LEARNING LESS IS MORE: PREMATURE UPPER-LAYER ATTENTION SPECIALIZATION HURTS LANGUAGE MODEL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify *premature upper-layer attention specialization* as a failure mode in GPT-style pretraining: upper attention becomes sharp and causally relied upon while lower-layer routing features are still developing. This timing mismatch occurs within the decoder hierarchy, where lower layers build the representations over which upper layers attend. Temporarily reducing the learning rate of upper-layer query and key projections delays specialization while leaving all other parameter groups on their normal schedules. Across 270M, 700M, and 1B decoders trained on 2.5B, 7B, and 30B tokens, respectively, the intervention improves final validation loss. At 270M, it improves three-run average perplexity by 0.50, reaches the baseline final loss with 13.17% fewer training tokens, and improves the seven-task downstream average by 0.41 percentage points. Mechanistic measurements show that lower routing develops on schedule while upper attention sharpening and strong early Q/K dependence are delayed. A pathwise analysis relates the upper-Q/K learning-rate multiplier to the local attention-logit update. Together, these results identify the timing of upper-layer attention learning as a concrete determinant of pretraining quality.

1 INTRODUCTION

Transformer pretraining is usually described through scale, data, and optimizer schedules (Vaswani et al., 2017; Brown et al., 2020; Kaplan et al., 2020; Hoffmann et al., 2022). Within a decoder, however, different computations develop at different times. Lower layers construct a residual basis, and upper layers use attention to compare, retrieve, and recombine information in that basis. Upper-layer query/key projections determine which past positions the upper network treats as relevant. Their learning schedule therefore controls when upper attention commits to matching patterns over representations that are themselves still changing.

We study a failure mode in GPT-style decoder blocks that we call ***premature upper-layer attention specialization***: upper-layer attention becomes sharp and causally relied upon before lower-layer routing and copy-like features have stabilized. Early in pretraining, lower layers are still building routing and copy-like features associated with induction behavior (Olsson et al., 2022; Edelman et al., 2024). If upper-layer Q/K logits grow too quickly, upper attention can become low-entropy on a moving, immature representation. Once a row of attention assigns most probability mass to one key, the softmax gradient available to raise alternative keys is proportional to their small probabilities. A wrong or noisy early match can therefore persist as an optimization attractor.

We address this failure mode with a **selective learning intervention**: reduce the early learning rate only for upper-half Q/K, then restore it as lower routing develops. The intervention leaves lower Q/K, values, FFNs, embeddings, and all other parameter groups on the normal optimizer schedule. It improves final validation loss across 270M, 700M, and 1B decoders trained on 2.5B, 7B, and 30B tokens, respectively. In the three-run 270M comparison, it also improves perplexity, token efficiency, and the average score on a seven-task downstream suite.

The **mechanism evidence** connects these gains to a change in upper attention development. Lower routing reaches maturity on schedule, while the intervention delays upper entropy collapse, suppresses early upper Q/K logit growth, reduces early concentration of the upper-Q/K bilinear form, and sharply

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

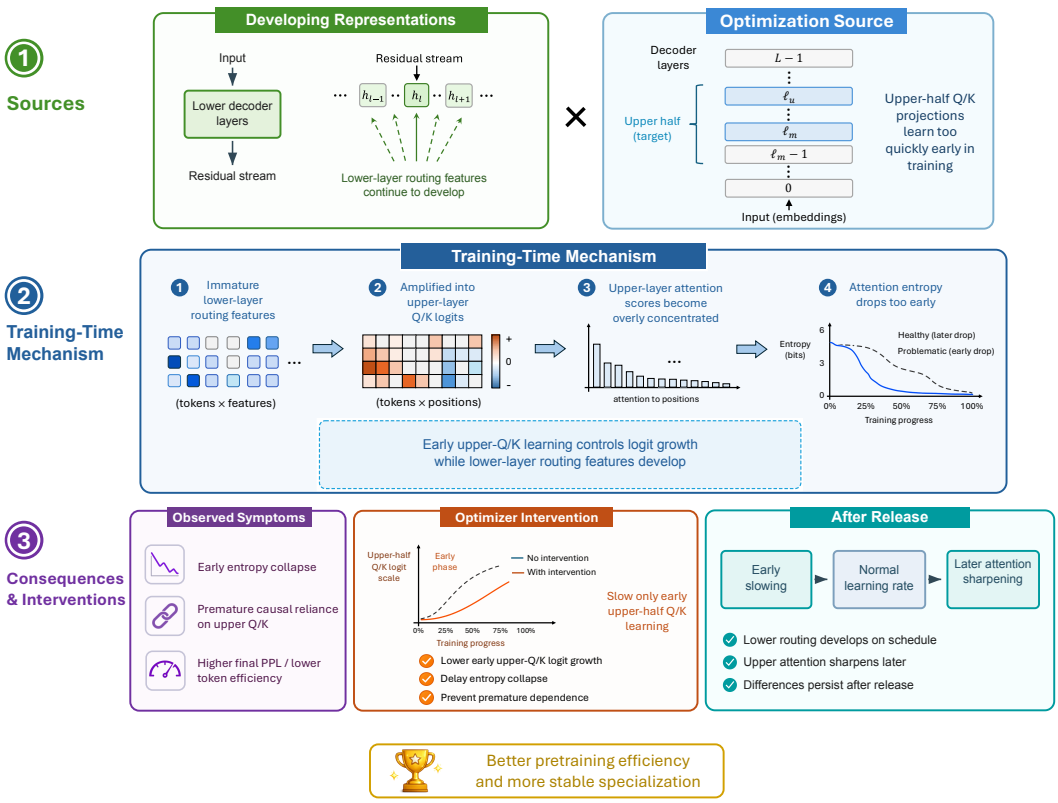

Figure 1: Premature upper-layer attention specialization and the learning-rate intervention. Lower-layer routing features are still developing when rapid early upper-Q/K learning produces concentrated attention logits. Attention sharpens and becomes causally important early in training. Selectively slowing upper-half Q/K delays this process while the remaining parameters follow their normal schedules. After release, upper Q/K returns to the normal learning rate, and the difference in attention development persists. Curves illustrate the mechanism schematically; Figure 2 reports the measured trajectories and causal ablations.

reduces early causal dependence on upper Q/K. The difference persists after the learning rate returns to normal: upper attention sharpens later, and the final model remains less sensitive to upper-Q/K ablation. A brief early intervention therefore changes the subsequent learning trajectory while preserving lower-layer development.

