap

The OLMo-1 endpoint audit is reported in Appendix C.4, where the D64 profiles provide the terminal cross-section. Across all 108 matched model, ratio and budget endpoints, student CE falls from $L = 256$ to $L = 2048$, while direct KL rises. Certified direct KL also rises in all 102 endpoints for which it is available. The student therefore makes better absolute predictions with more context, but captures less of the teacher's improvement.

This contrast can disappear in aggregate validation loss, which motivates our context-conditioned KL. We interpret the growing gap as evidence that the teacher uses the added context more effectively than the recurrent student.

7 Discussion

What scales, and what does not. Within a fixed setup, early-budget power laws can forecast a later Phase-2 endpoint well enough to support run planning. They do not give a universal exponent, a

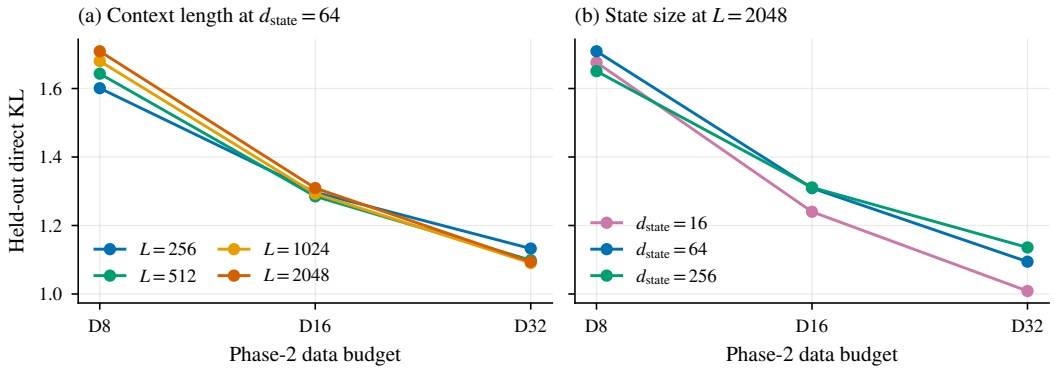

Figure 4: **OLMo-2 1B context and state-size interventions.** All curves use $r = 0.5$ and the matched D8–D32 Phase-2 budgets. The left panel compares evaluation contexts for $d_{\text{state}} = 64$. The right panel compares $d_{\text{state}} \in \{16, 64, 256\}$ at $L = 2048$. Direct KL falls with data, but larger states do not produce a monotonic ordering at D32.

monotonic effect of state size, a separable effect of replacement ratio, or an identifiable positive KL floor. These failures limit the supported forecasts.

How the spectral view can become predictive. Raw spectral tails weight all directions and downstream uses equally. For $\Delta H_{\text{cross}} = (\widetilde{B} - B)U$ and downstream logit Jacobian J, Theorem 1 motivates

$$Q = \mathbb{E}\big[(J\Delta H_{\text{cross}})^{\top}\big(\operatorname{diag}(p_T) - p_T p_T^{\top}\big)(J\Delta H_{\text{cross}})\big] \tag{7}$$

which is closer to local KL. A held-out test could fit one calibration constant on some layers or state sizes and predict the rest. Comparing cross-cut and logit residuals would also estimate η and test whether the componentwise statement extends.

Limitations and future work. The primary grid covers three sub-1B teachers and one recurrent family, without Phase-0 or Phase-1 budget ablations. The state-size intervention covers only OLMo-2 1B at $r = 0.5$. Future work should compare candidate laws on declared endpoints, calibrate longer-budget learning rates and repeat the grid with other recurrent families, including GLA and Gated DeltaNet. Value-weighted predictors and η should be tested while holding out a state size, ratio and model family.

8 Conclusion

Within a fixed setup, attention-to-recurrence distillation improves with data, but not as one law of data, context and state size. Matrix spectra expose a representation constraint, yet raw state size does not predict full-model KL. Under a fixed setup, early Phase-2 budgets forecast later quality, while context-conditioned evaluation reveals what aggregate loss hides. This provides a protocol for future studies.

Reproducibility statement. The appendix records layouts, budgets, optimisation and evaluation details. We plan to release endpoint manifests, figure source tables and scripts that rebuild the reported summaries.

Use of large language models. We used large language model (LLM) assistants in three ways. To aid and polish writing, they suggested edits to wording, structure and length throughout the paper and drafted parts of the related-work and appendix text, which the authors then revised. For retrieval and discovery, they helped search for related work and supporting references, and every citation was verified against its original source. For research ideation and execution, they assisted with the experiments and with reviewing them, and reviewed the proof and suggested improvements to it. The authors directed the research, reviewed every AI-assisted output, checked all reported numbers against the logged results, and take full responsibility for the content of the paper.

