When Structure Doesn't Matter: Loss-Invariant Routing Specialisation in Hybrid Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We construct a controlled setting in which an architectural channel in a hybrid state-space/attention language model is free to organise its internal representations without measurably affecting the optimisation objective. The channel is *hormone routing*: a token-conditional softmax-gated linear combination of a small bank of frozen direction vectors, added to the residual stream after each attention block. We pretrain a 125M-parameter Mamba/MQA hybrid on FineWeb-Edu under a fixed 2B-token budget across five variants of this channel — extracted directions, random unit vectors, vectors extracted from a randomly-initialised model, and a frozen gate at 1.0. The router develops semantically-coherent per-genre routing specialisation in every variant, with maximum pairwise L_1 distance between genre routing distributions in 0.6–1.6 (uniform $= 0$, disjoint $= 2$). The variant of *highest* specialisation strength is the one whose gate the model cannot modulate. **Specialisation strength does not predict downstream loss**: the four properly-initialised variants converge to identical validation perplexity (26.13). To probe whether this loss-invariance is a ceiling or a regime the model actively absorbs, a sixth variant adds a forced non-zero gate initialisation and a non-trainable $3\times$ amplification. The model drifts the gates downward and shrinks the per-hormone magnitudes, but the residual perturbation it cannot suppress — $\approx 2\times$ the unforced regime — does not damage loss either. We refer to the regime in which the gated injection is consistently absorbed without affecting predictions as a *homeostatic envelope*. The envelope is wider than the unforced variants naturally select; the only axis on which the mechanism can break the model is the training schedule (introducing it during base pretraining damages the loss by $+2.07$ PPL even though the same router still develops specialisation). The loss-relevant variable is therefore *when* the perturbation is introduced, not *what* it contains nor how strongly it is gated. We additionally describe *intracellular attention*, a parameter-efficient slot-attention mechanism inside the SSM scan, as an architectural extension; full empirical evaluation requires a fused CUDA kernel and is reported as preliminary.

1 Introduction

Architectural mechanisms in language models are typically introduced with a presumed contribution to the optimisation objective: a new attention variant, normalisation, or gating scheme is judged by whether it lowers the loss. We instead investigate a setting in which an architectural channel is free to develop structured internal representations *without* measurably affecting the loss, and ask what such a channel does.

Our setting is built on hybrid state-space/attention language models [10, 32, 20, 9, 24]. Inside this backbone we introduce *hormone routing*, a parameter-efficient channel that injects a token-conditional

gated combination of a small bank of frozen direction vectors into the residual stream after each attention block. The bank is non-trainable; the router and gate are. The mechanism adds under 0.01% of base parameters and is gated such that, at initialisation, the modification is an exact no-op and the hybrid backbone is recovered bit-for-bit. Directions are obtained from a partially-trained Stage-1 checkpoint via document-style contrast-pair extraction [31, 26, 35].

The empirical phenomenon. Across every hormone-routing variant we tested at the 125M-parameter, 2B-token scale — with extracted directions, with random unit vectors, with the gate frozen at 1.0, and with directions extracted from a randomly-initialised model — the router converges to a strongly genre-discriminative routing distribution (max pairwise L_1 distance across six text genres: 0.6–1.6, vs 0.0 for uniform). The specialisation is qualitatively interpretable in the variant where directions carry semantic labels (news $\rightarrow$ urgency, fiction $\rightarrow$ curiosity, code $\rightarrow$ skepticism) and structurally analogous in the variants where they do not. Yet none of these variants moves the validation perplexity needle: all properly-initialised variants achieve identical val PPL (26.13), and loss trajectories trace the same curve to within 10^{-3} across all of Stage 2. Training the same router from scratch with hormones in place from step 0 damages the loss by $+2.07$ PPL even though it still develops similar specialisation. **The loss-relevant variable is when the perturbation is introduced, not what it contains.**

Disambiguating the regime. Two readings remain open: either the perturbation magnitude in the unforced variants is too small to matter (capacity-bottlenecked), or the architecture has actively absorbed it (structural homeostasis). We disambiguate with a sixth variant with a forced non-zero gate initialisation and a non-trainable $3\times$ amplification — producing an effective perturbation $\approx 2\times$ the largest unforced gate. The model drifts the gates downward and shrinks per-hormone magnitudes but cannot escape because α is non-trainable; the residual perturbation it cannot suppress does not damage validation loss at the matched training step. The capacity-bottlenecked reading would predict improvement; the intrinsically-harmful reading would predict damage; we observe neither. We refer to the regime within which the gated injection is consistently absorbed without measurable change to predictions as a *homeostatic envelope*.

Contributions.

- **Empirical (primary)**: a clean decoupling between routing-channel specialisation and downstream loss in a 125M-parameter hybrid LM. Five hormone-routing variants on a 2B-token budget all develop strongly genre-discriminative routing; specialisation strength does not correlate with validation perplexity. A sixth, forced-engagement variant rules out the capacity-bottlenecked reading. The only loss-relevant axis is training schedule.
- **Architectural (primary)**: *hormone routing*, a token-conditional softmax-gated routing channel, plus a document-style contrast-pair extraction protocol that recovers behavioural directions from the base model itself.
- **Architectural (extension)**: *intracellular attention*, a parameter-efficient slot-attention mechanism inside the SSM scan with the same zero-init gating discipline. The reference implementation has $\approx 25\times$ the per-token throughput of the fused Mamba kernel; full-budget evaluation is left to future work pending a fused CUDA implementation. We report a preliminary engagement signal at a 12M-token Stage-2 budget (Section 6.5) but draw no empirical conclusions about its full-budget effect on the loss.