Figure 1 summarizes the intervention and its training-time effect. A pathwise analysis formalizes how the upper-Q/K learning rate controls the local logit update. The multiplier sweep, fixed-release comparisons, and training-time probes connect this local analysis to the observed changes in specialization and final model quality.

Our contributions are summarized as follows:

- We identify **premature upper-layer attention specialization** in GPT-style decoder pretraining. Attention sharpens and acquires strong causal importance while lower-layer routing features are still developing; selective early slowing separates their development.
- We introduce a **selective learning-rate intervention** that improves final validation loss across 270M, 700M, and 1B models. At 270M, it improves three-run average perplexity by 0.50, saves 13.17% of the training tokens needed to reach the baseline final loss, and improves the seven-task downstream average by 0.41 percentage points.
- We connect the intervention to attention development through **training-time measurements and pathwise analysis**. Routing, entropy, logit, bilinear-form, and causal-ablation measure-

ments track the altered trajectory. The analysis relates the learning-rate multiplier to local logit growth, and fixed-release controls retain most of the gain in the 270M comparison.

2 Related Work

Decoder pretraining and training-time analysis. Causal decoder pretraining is the standard setting for studying language-model scaling, from the Transformer and GPT-style decoders to GPT-3, Megatron-LM, PaLM, OPT, BLOOM, LLaMA, and LLaMA 2 (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020; Shoeybi et al., 2019; Chowdhery et al., 2023; Zhang et al., 2022; BigScience Workshop et al., 2022; Touvron et al., 2023a;b). Scaling laws characterize loss as a function of model size, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022); Pythia, OLMo, and PolyPythias make training-time development observable through open checkpoints, data, code, and repeated runs (Biderman et al., 2023; Groeneveld et al., 2024; van der Wal et al., 2025). We take this developmental view inside the decoder block: upper attention can become confident before lower copy and routing features are ready.

Transformer architecture, normalization, and stability. Transformer stability depends on normalization placement, residual scaling, and positional encoding. LayerNorm and RMSNorm are standard (Ba et al., 2016; Zhang & Sennrich, 2019); PreNorm, ScaleNorm, NormFormer, DeepNorm, and related analyses study how normalization and residual parameterization affect depth and optimization stability (Nguyen & Salazar, 2019; Xiong et al., 2020; Shleifer et al., 2021; Wang et al., 2024; Liu et al., 2020). RoPE is now common in decoder LMs (Su et al., 2024), while broad component studies show that Transformer modifications do not transfer uniformly across implementations (Narang et al., 2021). We study the training-time development of upper attention under a fixed decoder architecture.

Mechanistic development of attention and routing. Transformer-circuits work studies attention heads and residual streams as compositional circuits (Elhage et al., 2021). Induction heads emerge during training and support copy-like routing behavior (Olsson et al., 2022; Edelman et al., 2024); pruning, probing, and causal interventions identify which attention heads matter for model behavior (Michel et al., 2019; Voita et al., 2019; Clark et al., 2019; Wang et al., 2022). This motivates our lower-routing readout and our causal upper-Q/K ablation. The new finding is timing: upper Q/K can become sharp and causally relied upon before lower routing has stabilized.

Attention entropy, Q/K logits, and closest interventions. Sharp attention is useful when it is mature; the failure here is premature sharpness. Prior work links low attention entropy and large logits to Transformer instability: entropy collapse motivates spectral reparameterization, QKNorm normalizes query/key activations, and ViT-22B uses QK normalization as part of a stable scaling recipe (Zhai et al., 2023; Henry et al., 2020; Dehghani et al., 2023). Most closely, attention-logit change can be controlled with query/key-specific learning rates (Anson & Aitchison, 2025). Our intervention targets the timing of upper-layer attention learning: it slows only early upper-half Q/K, preserves lower-layer learning, and relates the resulting specialization trajectory to final pretraining quality.

3 Premature Upper Attention Specialization

3.1 Training-Time Specialization

For an upper decoder layer, let $X \in \mathbb{R}^{n\times d}$ be the layer-normalized hidden matrix for a sequence of length n. For one head,

$$Q = XW_Q, \qquad K = XW_K, \qquad Z = \frac{QK^\top}{\sqrt{d_k}} = \frac{XW_QW_K^\top X^\top}{\sqrt{d_k}}, \tag{1}$$

where Z is the pre-softmax attention-logit matrix. Define $B = W_QW_K^\top$. Upper Q/K specialization is the growth of a structured bilinear form $XBX^\top$ that makes upper attention selective over positions.

Specialization is useful when the residual basis X already contains stable routing features. It is harmful when a component of X is immature. We refer to these moving components as immature

residual directions. Formally, let P be an orthogonal projector onto residual directions whose lower-layer routing features have not yet stabilized, and let $X_P = XP$. If upper Q/K learns quickly on X_P, the model can form confident attention matches on features that later move or disappear. We measure this event with upper attention logit magnitude, upper attention entropy, upper/lower logit ratio, and causal dependence on upper Q/K.

3.2 Selective Upper-Q/K Learning

We use one selective intervention to expose and correct the failure. During the early window, the intervention multiplies the learning rate of upper-half W_Q and W_K by 0.25 while leaving lower Q/K, values, FFNs, embeddings, and all other optimizer behavior unchanged.

Release schedule. The main experiments use the same release rule: the lower-copy score, defined as the lower-half attention mass assigned to the nearest previous occurrence of the same token and averaged over layers, heads, and valid repeated-token positions, must be at least 0.005 for three consecutive evaluations. The earliest release is at 3% of training; if the condition has not fired by 12%, release is forced. After release, the multiplier ramps from 0.25 to 1.0 over 1% of total steps. Section 5.6 compares this rule with fixed early release times; Appendix D examines its threshold and patience settings.

4 Experimental Setup

270M decoder pretraining. The 270M experiments use a 20-layer causal decoder with width 960, 15 attention heads, sequence length 1024, RoPE position encoding, tied input/output embeddings, and roughly 270M parameters. These runs use 2.5B FineWeb-Edu tokens (Penedo et al., 2024) and the same GPT-2 tokenizer, packed-token data pipeline, AdamW optimizer family, cosine schedule, batch geometry, and evaluation protocol. The GPT-style baseline uses LayerNorm, biased linear projections, and a GELU FFN of hidden width 3840.