REFERENCES

AI21 Labs. Jamba: A hybrid transformer-Mamba language model. arXiv preprint arXiv:2403.19887, 2024. URL https://arxiv.org/abs/2403.19887. Accessed: 2026-05-22.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. arXiv preprint arXiv:2312.04927, 2023. URL https://arxiv.org/abs/2312.04927.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. arXiv preprint arXiv:2404.10102, 2024. URL https://arxiv.org/abs/2404.10102.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2412.04403.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. arXiv preprint arXiv:2002.07028, 2020. URL https://arxiv.org/abs/2002.07028.

Aviv Bick, Kevin Y. Li, Eric P. Xing, J. Zico Kolter, and Albert Gu. Transformers to ssms: Distilling quadratic knowledge to subquadratic models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 31788–31812. Curran Associates, Inc., 2024. doi: 10.52202/079017-0999. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3848fef259495bfd04d60cdc5c1b4db7-Paper-Conference.pdf.

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, 2025a. URL https://arxiv.org/abs/2502.14458.

Aviv Bick, Eric P. Xing, and Albert Gu. Understanding the skill gap in recurrent models: The role of the gather-and-aggregate mechanism. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=hWYisuBbp7.

Aviv Bick, Eric P. Xing, and Albert Gu. Retrieval-aware distillation for Transformer–SSM hybrids. arXiv preprint arXiv:2602.11374, 2026. URL https://arxiv.org/abs/2602.11374. Accessed: 2026-05-22.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, et al. Pythia: A suite for analyzing large language models across training and scaling. arXiv preprint arXiv:2304.01373, 2023. URL https://arxiv.org/abs/2304.01373.

Johan Bjorck, Alon Benhaim, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal LR across token horizons. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.19913.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.08606.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.14891.

Aakshita Chandiramani et al. Nemotron 3 super: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. arXiv preprint arXiv:2604.12374, 2026. URL https://arxiv.org/abs/2604.12374.

Yingfa Chen, Xinrong Zhang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Stuffed Mamba: Oversized states lead to the inability to forget. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2410.07145.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2410.11840.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? arXiv preprint arXiv:2410.23771, 2025. URL https://arxiv.org/abs/2410.23771.

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. URL https://arxiv.org/abs/2505.03005.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, et al. OLMo: Accelerating the science of language models. arXiv preprint arXiv:2402.00838, 2024. URL https://arxiv.org/abs/2402.00838.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.00396.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-policy distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.08543.

Yuxian Gu, Qinghao Hu, Shang Yang, Haocheng Xi, Junyu Chen, Song Han, and Han Cai. Jet-Nemotron: Efficient language model with post neural architecture search. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2508.15884.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.18392.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.15556.

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? In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.06654.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. arXiv preprint arXiv:2404.06395, 2024. URL https://arxiv.org/abs/2404.06395.

Sukjun Hwang, Aakash Lahoti, Tri Dao, and Albert Gu. Hydra: Bidirectional state space models through generalized matrix mixers. arXiv preprint arXiv:2407.09941, 2024. URL https://arxiv.org/abs/2407.09941.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. arXiv preprint arXiv:2402.01032, 2024. URL https://arxiv.org/abs/2402.01032.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/2006.16236.

Hyunji Lee, Wenhao Yu, Hongming Zhang, Kaixin Ma, Jiyeon Kim, Dong Yu, and Minjoon Seo. Understanding and enhancing Mamba-transformer hybrids for memory recall and language modeling. arXiv preprint arXiv:2510.26912, 2025. URL https://arxiv.org/abs/2510.26912.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. Datadecide: How to predict best pretraining data with small experiments. arXiv preprint arXiv:2504.11393, 2025. URL https://arxiv.org/abs/2504.11393.

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. URL https://arxiv.org/abs/2405.06640.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2404.08819.

Michael Poli, Armin W. Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, et al. Mechanistic design and scaling of hybrid architectures. arXiv preprint arXiv:2403.17844, 2024. URL https://arxiv.org/abs/2403.17844.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bIlnpVM4bc.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2015. URL https://arxiv.org/abs/1412.6550.

Xuyang Shen, Dong Li, Ruitao Leng, Zhen Qin, Weigao Sun, and Yiran Zhong. Scaling laws for linear complexity language models. arXiv preprint arXiv:2406.16690, 2024. URL https://arxiv.org/abs/2406.16690.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, et al. 2 OLMo 2 furious. arXiv preprint arXiv:2501.00656, 2025. URL https://arxiv.org/abs/2501.00656.