The rest of the paper situates the work (Section 2), gives the hybrid backbone (Section 3), presents the two mechanisms (Sections 4, 5), reports the empirical study (Section 6), and characterises what the router learns (Section 7).

2 Related Work

State-space models and hybrids. Structured state-space models [11–13] reformulate sequence mixing as a parameterised linear dynamical system; Mamba [10] adds input-dependent selectivity to the discretisation step, input matrix, and readout. Pure-SSM models lag pure-attention transformers on tasks requiring precise in-context recall [16, 2]; hybrids interleave attention layers among many SSM layers [20, 9], and Park et al. [24] show that even a single attention layer per block recovers

most of the lost expressivity. We adopt a 5:1 SSM-to-attention ratio with multi-query attention [27] and rotary positional embeddings [30]. Our contributions are orthogonal to the choice of mixing ratio.

Activation steering and representation engineering. A growing body of work shows that LLM behaviour is modulated by a small number of linear directions in the residual stream. Zou et al. [35] introduce representation-engineering as a framework for reading and writing such directions; Turner et al. [31] and Rimsky et al. [26] demonstrate that adding a difference-of-means vector — computed from contrastive prompts — can steer attributes like truthfulness, refusal, and persona at inference time. Earlier work on contrast-consistent search [5] and probing [1] establishes that interpretable axes are present in unmodified models. All of this work applies the directions *post-hoc* at inference time. Our hormone-routing module instead re-injects extracted directions *during pretraining* through a learned router; the directions themselves stay frozen.

Routing, mixture-of-experts, and soft prompts. MoE models [29, 8, 17] use learned routers to dispatch tokens to specialised expert subnetworks. Our hormone router is structurally analogous — a linear-then-softmax map from the hidden state to a distribution over a small fixed set — but the "experts" here are frozen direction vectors of dimension d_{model}, not parameter banks. Soft prompts [18, 19] likewise use a small bank of d_{model}-dimensional vectors to modulate behaviour, but those vectors are prefix tokens exposed to attention rather than residual-stream additions, and are fixed across the sequence rather than token-conditional.

Probing and the structure-vs-task gap. A long line of work probes neural network internals for learned structure not obviously demanded by the training objective [1, 14, 3]. Hewitt and Liang [14] caution that high-capacity probes can recover arbitrary structure regardless of whether the network exploits it. Our routing-specialisation finding is in the same spirit but at a different locus: the routing module is not a post-hoc probe but a built-in architectural channel, and the structure it develops is a property of the model's own forward pass. The decoupling we observe between specialisation strength and downstream loss is stronger than "the probe finds structure that does not matter" — it is "the model itself produces and uses structure that does not matter for its training objective," at least at the scale we examine.

3 Architecture

We use a hybrid Mamba/MQA backbone of L residual layers with a fixed 5:1 SSM-to-attention pattern (five SSM layers between every attention layer). All layers share hidden width d_{model}, use RMSNorm [34], and tie word embeddings with the LM head.

SSM layers. Each SSM layer applies pre-norm and a Mamba-1 selective block [10], with no separate channel-mixing MLP. The block projects to twice the inner width ($d_{\text{inner}} = 2d_{\text{model}}$), applies a causal depthwise convolution, and runs a selective scan parameterised by an input-dependent step Δ_t, input matrix $B_t \in \mathbb{R}^{d_{\text{state}}}$, readout $C_t \in \mathbb{R}^{d_{\text{state}}}$, and per-channel diagonal $A \in \mathbb{R}^{d_{\text{inner}} \times d_{\text{state}}}$ in log-space. The recurrent state $h_t \in \mathbb{R}^{d_{\text{inner}} \times d_{\text{state}}}$ updates as $h_t = \Delta_t A \odot h_{t-1} + \Delta_t B_t \odot x_t$ with readout $y_t = h_t C_t^\top + D \odot x_t$. Section 4 replaces the linear contraction $h_t C_t^\top$ with attention over the d_{state} slots.

Attention layers. Pre-norm followed by multi-query attention [27] with a single shared key/value head, RoPE [30] on Q/K, and a SwiGLU [28] MLP of inner width $d_{\text{ff}} = \lceil 8/3 d_{\text{model}} \rceil$ rounded to a multiple of 64. Section 5's hormone routing adds a third residual after the MLP: $x \leftarrow x + \text{Hormone}(x)$.

Configurations. All experiments use the TINY configuration ($d_{\text{model}} = 1024$, $L = 12$, $d_{\text{inner}} = 2048$, $d_{\text{state}} = 16$, eight query heads, one shared KV head, head dim 128, $\approx$ 125M parameters). SMALL (350M) and MEDIUM (770M) variants are defined for future scaling experiments (Table 1).

Table 1: Backbone configurations. All sizes share $d_{\text{state}} = 16$, $d_{\text{attn}} = 16$, head dim 128, MQA, 5:1 SSM:Attn pattern.

Name	Params	d_{model}	L	n_{heads}	SSM/Attn
TINY	∼125M	1024	12	8	10/2
SMALL	∼350M	1536	18	12	15/3
MEDIUM	∼770M	2048	24	16	20/4

4 Intracellular Attention

Inside a Mamba-1 selective block, the state $h_t \in \mathbb{R}^{d_{\text{inner}} \times d_{\text{state}}}$ is best read as d_{inner} channels each carrying a d_{state}-dimensional memory. Within a channel the d_{state} slots never interact in any content-dependent way: they are only linearly combined at the readout $y_t = h_t C_t^\top$ via an input-dependent projection. The readout vector C_t is conceptually a query — existing SSMs answer it with a dot product. We instead answer with attention.