Larger-scale pretraining. We also evaluate the intervention in a 700M decoder trained for 7.0B tokens and a 1B decoder trained for 30B tokens. Table 1 brings together the pretraining results at all three scales. Appendix E gives the 700M configuration and training-time measurements.

Intervention comparisons. At each model scale, the control and intervention use the same architecture and training protocol; only the early upper-Q/K learning-rate schedule changes. Difference rows report intervention minus control. Where a control experiment reports PPL reduction, it is defined as control perplexity minus intervention perplexity, so positive values denote an improvement.

Sensitivity and alternative interventions. The multiplier sweep and fixed-release experiments each use one seed in the 270M decoder, with the data and training budget held fixed within each comparison. Global learning-rate reduction and entropy regularization are evaluated against their respective controls. Their tables report individual-run results separately from the three-run averages in Table 1.

Mechanism measurements. We record upper attention logit magnitude, upper attention entropy, upper/lower logit ratio, lower copy-routing maturity, and causal sensitivity to zeroing upper Q/K. Entropy is computed per attention row, divided by $\log \max(N, 2)$ for N available causal keys, and averaged over positions, heads, and upper layers. The sharp-attention event is the first measured checkpoint at which this normalized entropy falls below 0.75. Logit magnitude averages absolute pre-softmax logits over causal query–key pairs. These readouts track concentration and scale along the training trajectory.

The causal probe zeros upper-half Q, K, or both at evaluation time. Each operation makes the affected attention rows uniform over the available causal keys, while preserving the value and output projections. We report the increase in PPL relative to the unmodified checkpoint. This measures prediction dependence on learned query–key matching. The three operations give the same upper-wide causal readout because zeroing either factor sets their dot product to zero.

Table 1: GPT-style pretraining across model scales and token budgets. The intervention slows only upper-half Q/K in the early window. Results at 270M and 700M are mean ± sample standard deviation across three full training runs, with paired-seed differences. Difference rows report intervention minus control. Target loss is the final control loss at each scale. For 1B, PPL is calculated as the exponential of validation loss; dashes denote inapplicable entries.

Params	Training tokens	Method	Loss	PPL	Tokens to target loss	Token saving
270M	2.5B	Control	3.2883 ± 0.0021	26.7981 ± 0.0560	2.500B	–
		Intervention	3.2696 ± 0.0051	26.3014 ± 0.1344	2.171 ± 0.035B	329.4M / $13.17 \pm 1.41\%$
		Difference	-0.0187 ± 0.0030	-0.4967 ± 0.0789	-0.329 ± 0.035B	–
700M	7.0B	Control	2.9028 ± 0.0003	18.2287 ± 0.0088	7.000B	–
		Intervention	2.8958 ± 0.0004	18.1013 ± 0.0119	6.461 ± 0.017B	0.5B
		Difference	-0.0071 ± 0.0004	-0.1274 ± 0.0073	-0.539 ± 0.017B	–
1B	30B	Control	2.6520	14.1828	30.000B	–
		Intervention	2.6474	14.1177	28.482B	1.5B / 5.06%
		Difference	−0.0046	−0.0651	−1.518B	–

5 Results

5.1 Pretraining Gains Across Model Scales

At 270M, Table 1 shows that the upper-Q/K learning-rate intervention improves three-run average final validation loss by 0.0187 and average perplexity by 0.4967. It also reaches the matched baseline final loss after 2.171B tokens on average, saving 329.4M tokens out of the 2.5B-token budget.

At 700M and 7.0B training tokens, the intervention improves three-run average final validation loss by 0.0071 and perplexity by 0.1274, saving 0.539B same-loss tokens on average. These results are included in Table 1; Appendix E gives the model configuration and training-time measurements.

At 1B and 30B training tokens, the intervention reduces final validation loss by 0.0046 and reaches the baseline final loss after 28.482B tokens, saving 1.518B tokens or 5.059% of the training budget (Table 1).

5.2 Downstream Benchmark Evaluation

Table 2 reports the downstream evaluation for the same 270M GPT-style intervention comparison. The upper-Q/K learning-rate intervention improves the three-run average score by 0.41 percentage points and wins 14 of 21 run-task comparisons. The gains are largest on LAMBADA, RACE, and ARC-Easy, while ARC-Challenge is nearly flat. Together with the validation and token-efficiency results, this shows that reducing premature upper-attention specialization improves the trained base model's general evaluation behavior.

5.3 Early Causal Dependence on Upper Q/K

The GPT-style control depends strongly on upper Q/K by 3% of training. Zeroing upper-half Q/K at this point increases perplexity by 82.3; the increase reaches 146.1 at the end of training (Figure 2D). With the intervention, the corresponding increases are 0.35 and 17.4. Upper attention in the control therefore becomes causally important at the same early stage when its logits grow and its entropy falls.

Under the intervention, reliance on upper Q/K develops gradually. The smaller ablation penalty persists through the end of training, when the intervention also achieves lower validation perplexity. Early learning-rate reduction changes both the onset and the final extent of this dependence.

Layer-band ablations locate the control's early dependence near the top of the decoder. At 3% of training, zeroing Q/K in layers 10–12, 13–15, or 16–19 of the 20-layer model, indexed from zero, increases PPL by 3.526, 14.052, and 35.773, respectively. Under the intervention, all three increases are below 0.13. By the final checkpoint, each band contributes to prediction in the intervention, with

Table 2: Downstream benchmark evaluation for the GPT-style intervention comparison. Results are mean ± sample standard deviation over three runs on the pre-specified seven-task suite. Delta is the paired intervention-minus-control difference in percentage points; the mean-row deviation is computed over seed-level suite averages.