Raf Vandebril, Marc Van Barel, and Nicola Mastronardi. *Matrix Computations and Semiseparable Matrices, Volume 1: Linear Systems*. Johns Hopkins University Press, Baltimore, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

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

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M. Rush, and Tri Dao. The Mamba in the Llama: Distilling and accelerating hybrid models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2408.15237.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768, 2020. URL https://arxiv.org/abs/2006.04768.

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. URL https://openreview.net/forum?id=h3wbI8Uk1Z.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. arXiv preprint arXiv:2309.16039, 2023. URL https://arxiv.org/abs/2309.16039.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2312.06635.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.06464.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2402.04347.

Michael Zhang, Simran Arora, Rahul Chalamala, Alan Wu, Benjamin Spector, Aaryan Singhal, Krithik Ramesh, and Christopher Ré. LoLCATs: On low-rank linearizing of large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.10254.

A Proof of the conditional one-head KL lower bound

We give the complete argument used in Theorem 1. The result is intentionally componentwise. Its purpose is to separate the unconditional linear-algebra step from the assumptions needed to reach KL loss.

A.1 Set-up

Fix a sequence length L and cut k, with suffix length $n = L - k$. Let $\mathcal{M}_d$ be the class of d-semiseparable student mixers considered in the theorem. The teacher uses a causal mixing matrix $P(X)$ and a student $M \in \mathcal{M}_d$ uses the matrix $M(X)$. Let

$$B(X) = P(X)_{k+1:L,1:k}, \qquad \widetilde{B}(X) = M(X)_{k+1:L,1:k}. \tag{8}$$

By d-semiseparability, $\operatorname{rank}(\widetilde{B}(X)) \le d$. Let $U(X) \in \mathbb{R}^{k\times d_v}$ be the frozen matrix of prefix values. We isolate their contribution to the suffix outputs,

$$H_T = BU, \qquad H_M = \widetilde{B}U. \tag{9}$$

This is an interventional comparison. Contributions from suffix values, residual paths and other heads are held identical or removed rather than added to these outputs. For each input, the same frozen map

G sends the component outputs to logits Z_T and Z_M. The operator Π removes the logit direction that is invisible to softmax.

For suffix position t, define the interpolation $z_t(s) = z_{T,t} + s(z_{M,t} - z_{T,t})$ and $p_t(s) = \mathrm{softmax}(z_t(s))$. Its categorical Fisher matrix is

$$F_t(s) = \mathrm{diag}(p_t(s)) - p_t(s)p_t(s)^\top. \tag{10}$$

We write $p_{T,t} = p_t(0)$ and $p_{M,t} = p_t(1)$. The mean componentwise suffix risk is

$$\mathcal{R}_{\mathrm{KL}}(M) = \frac{1}{n}\mathbb{E}\sum_{t=k+1}^{L} D_{\mathrm{KL}}(p_{T,t} \,\|\, p_{M,t})\,. \tag{11}$$

Below, $\Delta B = \widetilde{B} - B$, $\Delta H = H_M - H_T = \Delta B\, U$, and $\Delta Z = Z_M - Z_T$. We suppress the dependence on X.

A.2 Assumptions

The three positive constants in the main statement can depend on the data distribution and $\mathcal{M}_d$. They apply throughout that class.

A1: residual excitation. For every admissible student residual,

$$\mathbb{E}\|\Delta H\|_F^2 \geq \nu_{\mathrm{eff}}\mathbb{E}\|\Delta B\|_F^2. \tag{12}$$

This rules out the case where the matrix error lies entirely in value directions that are absent from the data.

A2: downstream observability. For the component-output differences induced by admissible students,

$$\mathbb{E}\|\Pi\Delta Z\|_F^2 \geq \mu^2\mathbb{E}\|\Delta H\|_F^2. \tag{13}$$

This requires the head-output error to reach the observable logits on average.

A3: restricted Fisher curvature. For every admissible student, evaluated input and suffix position, let $\delta z_t = z_{M,t} - z_{T,t}$. Along its logit segment, assume

$$\delta z_t^\top F_t(s)\delta z_t \geq \lambda_F\|\Pi\delta z_t\|_2^2 \quad \text{for all } s \in [0,1]. \tag{14}$$

The projection is needed because adding the same scalar to every logit leaves a softmax distribution unchanged.