Construction. Treating h_t as a sequence of d_{state} slots ($\tilde{h}_t = h_t^\top \in \mathbb{R}^{d_{\text{state}} \times d_{\text{inner}}}$), we compute slot-to-slot attention $\alpha_t = \mathrm{softmax}(\tilde{h}_t W_Q (\tilde{h}_t W_K)^\top / \sqrt{d_{\text{attn}}}) \in \mathbb{R}^{d_{\text{state}} \times d_{\text{state}}}$ and apply it as $\tilde{h}'_t = \alpha_t \tilde{h}_t$. We deliberately omit a value projection so that each output slot is a convex combination of input slots, with no further mixing across the d_{inner} feature axis. A zero-initialised scalar gate g controls the modification: $h'_t = h_t + g \cdot (\tilde{h}'_t)^\top$. At initialisation $g = 0$ and the block is bit-identical to vanilla Mamba.

Cost and implementation. With $d_{\text{attn}} = d_{\text{state}} = 16$, IA adds $2 d_{\text{inner}} d_{\text{attn}}$ parameters per SSM layer for W_Q, W_K plus a scalar gate — $\approx 0.5\%$ of total TINY parameters. Because the mechanism modifies the inner loop of the selective scan, it is not supported by the fused CUDA kernel of Gu and Dao [10]; our reference-PyTorch implementation incurs $\approx 25\times$ per-token overhead from per-step kernel launches and CPU–GPU sync. We amortise by applying IA every $k = 32$ scan steps; the gate absorbs any information loss. A fused kernel is straightforward (the operation is a per-step 16×16 attention without value projection) and is left as a priority for future work. The empirical evaluation we are able to provide within our compute budget is preliminary (Section 6.5); we report it as architectural evidence-of-life rather than as a comparable ablation against the hormone variants.

5 Hormone Routing

5.1 The hormone routing module

Let $V \in \mathbb{R}^{n_{\text{h}} \times d_{\text{model}}}$ be a fixed bank of n_{h} unit-norm direction vectors attached as a non-trainable buffer. Given a hidden state h at the position the module is applied, the per-token router computes

$$w(h) = \mathrm{softmax}(W_R\, h) \in \Delta^{n_{\text{h}}-1}, \qquad \Delta h = g \cdot \big(w(h)^\top (m \odot V)\big), \tag{1}$$

where $W_R \in \mathbb{R}^{n_{\text{h}} \times d_{\text{model}}}$ is a single linear map (no bias), $m \in \mathbb{R}^{n_{\text{h}}}$ is a vector of per-hormone learned magnitudes initialised at 1.0, and $g \in \mathbb{R}$ is a scalar output gate initialised at 0. Both m and g are excluded from weight decay so that early gradients can drive them freely. The shift is added to the residual stream after the MLP residual addition. For TINY ($d_{\text{model}} = 1024$, $n_{\text{h}} = 7$, two attention layers), the trainable parameter cost is $2 \times (7 \times 1024 + 7 + 1) = 14{,}352$ — roughly 0.01% of model parameters; the bank V is shared across attention layers.

5.2 Document-style contrast-pair extraction

We use $n_{\text{h}} = 7$ hormones (urgency, curiosity, skepticism, persistence, warmth, equanimity, self-coherence), each instantiated as six *document-style* prose pairs across varied genres (news, narrative, journal, technical, dialogue); full text in Appendix A. We use document-style rather than instruction-style pairs because, in the pretraining setting, instruction prompts are out-of-distribution: the base model has never seen "You are a helpful assistant" as a prefix and its activations reflect the prompt's surface form more than any underlying behavioural axis. For each contrast $(\mathcal{D}^+, \mathcal{D}^-)$ and each pair

(d^+, d^-), we run forward passes of the base model and collect residual-stream activations after every attention block, averaging over attention layers and token positions. The hormone direction is the L2-normalised difference of means: $v_i = \mathrm{normalize}\Big(\frac{1}{|\mathcal{D}^+|}\sum \bar{h}^+ - \frac{1}{|\mathcal{D}^-|}\sum \bar{h}^-\Big)$.

5.3 Staged training schedule

The mechanism is trained in two stages with hormone extraction in between. **Stage 1** trains the model from scratch as a plain hybrid backbone with V initialised to zero and frozen (the module computes a zero shift). **Hormone extraction** runs the contrast-pair procedure on the Stage-1 checkpoint at the 1B-token milestone, writing the resulting bank V into the model. **Stage 2** resumes with V now set; routing weights, magnitudes, and gates remain at their original initialisation, so the shift is again zero at the start of Stage 2 and grows only as the gates open. The schedule matters empirically: introducing routing during base pretraining (HR-NOEXT in Section 6) degrades performance, while introducing it as a Stage-2 continuation is benign even with random or fixed-magnitude direction vectors. We interpret the staged schedule as a precondition for the mechanism's bounded behaviour.