Task	Control	Intervention	Delta pp
PIQA	0.6065 ± 0.0056	0.6103 ± 0.0013	$+0.381 \pm 0.544$
HellaSwag	0.3011 ± 0.0013	0.3024 ± 0.0020	$+0.133 \pm 0.314$
ARC-Easy	0.4315 ± 0.0125	0.4373 ± 0.0034	$+0.575 \pm 1.582$
ARC-Challenge	0.2466 ± 0.0023	0.2460 ± 0.0125	-0.057 ± 1.258
RACE	0.2695 ± 0.0049	0.2778 ± 0.0056	$+0.829 \pm 1.023$
LAMBADA	0.1974 ± 0.0045	0.2073 ± 0.0004	$+0.996 \pm 0.485$
MMLU	0.2291 ± 0.0009	0.2294 ± 0.0001	$+0.028 \pm 0.082$
Mean	–	–	$+0.412 \pm 0.143$

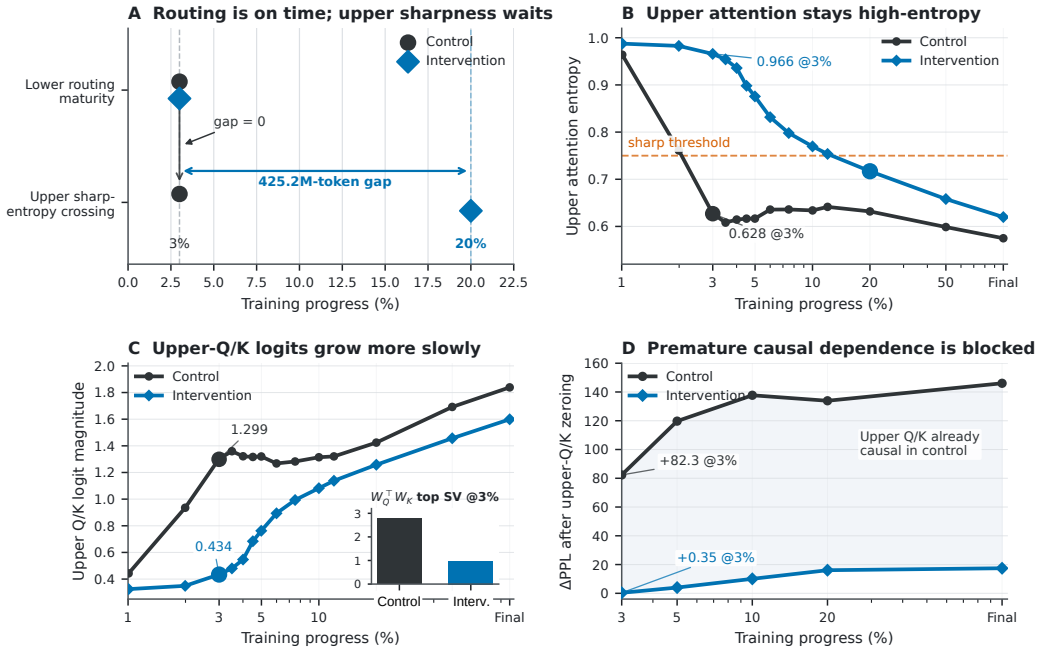

Figure 2: The intervention delays upper attention specialization without delaying lower routing. Control denotes default training; Intervention denotes reducing only early upper-half Q/K learning by a factor of 0.25 before release. (A) Lower routing matures at 3% in both runs, while upper sharp-entropy crossing moves from 3% in the control to 20% under the intervention. (B) Upper attention entropy stays high longer under the intervention. (C) Upper-Q/K logit growth is suppressed, with the 3% upper $W_Q^\top W_K$ top singular value shown as an inset. (D) Zeroing upper-half Q/K reveals that the control becomes prematurely dependent on upper Q/K, whereas the intervention sharply reduces that dependence.

penalties between 2.181 and 3.289. Upper-layer matching becomes useful during subsequent training across all three bands.

5.4 Lower Routing and Upper Specialization

The delay in upper attention sharpening persists after the learning rate returns to normal. Lower routing reaches the maturity threshold at 3% of training in both conditions (Figure 2A). The patience requirement places release at 4.01%, followed by the 1% ramp back to the normal rate. Upper attention crosses the sharp-entropy threshold at 3% in the control and at the measured 20% checkpoint under

Table 3: Upper-Q/K learning-rate multipliers in the 270M decoder, using one seed. Entropy, logit magnitude, and the Frobenius displacement of $W_Q^\top W_K$ from initialization are measured at 3% of training; PPL is measured after training.

Setting	Multiplier	Final PPL	Upper entropy	Upper logit	Q/K bilinear displacement
Control	1.0	26.712	0.627	1.302	3.272
Intervention	0.5	26.337	0.836	0.775	1.691
Intervention	0.25	26.218	0.966	0.435	0.705
Intervention	0.125	26.226	0.984	0.342	0.333

Table 4: Release schedules in the 270M decoder, using one shared seed. Each intervention uses a 0.25 upper-Q/K multiplier and a 1% ramp to the normal learning rate. PPL reduction and token savings are relative to the control below. Entropy and logit magnitude are measured at 3% of training.

Setting	Release rule	Final PPL	PPL reduction	Tokens saved	Entropy @3%	Logit @3%
Control	None	26.7382	–	–	0.6215	1.3081
Adaptive lower-copy	Lower-copy	26.1524	0.5858	14.80%	0.9657	0.4287
Fixed release at 3%	Fixed	26.2274	0.5108	13.92%	0.9655	0.4343
Fixed release at 6%	Fixed	26.2703	0.4680	13.07%	0.9665	0.4308

the intervention. Much of this separation therefore develops with upper Q/K already back on the normal schedule.

The early measurements locate the start of this altered trajectory. At the routing-maturity checkpoint, the intervention has higher upper attention entropy, lower logit magnitude, and a smaller top singular value of $W_Q^\top W_K$ (Figure 2B–C). Upper attention entropy stays above 0.75 until the measured 20% checkpoint, leaving 425.2M training tokens between routing maturity and sharpness. Lower-layer features have time to develop before upper attention concentrates on them. Together with the final ablation results, these measurements show a lasting difference in attention development and causal dependence following a brief change to early optimization.

5.5 Learning-Rate Multiplier and Logit Growth

Smaller early upper-Q/K learning rates produce progressively weaker upper-logit growth. Table 3 varies the multiplier while holding the architecture, data, token budget, and release rule fixed. At 3% of training, reducing the multiplier from 1.0 to 0.125 lowers upper-logit magnitude from 1.302 to 0.342 and raises entropy from 0.627 to 0.984. The displacement of $W_Q^\top W_K$ from initialization falls from 3.272 to 0.333. These measurements connect the learning-rate change to attention sharpness and to the weights that produce the logits.

All three reduced multipliers improve final perplexity. The lowest value in this sweep is 26.218 at $\alpha = 0.25$, close to 26.226 at $\alpha = 0.125$. Early logit growth continues to fall across these two settings, while final perplexity levels off. The sweep supports the learning-rate dependence analyzed in Section 6 and shows similar final performance over a range of suppression strengths.