A.3 Proof

For the log-partition function $A(z) = \log\sum_j e^{z_j}$, categorical KL is the Bregman divergence generated by A. Taylor's theorem with integral remainder gives

$$D_{\mathrm{KL}}(p_{T,t}\|p_{M,t}) = \int_0^1 (1-s)\,\delta z_t^\top F_t(s)\delta z_t \,\mathrm{d}s. \tag{15}$$

Applying Equation (14) and integrating $1-s$ yields

$$\sum_{t=k+1}^{L} D_{\mathrm{KL}}(p_{T,t}\|p_{M,t}) \geq \frac{\lambda_F}{2}\|\Pi\Delta Z\|_F^2. \tag{16}$$

Taking expectations and applying downstream observability now gives

$$\mathbb{E}\sum_{t=k+1}^{L} D_{\mathrm{KL}}(p_{T,t}\|p_{M,t}) \geq \frac{\lambda_F\mu^2}{2}\mathbb{E}\|\Delta H\|_F^2. \tag{17}$$

By Equation (9), $\Delta H = (\widetilde{B} - B)U$. Applying residual excitation gives

$$\mathbb{E} \sum_{t=k+1}^{L} D_{\mathrm{KL}}(p_{T,t} \| p_{M,t}) \geq \frac{\lambda_F \mu^2 \nu_{\mathrm{eff}}}{2} \mathbb{E} \| \widetilde{B} - B \|_F^2. \tag{18}$$

Finally, $\widetilde{B}$ has rank at most d. The minimum squared Frobenius error of a rank-d approximation is the spectral tail, so

$$\| \widetilde{B} - B \|_F^2 \geq \sum_{j>d} \sigma_j(B)^2 = \tau_d(B). \tag{19}$$

Dividing by $n = L - k$ proves Equation (5). □

Scope and open problem. The theorem is deliberately componentwise. It does not yet give a bound on the KL of the complete models, because suffix values, residual paths, other heads and later layers can compensate for the selected cross-cut error. End-to-end training may also change the values or representations enough to avoid the measured residual.

These gaps are measurable. One can compare the energy in $(\widetilde{B} - B)U$ with the change in the actual suffix logits, giving an empirical transmission coefficient η. Studying how η behaves across checkpoints, layers and state sizes would show where the spectral obstruction remains visible and where the model routes around it. This is the natural next step towards a full-model theory, and we hope the componentwise result provides a useful starting point.

B Experimental details

B.1 Replacement layouts

All primary sub-1B teachers have 16 layers. Blue A cells denote attention layers, and orange S cells denote Mamba-2 replacements. The layouts distribute attention through depth:

r	Layer, bottom to top															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0.25	A	A	A	S	A	A	A	S	A	A	A	S	A	A	A	S
0.50	A	S	A	S	A	S	A	S	A	S	A	S	A	S	A	S
0.625	A	S	A	S	S	A	S	S	A	S	A	S	S	A	S	S
0.75	A	S	S	S	A	S	S	S	A	S	S	S	A	S	S	S
0.875	A	S	S	S	S	S	S	S	A	S	S	S	S	S	S	S

The model widths are 1024, 1344 and 1536 for 300M, 530M and 750M. Each model uses 16 attention heads. The Mamba-2 configuration has state size 64, convolution width 4 and scan chunk length 128 for the primary data ladder.

B.2 Teacher architectures

Both teacher families are decoder-only Transformers with 16 layers, 16 attention heads, SwiGLU MLPs, RMSNorm, rotary position embeddings and untied input and output embeddings (Table 2), so the attention layers we replace have the same form, Equation (2), in both. The architectural differences are modest: OLMo 2 moves the normalisation from the inputs to the outputs of each attention and MLP sublayer, adds RMSNorm on queries and keys, and raises the RoPE base from 10^4 to 5×10^5 (Team OLMo et al., 2025). These changes are why we chose OLMo 2 for the cross-family check. The DataDecide suite, including its 1B checkpoint, shares the earlier OLMo 1 design, and among the fully open families that release data and checkpoints at several sizes, OLMo 2 is the one whose architecture follows current practice, whereas Pythia (Biderman et al., 2023) and OLMo 1 predate these changes. The larger differences, however, lie in how the teachers were trained. The DataDecide checkpoints use the OLMo 1 codebase and its GPT-NeoX-derived tokenizer (Groeneveld et al., 2024), a 2048-token context and about 100 training tokens per parameter on Dolma 1.7 (Magnusson et al., 2025). OLMo-2 1B instead uses a tokenizer built on the cl100k vocabulary, a 4096-token context and

about 4T pretraining tokens followed by a mid-training stage on Dolmino-mix-1124. These choices change the distribution the student imitates, the tokens over which KL is measured and the attention patterns Phase 0 must match, which is why we treat the 1B study as a cross-family check rather than as another point on the OLMo-1 curves.