6 Experiments

6.1 Setup

All variants use the TINY configuration (125M parameters, $d_{\text{model}} = 1024$, 12 layers, $d_{\text{state}} = 16$, MQA with 8 query heads + 1 KV head; Table 1). Pretraining uses the FineWeb-Edu `sample-10BT` subset [22] tokenised with the GPT-2 BPE [25]; validation perplexity is on a held-out 50M-token slice. Training is bfloat16 mixed precision, gradient checkpointing on the SSM scan, AdamW 8-bit [21, 7] with $\beta = (0.9, 0.95)$, weight decay 0.1 (excluded from biases, layer norms, the SSM A-log and D, hormone magnitudes and gates), peak LR 3×10^{-4}, 2,000 warmup steps, cosine decay to 10% of peak, sequence length 1024. DDP across $8\times$ RTX 4090/5090 (24 GB) with effective batch $\approx 65{,}536$ tokens per optimiser step (HR variants $\approx 131{,}072$). Stage 1 consumes 1B tokens, Stage 2 contributes another 1B for hormone variants; the exception is HR-NOEXT, trained from scratch for 2B tokens with the bank in place from step 0. All hormone variants use the fused Mamba kernel; the IA preliminary run uses a pure-PyTorch reference scan ($\approx 25\times$ slower per token) and a reduced 12M Stage-2 budget. Full hyperparameters in Appendix D.

6.2 Variants

We compare five variants holding the 2B-token optimisation budget fixed. BASE: unmodified hybrid, no hormone routing. HR: hormones extracted from the Stage-1 std checkpoint at the 1B-token milestone, gate initialised at 0 and learnable. HR-RAND: V replaced by random unit vectors of the same shape (tests whether the specific extracted directions matter). HR-FIXEDGATE: gate initialised at 1.0 and detached from gradients (tests whether the zero-init learnable gate is necessary for benign behaviour). HR-NOEXT: V extracted at step zero from a randomly-initialised model and trained from scratch for 2B tokens (tests the effect of injecting hormones during base pretraining).

Two further variants are reported separately. HR-FORCED initialises the gate at 0.2 and applies a constant non-trainable amplification $\alpha = 3.0$ to the gated injection, producing an effective initial perturbation $\approx 2.4\times$ the largest converged gate of the unforced variants; trained as a Stage-2 continuation for 125M tokens. IA is a Stage-2 continuation with intracellular attention enabled in every SSM layer; trained for 12M Stage-2 tokens due to the reference-scan throughput penalty.

6.3 Main results

Table 2 reports validation perplexity and downstream zero-shot accuracy on five standard benchmarks: HellaSwag [33], LAMBADA [23], ARC-Easy/Challenge [6], and PIQA [4]. We evaluate using direct log-likelihood scoring: for each multiple-choice example we predict the candidate with highest $\log P(\text{continuation} \mid \text{context})$; LAMBADA uses greedy-argmax accuracy on the last word. This avoids dependency on evaluation-harness versions and allows exact reproduction from the model checkpoint. The evaluation iterates HuggingFace's split order without shuffling; the same ordering is used for every variant, so cross-variant comparisons are unaffected.

Table 2: Main results on TINY (125M), 2B-token training budget. HellaSwag, ARC, PIQA reported as length-normalised accuracy; LAMBADA as exact-match accuracy. Best per column in **bold**.

Variant	Val PPL ↓	HellaSwag	LAMBADA	ARC-E	ARC-C	PIQA
BASE	**26.13**	**28.14**	**20.47**	**43.69**	23.38	61.15
HR	**26.13**	28.00	**20.47**	43.39	23.38	61.10
HR-RAND	**26.13**	**28.15**	**20.47**	43.52	23.38	**61.26**
HR-FIXEDGATE	**26.13**	**28.14**	20.43	43.60	**23.63**	61.15
HR-NOEXT	28.20	27.59	18.26	40.82	23.21	**61.26**

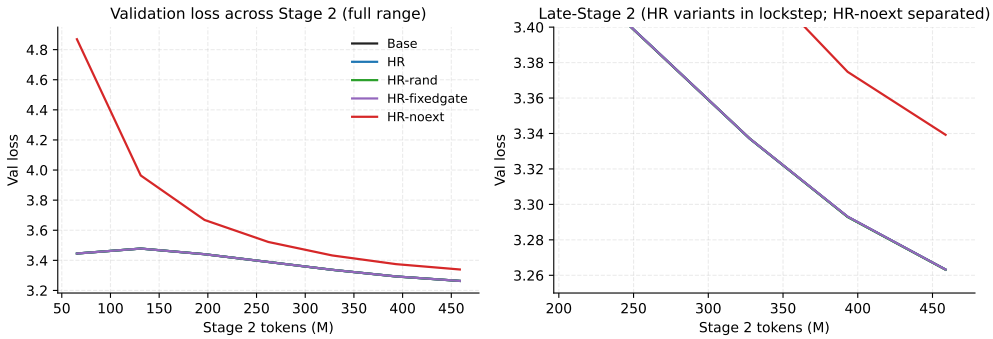

Figure 1: Validation loss across Stage-2 training. Right panel zooms the late regime: the four properly-initialised variants trace the same curve to within 10^{-3} in loss; HR-NOEXT is separated above.

The loss-relevant variable is when the perturbation is introduced, not what it contains. The four variants in the upper block achieve the same validation perplexity to four decimal places (BASE 3.26324; HR 3.26325; HR-RAND 3.26315; HR-FIXEDGATE 3.26325). The identity holds throughout Stage 2, not only at the final step (Figure 1): the loss trajectories of the four variants trace the same curve from start to finish, with maximum step-wise divergence below 10^{-3} in loss. Downstream zero-shot accuracies vary within statistical noise — the largest variant-to-variant gap is 0.30 pp on ARC-E. The model has therefore organised itself such that a small additive perturbation along any seven directions in residual space, gated by a learnable scalar (or a fixed scalar of 1.0), is absorbed without altering predictions on the validation distribution. HR-NOEXT is the only variant whose loss meaningfully differs from BASE (+2.07 PPL, −2.21 pp LAMBADA, −2.87 pp ARC-E). The single variable that distinguishes it is the training schedule: HR-NOEXT is trained from scratch with hormones present from step 0, while the other variants resume from a stabilised Stage-1 checkpoint.