5.6 Release Timing

Fixed early release times preserve most of the gain from the adaptive rule. Table 4 compares the lower-copy trigger with releases at 3% and 6% of training. All three schedules use a pre-release multiplier of 0.25 and a 1% ramp back to the normal rate. The adaptive rule reduces PPL by 0.5858; fixed releases reduce it by 0.5108 and 0.4680, respectively. Their token savings remain 13.92% and 13.07%, compared with 14.80% for adaptive release.

The adaptive rule gives the lowest PPL in this comparison, and both fixed schedules retain substantial improvements. Nearby threshold and patience choices also select releases within the early 3–6% window when replayed on the recorded lower-copy scores (Appendix D). The fixed-release training

Table 5: Global learning-rate reduction and entropy regularization in two 270M comparisons. Each block uses its own control; the columns report final PPL. Full measurements are in Appendices C and B.

Training intervention	Control PPL	Intervention PPL
Global learning-rate comparison		
Global $0.5\times$ learning rate	26.849	31.418
Selective upper-Q/K slowing	26.849	26.413
Entropy-floor comparison		
Entropy-floor control	26.738	26.678
Upper-Q/K slowing	26.738	26.152

results establish the persistence of the gain across release times; the replay describes how the trigger responds to its hyperparameters.

5.7 Global Learning Rate and Entropy Regularization

Selective upper-Q/K slowing improves PPL while preserving the normal learning schedule of the remaining parameters. In the global learning-rate comparison in Table 5, halving the rate for all parameters raises final PPL from 26.849 to 31.418. Selective slowing lowers it to 26.413. At 20% of training, the lower-copy score is 0.0164 under selective slowing and 0.0103 under global slowing. Retaining lower-layer learning accompanies the better final performance (Appendix C).

Direct entropy regularization has a smaller effect in the second comparison. The entropy-floor penalty uses $h_0 = 0.80$ and $\lambda = 0.10$, with the same lower-copy release rule as the intervention (Appendix B). It raises early upper attention entropy from 0.621 to 0.799 and improves final PPL by 0.061. Upper-Q/K slowing reaches entropy 0.966 and improves PPL by 0.586. Within these settings, the learning-rate intervention produces a larger change in both early attention and final perplexity.

6 Theory

We analyze how the early upper-Q/K learning rate controls the local attention-logit update. Let $X \in \mathbb{R}^{n\times d}$ be the normalized residual stream entering an upper attention head, $B = W_Q W_K^\top$, and $Z = XBX^\top/\sqrt{d_k}$. Let $P = P^\top = P^2$ project onto immature residual directions and write $X_P = XP$. The immature logit component is

$$Z_P = \frac{X_P P B P X_P^\top}{\sqrt{d_k}}. \tag{2}$$

Theorem 1 (Pathwise premature-logit growth)**.** *For one effective gradient step on W_Q, W_K with upper step sizes η_Q, η_K, the first-order immature-logit update obeys a deterministic pathwise bound*

$$\|\Delta Z_P\|_F \le \frac{\|X_P\|_{\mathrm{op}}^2 \|X_P\|_F \|E\|_{\mathrm{op}}}{d_k}\Big(\eta_Q \|XW_K W_K^\top P\|_F + \eta_K \|XW_Q W_Q^\top P\|_F\Big) + O(\eta_Q\eta_K), \tag{3}$$

where $E = \partial\mathcal{L}/\partial Z$ is the masked-softmax adjoint. Without extra assumptions this gives an unconditional bound controlled by the total residual scale. Under the measurable immature-channel locality condition

$$\|XW_S W_S^\top P\|_F \le \lambda_S \|X_P\|_F \|W_S\|_{\mathrm{op}}^2, \qquad S \in \{Q, K\}, \tag{4}$$

it sharpens to

$$\|\Delta Z_P\|_F = O\left(\eta_{\mathrm{upper}} \frac{\|X_P\|_F^4}{d_k}\right). \tag{5}$$

At 3% of training, the measured locality ratios (λ_Q, λ_K) are $(0.63, 0.58)$ for the control and $(0.49, 0.46)$ for the intervention. Both pairs satisfy the condition with constants below one; Appendix F gives the diagnostic definition and measurements.

The intervention sets $\eta_Q = \eta_K = \alpha\eta$ for early upper-half Q/K, with $\alpha = 0.25$ in the main experiments. At a fixed checkpoint, Theorem 1 gives $\Delta Z_P^{(\alpha)} = \alpha \Delta Z_P^{(1)} + O(\eta^2)$. The multiplier directly scales the first-order update while leaving the schedules of lower layers, values, FFNs, embeddings, and the output head unchanged. Table 3 shows the corresponding decline in early logit magnitude and bilinear displacement as α decreases.

Low-entropy attention requires logit range: if one of N causal keys receives probability at least $1 - \epsilon$, then

$$\max_j z_j - \min_j z_j \geq \log \frac{(N-1)(1-\epsilon)}{\epsilon}. \tag{6}$$

The local update analysis identifies upper-Q/K learning rate as a direct control on logit change. Figure 2 measures the resulting delay in upper attention sharpening over training. Appendix A gives the complete proofs.

7 Discussion

The intervention's effect persists after the normal learning rate resumes. Upper attention remains diffuse after release, and its causal contribution develops gradually during subsequent training. Lower routing matures on schedule, giving upper attention a more developed representation over which to learn matching patterns. By the final checkpoint, upper Q/K contributes to prediction across all three tested layer bands. The final model has lower validation perplexity and a smaller upper-Q/K ablation penalty than the control.

The multiplier and release experiments support an early period of reduced upper-Q/K learning followed by a return to the normal rate. Early logit magnitude and bilinear displacement respond to the multiplier, while final perplexity levels off across the smaller multipliers tested. Both fixed-release schedules retain most of the adaptive rule's gain. Selectivity also matters in the learning-rate comparison: keeping the other parameter groups on their normal schedules preserves lower-layer development, whereas the global half-rate control slows lower routing and worsens final perplexity.

The cross-scale validation gains and the 270M downstream results motivate treating the early upper-Q/K schedule as a pretraining design choice. Changing this schedule requires no additional model parameters or inference-time computation.