Table 2: Teacher architectures and training regimes. OLMo-1 values are shared by the three DataDecide sizes except where listed per size.

	OLMo-1 (DataDecide)	**OLMo-2 1B**
Layers / heads	16 / 16	16 / 16
Width (head dimension)	1024, 1344, 1536 (64, 84, 96)	2048 (128)
MLP	SwiGLU	SwiGLU
RMSNorm placement	sublayer inputs	sublayer outputs
QK normalisation	none	RMSNorm
Positional encoding	RoPE, $\theta = 10^4$	RoPE, $\theta = 5 \times 10^5$
Tokenizer (vocabulary)	GPT-NeoX-derived (50,280)	cl100k-based (100,352)
Context length	2048	4096
Pretraining data	Dolma 1.7	OLMo-mix-1124 + Dolmino
Pretraining tokens	about 100 per parameter	about 4T

B.3 Phase-2 budgets and optimisation

Table 3 summarises the training setup of the OLMo-1 grid and the OLMo-2 1B cross-family check, and Table 4 lists the budget and schedule of every Phase-2 rung.

Table 3: Training setup for the primary OLMo-1 grid and the OLMo-2 1B cross-family check.

	OLMo-1 grid (300M, 530M, 750M)	**OLMo-2 1B**
Teacher	DataDecide OLMo-1 checkpoints	OLMo-2-0425-1B
Distillation data	Dolma 1.7	OLMo-mix-1124
Phase-0 / Phase-1 tokens	size-specific (Table 5)	133.3M / 266.7M
Phase-2 endpoints	D1–D64 (18.9M–1.21B tokens)	0.50B, 1.01B, 2.01B tokens
Sequence length	2048	2048
Batch size (sequences)	8	128
Peak learning rate	10^{-4}	10^{-4}
Schedule	branching WSD (Table 4)	WSD

Table 4: Phase-2 rungs of the OLMo-1 grid. Each rung branches from the shared trunk at the listed step and decays over its final 10% of steps. One optimiser step processes 16,384 tokens.

Rung	**Tokens**	**Total steps**	**Branch step**	**Decay steps**
D1	18,857,984	1,151	1,036	115
D2	37,715,968	2,302	2,072	230
D4	75,431,936	4,604	4,144	460
D8	150,863,872	9,208	8,288	920
D16	301,727,744	18,416	16,575	1,841
D32	603,455,488	36,832	33,149	3,683
D64	1,206,910,976	73,664	66,298	7,366

Branching schedule. The WSD schedule (Hu et al., 2024) lets a single constant-rate trunk serve every budget, and cooldowns branched from such a trunk scale like cosine schedules matched to each training length (Hägele et al., 2024). It also lets every rung of a trajectory enter the fit, which improves scaling-law estimates (Choshen et al., 2025). For each (N, r) cell, a single trunk warms up linearly from zero to the peak rate over 57 optimiser steps and then holds the rate constant. At each branch step, the trunk saves the full training state: weights, optimiser moments, scheduler, data-loader position and random-number state. The endpoint resumes from that state and decays the rate linearly to 10^{-8}. The D64 trunk continues the D32 trunk from its final state. Each endpoint therefore matches the terminal decay of an uninterrupted WSD run of its own length with the same warmup. All rungs in a cell share the trunk and its data order, so differences between neighbouring rungs, including the D32 to D64 changes in Section 6, are paired comparisons within one trajectory.

Learning-rate calibration. Before launching the grid, we calibrated the Phase-2 peak rate at each OLMo-1 size with the $r = 0.5$, $d_{\text{state}} = 64$ student, the settings in Table 3, and candidate rates $\{3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\}$. The calibration runs lasted 1,151, 2,033 and 2,877 optimiser steps at 300M, 530M and 750M (Table 1 gives the corresponding tokens). We selected the rate by the reduction of the training distillation objective without optimisation instability. The rate 10^{-4} was selected at all three sizes, and we froze it across model sizes, ratios and data budgets. Compute constraints limited calibration to one replacement ratio and short budgets, so some cells may lie away from their individual optimum, particularly late in the ladder. Follow-up runs at lower learning rates did not show the same late reversal. We keep them separate from the main fits because they no longer share the common learning-rate spine.

Earlier phases and the 1B cross-family check. Table 5 gives the budgets and peak rates used before Phase 2. Within each model size, these settings are fixed across replacement ratios. Both phases use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, weight decay 0.1 and gradient clipping at 1.0. The WSD schedule uses 5% warmup, 10% terminal decay and a minimum rate of 10^{-8}. Phase 0 minimises the mean Frobenius norm between student and teacher mixer matrices over replaced layers, while Phase 1 minimises mean tokenwise hidden-state ℓ_2 error. The terminal checkpoint from each phase initialises the next phase.