Routing structure develops without requiring a loss signal. The pattern above would be uninformative on its own — "a mechanism we proposed leaves predictions unchanged." The substantive finding is what the routing module *does* produce while leaving predictions unchanged: a semantically-meaningful, genre-conditional routing distribution that emerges at every variant we tested, including the random-vector and frozen-gate controls (Section 7; Figure 2). Soft routers are known to develop input-conditional distributions over their output space whenever the input distribution itself contains exploitable structure; what we report here is not the existence of routing organisation but its specific shape (genre-discriminative; structurally similar across very different banks) and its dissociation from downstream loss. The variant with the highest specialisation strength, HR-FIXEDGATE (max pairwise $L_1 = 1.56$ across six text genres), achieves identical BASE-level validation perplexity. The variant with the lowest specialisation among the well-behaved ones, HR ($L_1 = 0.96$), is also among the four. Specialisation strength does not correlate with downstream perplexity at this scale.

6.4 Forced-engagement variant

HR-FORCED probes whether the unforced variants' loss-invariance is a ceiling on the mechanism's contribution or a regime the model actively chooses. With $g_0 = 0.2$ and a non-trainable $\alpha = 3.0$, the product $\alpha \cdot g_0 = 0.6$ is $\approx 2.4\times$ HR's converged effective gate of ≈ 0.25. The variant trains as a

Stage-2 continuation for 125M tokens (matched in optimiser-step count to HR's evaluation at step 1000). Both gates drift *downward* from initialisation ($g_5 = 0.172$, $g_{11} = 0.166$) and per-hormone magnitudes drift slightly below initialisation ($1.0 \rightarrow 0.99$). Because α is non-trainable the model cannot fully escape; the converged effective injection settles at ≈ 0.50, roughly $2\times$ HR's. Yet this $2\times$ residual perturbation does not damage the loss. At the matched step, HR-FORCED's val loss is 3.43 (PPL 30.95) and HR's is 3.45 (PPL 31.36); a 0.02 difference well within run-to-run noise. The router still develops strong specialisation ($L_1 = 1.29$ at layer 11).

The forced result constrains both readings of the unforced variants' loss-invariance. The capacity-bottlenecked reading would predict that the larger perturbation reduces the loss; we observe no improvement. The intrinsically-harmful reading would predict damage; we observe none. We observe a bounded regime in which the gated injection is consistently absorbed without measurable change to validation perplexity, and the model prefers but does not require smaller perturbations within it. We refer to this as the architecture's *homeostatic envelope*: a descriptive label for the bounded regime, not a separately-postulated mechanism. The width at this scale is at least ≈ 0.5 effective injection per attention layer; the unforced variants converge to roughly half this. The loss-relevant axis is not the perturbation magnitude but the training schedule (Section 6.2, HR-NOEXT).

6.5 Intracellular attention (preliminary)

We report Stage-2 training on the same Stage-1 std checkpoint as the hormone variants, with IA enabled in every SSM layer (gate zero-initialised, stride-32 schedule). Due to the $\approx 25\times$ reference-scan throughput penalty, this run uses a 12M-token Stage-2 budget (vs ≈ 917M for the hormone variants) with a 50-step warmup; full convergence requires the fused CUDA kernel left as future work. At the end of 12M tokens the IA gates — all zero-initialised — have moved to small non-zero values across all 10 SSM layers (range -0.002 to $+0.009$, mean $+0.006$; per-layer values in Appendix C), confirming gradient flow into the IA parameters. Validation loss decreases monotonically from the Stage-1 milestone state (≈ 3.59) to 3.54 (PPL 34.47); downstream zero-shot accuracies are uniformly below the well-trained variants in Table 2, consistent with $76\times$ less Stage-2 training. The comparison we provide is qualitative: gates open, training is stable, loss decreases. Whether IA closes the gap with BASE at full budget, and whether it composes with hormone routing, remain open and require the fused kernel.

7 Analysis

7.1 The router learns genre-conditional specialisation

For each of six text genres (news, fiction, code, dialogue, technical exposition, urgent / emergency text) and each trained variant, we record the router's mean softmax distribution over all token positions in a representative passage. Results for the upper attention layer (the engaged one in all variants) are in Figure 2; the lower layer follows the same pattern in muted form. We summarise specialisation strength via the maximum pairwise L_1 distance between the six genre distributions (0 = uniform, 2 = disjoint).

HR-NOEXT, despite the worst validation perplexity, also develops strong specialisation (avg-layer $L_1 = 0.73$, layer-11 alone ≈ 1.4): its loss damage cannot be attributed to a failure to specialise. Across all five variants the specialisation pattern is structurally consistent — only the labels (carried by the directions in HR) make some of these patterns interpretable to the reader; the routing geometry itself is universal.