8 Conclusion

Premature upper-layer attention specialization is an optimization failure in GPT-style decoder pretraining. Selectively slowing early upper-Q/K learning improves final validation loss across three model scales, with additional gains in token efficiency and downstream evaluation. The intervention preserves lower routing development while delaying upper attention sharpening and reducing strong early Q/K dependence. The pathwise analysis connects this intervention to the local attention-logit update. These results identify the timing of upper-layer attention learning as a concrete determinant of pretraining quality.

Limitations. The experiments cover models up to 1B parameters and training budgets up to 30B tokens. Model size and token budget increase together across the three scales, so their individual effects on the size of the gain remain unresolved. Downstream evaluation covers the 270M model; extending the seven-task suite to the larger models would establish how the validation gains transfer at those scales. The multiplier and fixed-release comparisons each use one seed, leaving uncertainty in the ranking of nearby schedules. Lower-copy score measures one aspect of routing development, and the sharpness times are resolved at the sampled checkpoints. The theory describes a local update with the residual representation fixed during that step. Extending it to jointly evolving lower and upper layers is needed to characterize the full training trajectory.

AI USE STATEMENT

Generative AI tools assisted with manuscript preparation, writing, and language polishing. The authors take responsibility for the final text, claims, and results.

REPRODUCIBILITY STATEMENT

Section 3.2 specifies the upper-Q/K multiplier, release criterion, and ramp schedule. Section 4 describes the model and evaluation settings, while Appendices E and G provide the 700M configuration and the data, optimizer, precision, and hardware details for the 270M and 700M comparisons. Appendix A contains the theoretical derivations. Sections 5.5–5.7 report the multiplier, release-time, and training-objective comparisons; Appendix D details the trigger replay.

REFERENCES

Ben Anson and Laurence Aitchison. Controlling changes to attention logits. *arXiv preprint arXiv:2511.21377*, 2025.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*, 2023.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagne, Alexandra Sasha Luccioni, François Yvon, et al. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2019.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. *Advances in neural information processing systems*, 37:64273–64311, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, 2019.

Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5758–5773, 2021.

Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the International Conference on Spoken Language Translation*, 2019.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Sam Shleifer, Jason Weston, and Myle Ott. NormFormer: Improved transformer pretraining with extra normalization. *arXiv preprint arXiv:2110.09456*, 2021.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Oskar van der Wal, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. PolyPythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations*, 2025.

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6761–6774, 2024.

Kevin R. Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the International Conference on Machine Learning*, 2020.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse. *arXiv preprint arXiv:2303.06296*, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Full Theoretical Derivations

We derive the pathwise upper-Q/K update bound and its dependence on the learning-rate multiplier. The analysis uses the normalized residual stream, the Q/K bilinear form, and the masked-softmax adjoint, and isolates the logit change from one Q/K update at a fixed input.

Let $X \in \mathbb{R}^{n\times d}$ be the normalized residual stream entering one upper-layer attention head, and let

$$W_Q, W_K \in \mathbb{R}^{d\times d_k}, \qquad B = W_Q W_K^\top. \tag{7}$$

The pre-softmax attention logits are

$$Z = \frac{XBX^\top}{\sqrt{d_k}}. \tag{8}$$

Let $P = P^\top = P^2$ be an orthogonal projector onto immature residual directions and define $X_P = XP$. We isolate the immature-to-immature logit component

$$Z_P = \frac{X_P\, PBP\, X_P^\top}{\sqrt{d_k}}. \tag{9}$$

Let $E = \partial\mathcal{L}/\partial Z$ be the adjoint through the masked rowwise softmax. The gradients of the head logits with respect to W_Q, W_K are

$$G_Q = \frac{X^\top E X W_K}{\sqrt{d_k}}, \qquad G_K = \frac{X^\top E^\top X W_Q}{\sqrt{d_k}}. \tag{10}$$

A.1 Proof of Theorem 1

For completeness, we restate Theorem 1 with the explicit second-order remainder.

Full statement of Theorem 1. Suppose one effective gradient step updates

$$W_Q^+ = W_Q - \eta_Q G_Q, \qquad W_K^+ = W_K - \eta_K G_K. \tag{11}$$

Let $B^+ = W_Q^+ (W_K^+)^\top$, and let

$$\Delta Z_P = \frac{X_P\, P(B^+ - B)P\, X_P^\top}{\sqrt{d_k}}. \tag{12}$$

Then

$$\begin{aligned}\|\Delta Z_P\|_F \le \frac{\|X_P\|_{\mathrm{op}}^2 \|X_P\|_F \|E\|_{\mathrm{op}}}{d_k}\Big(&\eta_Q \|X W_K W_K^\top P\|_F \\ &+\eta_K \|X W_Q W_Q^\top P\|_F\Big) + R_2,\end{aligned} \tag{13}$$

where the second-order remainder satisfies

$$R_2 \le \frac{\eta_Q\eta_K \|X_P\|_{\mathrm{op}}^2 \|X_P\|_F^2 \|E\|_{\mathrm{op}}^2 \|XW_K\|_F \|XW_Q\|_F}{d_k^{3/2}}. \tag{14}$$

In particular, ignoring $O(\eta_Q\eta_K)$ terms,

$$\|\Delta Z_P\|_F = O\big((\eta_Q+\eta_K)\|E\|_{\mathrm{op}}\|X_P\|_{\mathrm{op}}^2\|X_P\|_F\big), \tag{15}$$

with the remaining dependence carried by how the current Q/K quadratic form maps into the immature subspace.

Proof. The bilinear matrix update is

$$\begin{aligned}B^+ - B &= (W_Q - \eta_Q G_Q)(W_K - \eta_K G_K)^\top - W_Q W_K^\top \\ &= -\eta_Q G_Q W_K^\top - \eta_K W_Q G_K^\top + \eta_Q\eta_K G_Q G_K^\top.\end{aligned} \tag{16}$$

For the first-order Q-term,

$$P G_Q W_K^\top P = \frac{P X^\top E X W_K W_K^\top P}{\sqrt{d_k}} = \frac{X_P^\top E X W_K W_K^\top P}{\sqrt{d_k}}. \tag{17}$$

Therefore,

$$\|PG_Q W_K^\top P\|_F \le \frac{\|X_P\|_F \|E\|_{\mathrm{op}} \|X W_K W_K^\top P\|_F}{\sqrt{d_k}}. \tag{18}$$

Similarly,

$$PW_Q G_K^\top P = \frac{PW_Q W_Q^\top X^\top E X_P}{\sqrt{d_k}}, \tag{19}$$

and hence

$$\|PW_Q G_K^\top P\|_F \le \frac{\|X W_Q W_Q^\top P\|_F \|E\|_{\mathrm{op}} \|X_P\|_F}{\sqrt{d_k}}. \tag{20}$$

Finally,

$$\|X_P\, P\Delta B P\, X_P^\top\|_F \le \|X_P\|_{\mathrm{op}}^2 \|P\Delta B P\|_F. \tag{21}$$

Dividing by the outer $\sqrt{d_k}$ in the definition of ΔZ_P gives the first-order bound.