Table 5: Phase-0 and Phase-1 settings for the primary OLMo-1 grid. Budgets and rates are fixed across replacement ratios within a model size.

Teacher	Phase-0 tokens	Phase-1 tokens	Phase-0 LR	Phase-1 LR
OLMo-1 300M	11,436,032	11,436,032	3×10^{-2}	10^{-3}
OLMo-1 530M	10,108,928	20,217,856	10^{-2}	3×10^{-4}
OLMo-1 750M	14,286,848	28,573,696	10^{-2}	3×10^{-4}

The OLMo-1 teachers are the 300M, 530M and 750M Dolma 1.7 checkpoints of DataDecide (Magnusson et al., 2025), and the cross-family teacher is OLMo-2-0425-1B (Team OLMo et al., 2025). All teacher checkpoints and the Dolma snapshot are pinned to fixed Hugging Face revisions, which we release with the evaluation manifests. The OLMo-2 1B check keeps the same Phase-2 peak rate. Its different data, batch size and schedule are why we do not pool it with the OLMo-1 scaling fits.

B.4 Evaluation suite

The canonical sub-1B source table contains 432 measured points. The complete-ladder analysis uses 420 of them: 15 (N, r) trajectories, seven Phase-2 budgets and four context lengths. The remaining 12 points are endpoint-only $r = 1$ diagnostics at D64 and are not used in the power-law fits. For each context, the evaluator uses 256 fixed documents and 64 fixed target tokens per document. All compared endpoints share the same suite manifest. Teacher logits are produced by the matched teacher on exactly the same tokens.

The manifest draws from the same pinned Dolma 1.7 snapshot. We use eight tail shards, one assigned to each rank in the original data layout. In each shard we scan the first 512 documents containing at

least 16,384 characters and retain the 32 with the lowest SHA-256 hashes. This gives 256 documents through a deterministic, outcome-independent rule. The selected shards lie beyond the data consumed by every run in the grid, including D64.

We draw one window per document at a deterministic position. The four context lengths end at the same point and share the same 64 target tokens, so each shorter context is a suffix of the longer ones and only the amount of preceding text changes. This yields 256 windows and 16,384 target tokens at every context length, and all reported metrics are computed on these contiguous windows. For the certified diagnostics below, the suite also stores counterfactual copies in which the context beyond the most recent 128 tokens is replaced by spans from three other documents of the same source. Tokenisation uses the matched teacher tokenizer.

For a document i and target position t, direct KL is computed from the full teacher and student next-token distributions. We average first over target tokens and then over documents. Student and teacher ground-truth cross-entropies use the same targets. The evaluator also records certified quantities, which weight each target position by the teacher's mean increase in loss when its distant context is replaced by the donor spans, clipped at zero. The weights are computed once at $L = 2048$ and reused at every context length, so certified metrics emphasise targets whose teacher prediction depends on genuine distant context. The main paper uses direct KL because it is available consistently across the full grid.

C Additional results

C.1 D64 forecast table

Table 6: Cold D64 prediction aggregated over the 15 primary cells at $L = 2048$. Fits are pure power laws and do not use D64.

Fit range	RMSE	Median absolute percentage error
D1–D16	0.023	11.1%
D1–D32	**0.017**	**8.5%**

The 300M predictions usually lie above the observed D64 direct KL, meaning that the students improve more than the early law predicts. The 530M and 750M predictions usually lie below the observation. This sign structure is why a pooled error alone is insufficient evidence for cross-size universality.

C.2 Per-cell exponents and curvature

Table 7 fits each complete D1–D64 ladder at $L = 2048$ separately. The exponent comes from ordinary least squares of $\log K$ on $\log D$. The curvature c is the coefficient of x^2 in a quadratic fit of $\log K$ on $x = \log_2 D - \overline{\log_2 D}$. Negative c means the log–log curve steepens with more data, and positive c means it flattens or turns upward. Intervals are 95% t intervals with five degrees of freedom, and p-values for $c = 0$ are two-sided with four. Both treat the seven rungs of a cell as independent observations. The rungs share a training trunk (Appendix B), so the uncertainty is conditional on one trajectory and is likely understated.