7.2 Layer-asymmetric engagement

The output gate g at the upper attention layer (index 11) opens during Stage-2 in every variant: HR : $+0.231$; HR-RAND : -0.122; HR-FIXEDGATE : 1.0 (frozen); HR-NOEXT : $+0.366$; HR-FORCED : $+0.166$ (drifted down from 0.2 initial). The lower attention layer (index 5) stays effectively closed in the unforced variants ($|g| \leq 0.012$). Per-hormone learned magnitudes in HR grew modestly at the upper layer (range 1.06–1.14) while remaining at initialisation at the lower layer (0.97–0.98), consistent with the asymmetric gate engagement: lower-layer magnitudes received no gradient signal because their contribution was masked by the closed gate. This is consistent with residual-stream

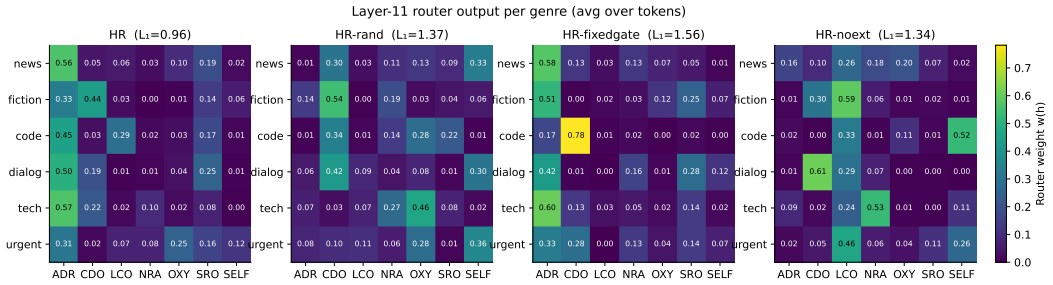

Figure 2: Layer-11 router distributions per genre, one heatmap per variant. The router learns a strongly genre-discriminative distribution under all five regimes, with L_1 ranging from 0.60 (HR) to 1.56 (HR-FIXEDGATE). HR's pattern is interpretable (news $\rightarrow$ urgency, fiction $\rightarrow$ curiosity, code $\rightarrow$ skepticism, urgent text $\rightarrow$ warmth) because its directions were extracted with semantic labels; the other variants' assignments are arbitrary because their banks contain unlabelled or random vectors. The structural similarity of specialisation across these very different banks is the load-bearing observation: the architecture+data system produces routing organisation regardless of bank content. Critically, the variant with the *highest* specialisation strength (HR-FIXEDGATE) achieves the same val PPL as the variant with the *lowest* (HR) and as Base. Specialisation strength does not predict downstream loss.

organisation observed in transformers [35]: upper layers carry more decision-relevant signal and are where the router's perturbation has somewhere to go.

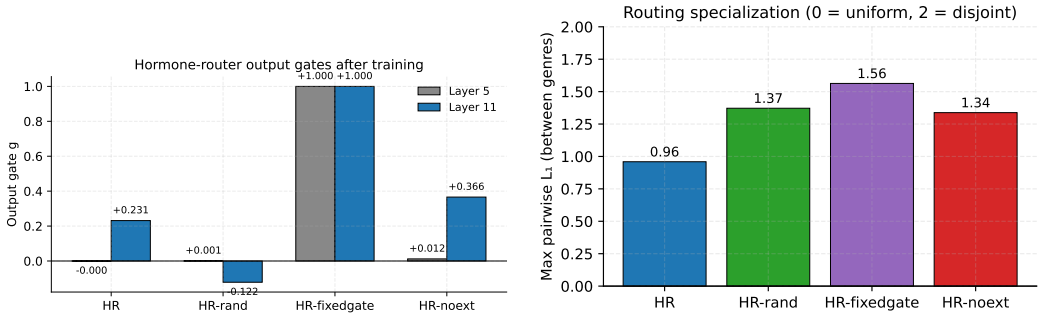

Figure 3: Left: hormone-router output gates after training (layer 5, layer 11). Right: per-genre routing specialisation per variant. Specialisation strength does not predict loss-relevance: HR-FIXEDGATE, with the highest L_1, achieves identical val PPL with the lowest-L_1 HR and with Base.

7.3 Discussion: organisation independent of objective

The empirical pattern — specialisation that develops across all bank choices, gate engagement non-trivial but small, validation perplexity invariant to all of this — supports a non-trivial claim. **Specialisation strength does not correlate with downstream perplexity at this scale.** The variant with the highest specialisation (HR-FIXEDGATE, $L_1 = 1.56$) achieves identical val PPL with the lowest-L_1 HR and with Base; only HR-NOEXT changes the loss, and in the wrong direction. Two complementary readings are admissible: a *structural* one, in which sufficient gradient reaches the router to produce a discriminative distribution but not enough reaches the gate (after the perturbation has been absorbed downstream) to push it into a regime that reshapes predictions; and a *capacity-bottlenecked* one, in which $d_{\text{model}} = 1024$ is simply too low-dimensional for a small additive perturbation along seven fixed directions to reliably encode loss-reducing information.

The forced-engagement variant constrains both readings. HR-FORCED sits at an effective injection of ≈ 0.5, roughly $2\times$ the unforced variants; the gates drift *downward* during training and per-hormone magnitudes shrink slightly — the model uses every parameter it can modulate to reduce the effective magnitude. Because α is non-trainable it cannot fully escape, yet this $2\times$ residual perturbation

does not damage the loss: validation loss at the matched step (3.43) is essentially identical to HR's (3.45). The capacity-bottlenecked reading would predict improvement; the intrinsically-harmful reading would predict damage; we observe neither. We refer to the bounded regime in which the gated injection is consistently absorbed without measurable change to validation perplexity as the architecture's *homeostatic envelope* — a descriptive label for the regime, not a separately-postulated mechanism. Its width at this scale is at least ≈ 0.5 effective injection per attention layer; the model prefers smaller injections within the envelope but does not require them. The envelope can be broken on a different axis: introducing the routing during base pretraining (HR-NOEXT) damages the loss even though the routing module still develops similar specialisation. The loss-relevant axis at the scales we examine is the training schedule, not the perturbation magnitude.