For the second-order term,

$$\|PG_Q G_K^\top P\|_F \le \|PG_Q\|_F \|G_K^\top P\|_F. \tag{22}$$

Using

$$\|PG_Q\|_F \le \frac{\|X_P\|_F \|E\|_{\mathrm{op}} \|X W_K\|_F}{\sqrt{d_k}}, \tag{23}$$

and

$$\|G_K^\top P\|_F \le \frac{\|X W_Q\|_F \|E\|_{\mathrm{op}} \|X_P\|_F}{\sqrt{d_k}}, \tag{24}$$

then multiplying again by the outer factor $\|X_P\|_{\mathrm{op}}^2/\sqrt{d_k}$ gives the claimed R_2. □

Corollary 1 (Unconditional pathwise bound)**.** *Without any additional structural assumption,*

$$\|X W_S W_S^\top P\|_F \le \|X\|_F \|W_S\|_{\mathrm{op}}^2, \qquad S \in \{Q, K\}. \tag{25}$$

Therefore,

$$\|\Delta Z_P\|_F \le \frac{\|E\|_{\mathrm{op}} \|X\|_F}{d_k} \Big(\eta_Q \|W_K\|_{\mathrm{op}}^2 + \eta_K \|W_Q\|_{\mathrm{op}}^2\Big) \|X_P\|_{\mathrm{op}}^2 \|X_P\|_F + R_2. \tag{26}$$

Using $\|X_P\|_{\mathrm{op}} \le \|X_P\|_F$ gives a cubic dependence on immature residual amplitude, modulated by the total residual scale. At a fixed checkpoint, the learning rates enter this bound linearly at first order, giving a direct control on the local upper-logit update.

Corollary 2 (Localized immature-channel bound)**.** *Assume the following locality condition holds for the immature subspace P:*

$$\|X W_S W_S^\top P\|_F \le \lambda_S \|X_P\|_F \|W_S\|_{\mathrm{op}}^2, \qquad S \in \{Q, K\}. \tag{27}$$

This says that the part of the upper Q/K quadratic form that lands in immature directions is not mainly driven by mature residual directions. Under this condition,

$$\|\Delta Z_P\|_F \le \frac{\|X_P\|_{\mathrm{op}}^2 \|X_P\|_F^2 \|E\|_{\mathrm{op}}}{d_k} \Big(\eta_Q \lambda_K \|W_K\|_{\mathrm{op}}^2 + \eta_K \lambda_Q \|W_Q\|_{\mathrm{op}}^2\Big) + R_2. \tag{28}$$

Using $\|X_P\|_{\mathrm{op}} \le \|X_P\|_F$, this gives

$$\|\Delta Z_P\|_F = O\left(\eta_{\mathrm{upper}} \frac{\|X_P\|_F^4}{d_k}\right). \tag{29}$$

This is the rigorous version of the localized $\eta_{\mathrm{upper}} \|X_P\|_F^4$ statement. Appendix F gives the corresponding diagnostic ratio.

Corollary 3 (Upper-Q/K learning-rate suppression). *Suppose Q and K use the same effective upper learning rate η. Apply an early upper-Q/K multiplier $0 < \alpha < 1$, so that*

$$\eta'_Q = \eta'_K = \alpha\eta. \tag{30}$$

At a fixed checkpoint with the same X, E, W_Q, W_K,

$$\Delta Z_P^{(\alpha)} = \alpha \Delta Z_P^{(1)} + O(\eta^2). \tag{31}$$

Consequently,

$$\|\Delta Z_P^{(\alpha)}\|_F \le \alpha \|\Delta Z_P^{(1)}\|_F + O(\eta^2). \tag{32}$$

For the intervention used in this paper, $\alpha = 0.25$, so the first-order immature upper-logit growth channel is reduced fourfold.

The selective intervention scales the first-order upper-Q/K update while lower layers and FFNs continue on their normal learning schedules. Section 5.5 reports the corresponding multiplier sweep.

A.2 Entropy Collapse Requires Logit-Range Growth

Finally, we connect logit growth to attention sharpness. For a row of attention over N causal keys, let

$$p_j = \frac{e^{z_j}}{\sum_{\ell=1}^{N} e^{z_\ell}}. \tag{33}$$

If one key obtains probability at least $1 - \epsilon$, then

$$\max_j z_j - \min_j z_j \ge \log \frac{(N-1)(1-\epsilon)}{\epsilon}. \tag{34}$$

Proof. Let $j^\star$ be the maximum-probability key, so $p_{j^\star} \ge 1 - \epsilon$. The remaining $N - 1$ keys have total mass at most ϵ, so at least one key $j_{\min}$ has

$$p_{j_{\min}} \le \frac{\epsilon}{N-1}. \tag{35}$$

Softmax ratios satisfy

$$\frac{p_{j^\star}}{p_{j_{\min}}} = e^{z_{j^\star} - z_{j_{\min}}}. \tag{36}$$

Therefore,

$$e^{z_{j^\star} - z_{j_{\min}}} \ge \frac{(N-1)(1-\epsilon)}{\epsilon}, \tag{37}$$

and taking logs gives the result. □

Thus, sharp attention requires sufficient logit range. The learning-rate corollary identifies how selective upper-Q/K slowing reduces the local update that contributes to this range. The training-time measurements in Figure 2 show the accompanying suppression of early upper-logit growth and delay in attention sharpening.

B Entropy-Floor Mediator Control

The entropy-floor comparison in Section 5.7 raises early upper attention entropy through a regularization term. The architecture and learning-rate schedule match the control, and the early training objective adds

$$\lambda \max(0, h_0 - H_{\text{upper}}), \tag{38}$$

where H_{upper} is the mean upper-half attention entropy, $h_0 = 0.80$, and $\lambda = 0.10$. The regularizer is released by the same lower-copy maturity rule as the main intervention.