C.3 Late-budget changes

From D32 to D64, the five 300M cells improve in direct KL by 10.8%, 12.9%, 15.6%, 17.8% and 20.9% as r increases. At 530M the corresponding changes are -15.1%, -6.5%, 1.3%, 1.3% and 5.0%. At 750M they are -17.8%, -6.4%, -0.7%, 0.9% and 5.7%. Positive values denote improvement. The pattern is gradual enough that no single D64 outlier explains it.

Table 7: Pure power-law exponents and curvature for the 15 primary cells at $L = 2048$, fitted on D1–D64. c is reported multiplied by 10^3.

Size	r	α	95% interval	$c \times 10^3$	$p(c = 0)$
300M	0.25	0.195	[0.182, 0.209]	−2.9	0.168
	0.50	0.214	[0.197, 0.231]	−5.1	0.037
	0.625	0.231	[0.210, 0.253]	−7.2	0.003
	0.75	0.248	[0.226, 0.270]	−7.6	<0.001
	0.875	0.245	[0.212, 0.279]	−11.5	<0.001
530M	0.25	−0.024	[−0.082, 0.034]	+19.4	0.002
	0.50	0.061	[0.022, 0.100]	+12.9	0.004
	0.625	0.128	[0.097, 0.159]	+10.4	<0.001
	0.75	0.131	[0.110, 0.152]	+5.8	0.054
	0.875	0.162	[0.144, 0.179]	+5.3	0.031
750M	0.25	−0.025	[−0.085, 0.036]	+20.4	0.001
	0.50	0.052	[0.016, 0.089]	+12.2	0.002
	0.625	0.096	[0.071, 0.120]	+8.0	0.005
	0.75	0.103	[0.086, 0.120]	+4.9	0.043
	0.875	0.139	[0.127, 0.151]	+3.9	0.009

C.4 Context profiles across the OLMo-1 grid

Figure 5 normalises the D64 direct KL of each OLMo-1 endpoint by its value at $L = 256$, so each curve shows how the imitation gap grows as the same targets receive more context.

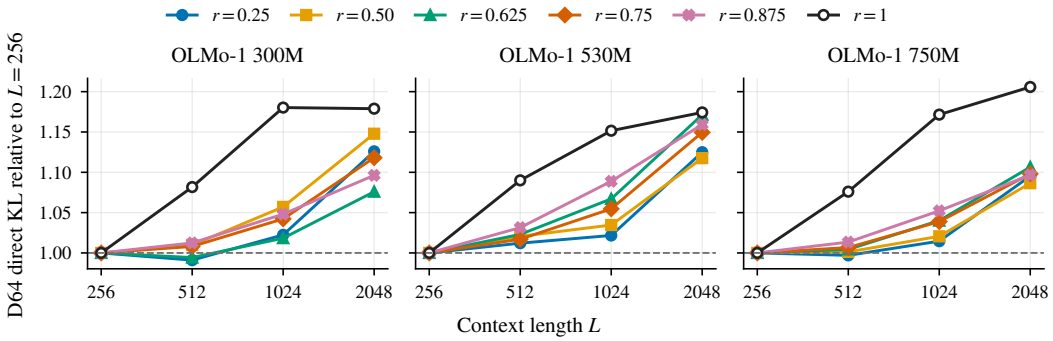

Figure 5: **Context profiles at D64.** Direct KL at each context is normalised by its value at $L = 256$ for the same OLMo-1 endpoint. Longer context generally exposes a larger teacher–student gap, especially at high replacement ratios. The $r = 1$ lines are endpoint-only diagnostics and are not part of the 15 learning-curve fits.

D Spectral diagnostics across softmax attention teachers

The main text uses Qwen3-4B and OLMo-2 1B to introduce the spectral coordinate. Here we test whether the same structure appears across a broader collection of softmax attention teachers. Table 8 lists the ten checkpoints. Each probe covers every layer and returned attention head at $L \in \{128, 256, 512, 1024, 2048\}$ and $d_{\text{state}} \in \{4, 8, 16, 32, 64, 128, 256\}$.

All probes use the WikiText-2 raw training split. We concatenate its non-empty text fields, retain the first 200,000 characters and tokenise that common text with each teacher's own tokenizer. At context length L, inputs are contiguous, non-overlapping token windows beginning at offsets $0, L, 2L, \ldots$. The number of windows is fixed within a checkpoint. The two main-text probes use four windows for

Table 8: Softmax attention teacher checkpoints used in the spectral study. The number of layer-head blocks is the product of the two preceding columns.