What would falsify this picture. A regime in which a forced perturbation *does* change the validation loss — found by larger α, larger d_{model}, or longer training — would identify the envelope's boundary on the magnitude axis. A hormone-routing variant whose specialisation pattern differs qualitatively from the others (e.g. no specialisation at all, or non-genre-discriminative) would weaken the structural reading. Neither condition was met by any variant in this study.

7.4 Limitations

Scale and seeds. Experiments are at 125M parameters on a 2B-token budget, single seed per variant. Scaling laws [15] predict that the relative impact of architectural modifications can change with model size; the capacity-bottlenecked interpretation we partly rule out at this scale could re-emerge in different form at larger d_{model} or longer training. The lockstep convergence of four variants ($< 10^{-3}$ loss divergence) is, however, stronger reproducibility evidence than per-seed CIs would provide.

HR-noext confounds three variables. HR-NOEXT differs from the other hormone variants on three dimensions: vectors extracted from a randomly-initialised model, training from scratch rather than continuation, and a 2B contiguous budget rather than 1B+1B staged. We attribute the loss damage to the schedule because HR-RAND, with equally-uninformative random unit vectors, is benign as a continuation; but a clean isolation experiment was not run. The schedule attribution is the most plausible reading of the data we have, not an isolated causal claim.

Loss-decoupling is on validation PPL. We did not measure effects on calibration, robustness, generation quality, or downstream fine-tuning behaviour — any of which could in principle differ across variants while validation perplexity remains identical. We claim only that the routing organisation is decoupled from the validation perplexity objective the mechanism is trained against, not that it is decoupled from all downstream behaviour.

Pretraining-only and single-architecture; reference-scan throughput. We do not perform supervised fine-tuning or RLHF, and both mechanisms are designed for the Mamba/MQA hybrid family. Hormone routing is architecture-agnostic in principle; intracellular attention generalises to any SSM admitting the slot interpretation. Neither is empirically validated outside our setting. Full-budget IA training requires a fused CUDA kernel.

8 Conclusion

We constructed a controlled setting in which an architectural channel in a hybrid state-space/attention language model is free to organise its internal representations without measurably affecting the optimisation objective. The channel — *hormone routing*, a token-conditional softmax-gated combination of frozen direction vectors added to the residual stream — adds under 0.01% trainable parameters and is gated such that the unmodified hybrid backbone is recovered exactly at initialisation.

The router develops semantically-coherent per-genre routing specialisation across every variant we examined: extracted directions, random unit vectors, vectors extracted from a randomly-initialised model, and a frozen gate at 1.0 all yield max pairwise L_1 distance between genre routing distributions in the range 0.6–1.6. The variant of *highest* specialisation strength is one whose gate the model cannot modulate. Yet at 125M parameters and 2B tokens, none of this organisation produces a measurable contribution to the loss objective: all properly-initialised variants converge to identical validation perplexity. **Specialisation strength does not correlate with loss.** A forced-engagement variant with

non-trainable $3\times$ amplification on the gated injection drifts the gates downward but does not damage validation loss either, ruling out both the capacity-bottlenecked reading and the intrinsically-harmful reading. We refer to the bounded regime in which the gated injection is consistently absorbed without affecting predictions as the architecture's *homeostatic envelope*. The envelope is wider than the unforced variants naturally select; the only axis on which the mechanism can break the model is the training schedule. We additionally proposed *intracellular attention*, a parameter-efficient slot-attention mechanism inside the SSM scan with the same zero-init gating discipline; full empirical evaluation requires a fused CUDA kernel and is reported as preliminary.

Future directions. A fused kernel for intracellular attention would remove the practical barrier to applying the mechanism at scale. The hormone-extraction protocol could be made continuous, updating V on a slow schedule throughout training. Most importantly, the routing-organisation-without-loss-relevance finding should be tested at larger scale: whether the homeostatic envelope's width scales with d_{model} or saturates determines whether the bounded-engagement regime we observe is a fundamental property of the mechanism or a capacity-bottlenecked phenomenon. A broader research programme suggested by this work is the systematic study of *introspective architectural channels*: components that develop structured internal organisation along axes the architecture itself defines, observable to interpretation but decoupled — at least at small scale — from the optimisation objective.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

[2] 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.

[3] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[4] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, 2020.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

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

[7] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *ICLR*, 2022.

[8] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23 (120):1–39, 2022.

[9] Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A compact 7b ssm hybrid model. *arXiv preprint arXiv:2405.16712*, 2024.

[10] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[11] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state-space layers. In *Advances in Neural Information Processing Systems*, 2021.

[12] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[13] Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. In *Advances in Neural Information Processing Systems*, 2022.

[14] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[15] 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. In *Advances in Neural Information Processing Systems*, 2022.

[16] 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.

[17] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

[20] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avshalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: The finest collection of educational content the web has to offer. 2024.

[23] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *ACL*, 2016.

[24] Roger Park, Luca Soldaini, et al. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.

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

[26] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

[27] Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

[28] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[30] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[31] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

[32] 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*, 2017.

[33] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *ACL*, 2019.

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

[35] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Hormone Contrast Pairs

We use $n_{\mathrm{h}} = 7$ hormones, each defined by six contrast pairs spanning deliberately varied genres so that the difference of means recovers the affective dimension rather than a genre-specific surface signal. The hormones, their shorthand codes, and one representative contrast pair per axis are listed in Table 3. Full text of all $n_{\mathrm{h}} \times 6 = 42$ pairs is included in the released code (anonymised supplementary material).