The penalty raises early upper attention entropy and lowers logit magnitude, with a final PPL reduction of 0.061. Selective upper-Q/K slowing produces a larger change in both early measurements and final PPL in this comparison, as reported in Table 5.

Table 6: Entropy-floor mediator control in the 270M GPT-style decoder. The entropy-floor control directly penalizes low upper attention entropy while leaving the architecture and upper-Q/K learning rate unchanged. PPL reduction is relative to the GPT-style control in the same matched comparison.

Setting	Entropy floor	Entropy @3%	Logit @3%	Final ppl	PPL reduction
GPT-style control	–	0.621	1.308	26.738	–
Entropy-floor control	0.80	0.799	0.861	26.678	0.061
Upper-Q/K slowing	–	0.966	0.429	26.152	0.586

C Global Learning-Rate Control

The global learning-rate comparison in Section 5.7 halves the rate for all parameters while keeping the 270M architecture, data, token budget, batch geometry, and schedule shape fixed. This applies the reduction to lower layers, FFNs, embeddings, and output learning alongside upper attention.

Table 7: Global learning-rate control in the 270M GPT-style setting. The selective intervention slows only early upper-half Q/K, while the global control halves the learning rate for all parameters.

Setting	Final loss	Final ppl	Δ loss	Δ ppl
GPT-style control	3.2902	26.849	–	–
Selective upper-Q/K slowing	3.2739	26.413	-0.0164	-0.436
Global 0.5$\times$ learning rate	3.4474	31.418	+0.1571	+4.569

Global slowing increases final PPL by 4.569, while selective upper-Q/K slowing reduces it by 0.436. At 20% of training, the selective intervention has lower-copy score 0.0164 and upper attention entropy 0.729; the global half-rate control has lower-copy score 0.0103 and upper entropy 0.660. The selective schedule preserves faster lower-copy development and yields better final model quality in this comparison.

D Release-Rule Robustness

The lower-copy release rule responds similarly to nearby threshold, patience, and window choices. Table 8 recomputes the release point from recorded lower-copy scores under each rule, without retraining. Section 5.6 reports the separate training experiments with fixed release times.

The replayed release points remain in the same early window; the 3% maturity event in Figure 2 is the first threshold crossing, while the 4.01% main-rule release in Table 8 occurs after the required three consecutive evaluations. Lowering the threshold does not move the release earlier because the patience condition and 3% minimum window already bind; raising the threshold or increasing patience moves release modestly later, but still within the early specialization period.

These rule variants select releases within the same early maturity window. The training comparisons in Table 4 measure final model quality after release at representative points in that window.

E 700M Configuration and Training-Time Measurements

The 700M results in Table 1 are averaged over three seeds. The model uses 22 layers, width 1536, 12 attention heads, FFN width 6144, sequence length 1024, and 700M parameters. All runs use the same data order, token budget, optimizer family, cosine schedule, and evaluation protocol. The intervention is unchanged: early upper-half Q/K uses a 0.25 learning-rate multiplier and the same lower-copy release rule defined in Section 3.

The result follows the 270M pattern. At 3% training, the control has lower upper attention entropy and larger upper Q/K logits, while the intervention leaves lower copy-routing comparable and suppresses upper sharpness. By the end of the 7.0B-token run, the intervention improves final loss, final perplexity, and same-loss token efficiency.

Table 8: Release checkpoints obtained by applying lower-copy rule variants to logged evaluation checkpoints from the main GPT-style intervention experiments. The main rule and nearby threshold, patience, and forced-window variants all select the same early maturity window.

Rule	Threshold	Patience	Release window	Release checkpoint
Main rule	0.005	3	3–12%	4.01%
Lower threshold	0.003	3	3–12%	4.01%
Higher threshold	0.007	3	3–12%	5.49%
No patience	0.005	1	3–12%	3.00%
Stricter patience	0.005	5	3–12%	4.99%
Shorter forced window	0.005	3	3–8%	4.01%

F Immature-Channel Locality Diagnostic

The locality condition in Section 6 is directly measurable from checkpoints. For $S \in \{Q, K\}$ define

$$\lambda_S = \frac{\|XW_S W_S^\top P\|_F}{\|X_P\|_F \|W_S\|_{\mathrm{op}}^2}. \tag{39}$$

Small or moderate values of λ_Q, λ_K indicate that the upper-Q/K quadratic form landing in immature directions is not primarily driven by mature residual directions. This is the empirical condition under which the unconditional pathwise bound sharpens to the localized $\eta_{\mathrm{upper}}\|X_P\|_F^4/d_k$ growth term. We also report $R_P/\|X\|_F^2$, where $R_P = \|X_P\|_F^2$, to show that the immature residual channel is present but does not cover the full residual stream.

Table 9: Immature-channel locality diagnostic at the 3% checkpoint used in the mechanism measurements. The locality ratios remain below one for both the control and intervention, supporting the localized immature-channel interpretation used in Corollary 2.

Setting	λ_Q	λ_K	$R_P/\|X\|_F^2$
Control	0.63	0.58	0.18
Intervention	0.49	0.46	0.16

The measured ratios support the condition used by the localized theorem. Both λ_Q and λ_K remain below one, so the locality constants are not absorbing an uncontrolled blow-up. The control has larger ratios than the intervention, matching the mechanism evidence that default early upper-Q/K learning forms stronger premature matches. The normalized immature energy is also substantial but bounded: $R_P/\|X\|_F^2$ is 0.18 in the control and 0.16 under the intervention, so the diagnostic isolates a meaningful residual channel rather than relabeling the entire residual stream as immature.

G Reproducibility and Compute Details

The 270M and 700M pretraining comparisons use packed GPT-2-tokenized FineWeb-Edu text (Penedo et al., 2024) with sequence length 1024 and a held-out validation stream sampled from the same preparation pipeline. The 270M runs use 2.5B training tokens, global batch size 524,288 tokens, AdamW with learning rate 2.5×10^{-4}, cosine decay to 10% of peak learning rate, 2% warmup, weight decay 0.1, betas $(0.9, 0.95)$, and bf16 training. The 700M replication uses 7.0B training tokens, the same global batch size and optimizer settings except peak learning rate 2.0×10^{-4}, and bf16 training.

The 270M and 700M runs use 8 H100 GPUs. Their paired comparisons keep data order, token budget, batch geometry, optimizer family, schedule shape, and evaluation protocol fixed within the pair.