Checkpoint	Layers	Heads	Blocks
OLMo-2-0425-1B	16	16	256
OLMo-2-1124-7B	32	32	1,024
OLMo-2-1124-13B	40	40	1,600
OLMo-2-0325-32B	64	40	2,560
Qwen3-4B	36	32	1,152
Qwen2.5-7B	28	28	784
Pythia-6.9B	32	32	1,024
BLOOM-7.1B	30	32	960
Phi-2	32	32	1,024
Mistral-7B-v0.1	32	32	1,024

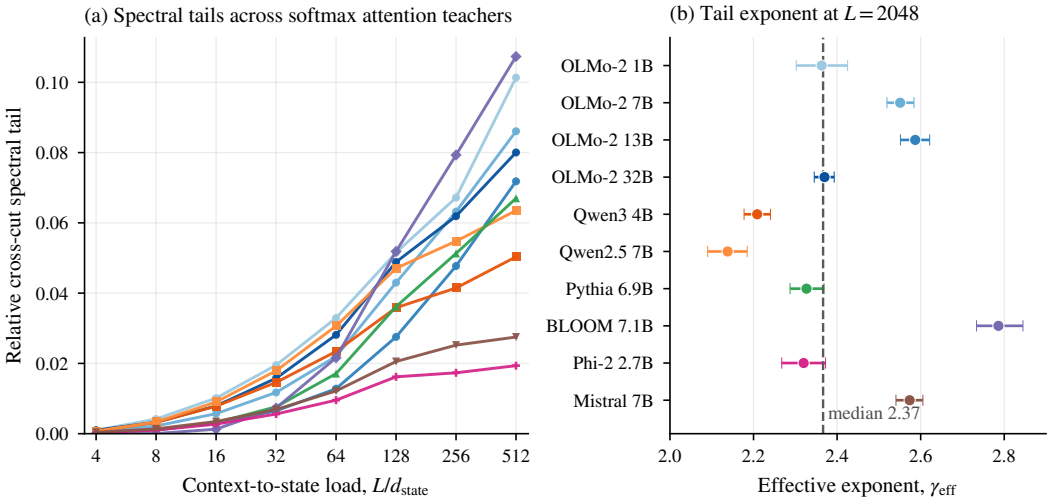

Figure 6: **Spectral diagnostics across softmax attention teachers.** (a) The median relative cross-cut spectral tail rises with L/d_{state} for all ten teacher checkpoints. Colours identify the checkpoints listed in panel b. (b) Effective tail exponents fitted at $L = 2048$. Bars are 95% layer-head block-bootstrap intervals. The estimates range from 2.14 to 2.79, with a median of 2.37. The recurring range supports a shared representation-level pattern, while the between-teacher differences argue against a single universal exponent.

OLMo-2 1B and two for Qwen3 4B. Models run in evaluation mode with eager attention. We cast the returned attention matrices to FP32 before computing singular values.

For a token window x, layer ℓ, head h and context length L, we extract the causal attention matrix $A^{(\ell,h)}(x; L)$. We cut it at the sequence midpoint and retain the block $B = A^{(\ell,h)}_{L/2+1:L,\,1:L/2}$, which records how positions in the second half use positions in the first. If $\sigma_j(B)$ are its singular values, the relative rank-d tail is

$$\phi_{\ell,h}(L, d) = \frac{\sqrt{\sum_{j>d} \sigma_j(B)^2}}{\|B\|_F}. \tag{20}$$

This is the relative Eckart–Young residual. It lower-bounds the Frobenius error of any causal mixer whose cross-cut rank is at most d. The probe is training-free: it computes the optimal rank-constrained residual directly from the teacher matrix. The stored value for each layer-head block is averaged over its probe windows before aggregation.

The study tests two representation-level claims used in Section 4. First, it asks whether the dimensionless load L/d_{state} organises ϕ across different context lengths and state sizes. Second, it asks whether the tail has a similar power-law shape across teacher architectures. Figure 6a groups observations with the same L/d_{state} and plots their median across layer-head blocks for $L \geq 512$. For panel b, we first average ϕ across layer-head blocks at $L = 2048$, then fit $\log \phi = a + s \log d_{\text{state}}$. If discarded spectral energy follows $\sum_{j>d} \sigma_j^2 \propto d^{1-\gamma}$, the fitted norm slope gives $\gamma_{\text{eff}} = 1 - 2s$. We use 1,000 block-bootstrap resamples of complete layer-head curves and report the 2.5th and 97.5th percentiles. These intervals measure variation across the probed blocks.

Nine estimates fall between 2.14 and 2.59, while BLOOM is higher at 2.79. The OLMo-2 size ladder is also not monotonic in model size. The result shows that the spectral motivation is not peculiar to the two teachers in Figure 1. It supports a recurring teacher-side representation pattern, while the conditional theorem and the end-to-end experiments address the separate question of how that pattern reaches KL.