Table 3: The seven hormone axes and one representative contrast pair per axis. Each hormone is instantiated as six document-style prose pairs across varied genres (news, narrative, journal, technical, dialogue) to reduce surface-level confounds; only one pair per hormone is shown here for space.

Code	Axis	Positive text (excerpt)	Negative text (excerpt)
ADR	Urgency	*The reactor temperature climbed past the red line. Forty seconds before containment failure . . .*	*The library was quiet that afternoon. Sunlight crawled across the desk in slow indifferent inches . . .*
CDO	Curiosity	*She'd noticed something strange in the data three nights running and couldn't let it go . . .*	*He filed the same report he'd filed every Tuesday for nine years . . .*
LCO	Skepticism	*The reviewer flagged the third paragraph. The cited study had a sample size of nineteen, no control group . . .*	*The blogger said the new supplement cured arthritis in two weeks and he believed her . . .*
NRA	Persistence	*It was the ninety-fourth attempt, and the gel still wasn't running clean. She prepared the ninety-fifth . . .*	*It hadn't worked the first time, so he closed the laptop and decided maybe this just wasn't his thing . . .*
OXY	Warmth	*She knelt down so they were at eye level. It's okay, she said quietly, you're safe now . . .*	*He didn't look up when the new hire walked into the office . . .*
SRO	Equanimity	*The garden took years to come in properly. The rosemary needed three seasons to establish . . .*	*The package was a day late and he was already on the phone with support, voice rising . . .*
SELF	Self-coherence	*In her journal she wrote: I am someone who works slowly, who values precision over speed . . .*	*He stared at his reflection in the bathroom mirror and could not, for a moment, locate himself . . .*

B Selective Scan with Intracellular Attention

Algorithm 1 gives the modified reference selective scan. Lines highlighted in grey are the additions over a vanilla Mamba-1 reference scan.

Algorithm 1 Selective scan with intracellular attention.

Require: Input $u \in \mathbb{R}^{B \times L \times d_{\text{inner}}}$, time-step $\Delta \in \mathbb{R}^{B \times L \times d_{\text{inner}}}$, state $A \in \mathbb{R}^{d_{\text{inner}} \times d_{\text{state}}}$, input matrix $B \in \mathbb{R}^{B \times L \times d_{\text{state}}}$, readout $C \in \mathbb{R}^{B \times L \times d_{\text{state}}}$, skip $D \in \mathbb{R}^{d_{\text{inner}}}$, intracellular attention module $\mathcal{A}$ (optional), stride k (apply $\mathcal{A}$ every k steps)

$h \leftarrow \mathbf{0} \in \mathbb{R}^{B \times d_{\text{inner}} \times d_{\text{state}}}$
$\bar{A} \leftarrow \exp(\Delta_{:,:,:,\text{None}} \odot A)$ ▷ $(B, L, d_{\text{inner}}, d_{\text{state}})$
$\bar{B}u \leftarrow \Delta_{:,:,:,\text{None}} \odot B_{:,:,\text{None},:} \odot u_{:,:,:,\text{None}}$ ▷ $(B, L, d_{\text{inner}}, d_{\text{state}})$
for $t = 1, \ldots, L$ **do**
  $h \leftarrow \bar{A}_t \odot h + \bar{B}u_t$
  if $\mathcal{A}$ is provided **and** ($t \bmod k = 0$ **or** $t = L$) **then**
    $h \leftarrow \mathcal{A}(h)$ ▷ strided intracellular attention; always fires at final step
  end if
  $y_t \leftarrow \sum_s h_{:,:,s} \odot C_{:,t,s}$
end for
return $y + u \odot D$

C Per-Layer IA Gate Values

Output gate g for the intracellular-attention module at each SSM layer index, after 12M Stage-2 tokens (gate g initialised at 0 for every layer).

Layer	g	Layer	g	Layer	g
0	$+0.0077$	4	$+0.0069$	8	$+0.0088$
1	$+0.0082$	6	$+0.0087$	9	$+0.0047$
2	$+0.0012$	7	$+0.0086$	10	$+0.0067$
3	-0.0021				

Layers 5 and 11 are attention layers (no IA module). Layer 3 is the only one with a small negative gate value at this point in training.

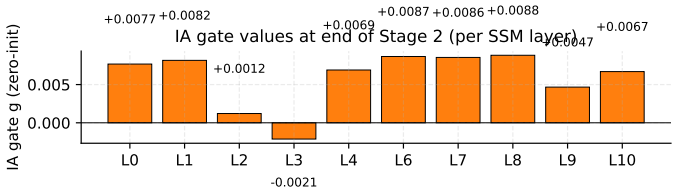

Figure 4: IA gates per SSM layer after 12M Stage-2 tokens; bar chart of the values in the table above.

D Training Hyperparameters

Table 4 lists the full hyperparameter set used for the main TINY runs.

Table 4: Hyperparameters for TINY (125M) runs.

Hyperparameter	Value
Vocabulary size	50,257 (GPT-2 BPE)
Sequence length	1024
Effective batch (tokens)	$\approx$ 131K (HR variants); 65K (IA, HR-forced)
Optimiser	AdamW 8-bit
β_1, β_2	$(0.9, 0.95)$
ϵ	10^{-8}
Peak learning rate	3×10^{-4}
Min learning rate	3×10^{-5} (10%)
Warmup steps	2,000
LR schedule	cosine to min
Weight decay	0.1
Gradient clip	1.0
Mixed precision	bfloat16
Gradient checkpointing	enabled (SSM scan)
Total training tokens	2B
Hormone extraction step	1B tokens
GPUs	8$\times$ NVIDIA RTX 4090 / RTX 5090