FIELDNET: A SOURCE-FREE NORMALIZER FOR ANY TRANSITION, AND CLOSED-FORM UNDER TIME HOMOGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Selectivity, bounded normalization and linear complexity are treated as a trilemma, reached in three steps: an input-dependent decay leaves the kernel lag-dependent, and a lag-dependent kernel has no closed-form mass, so the state cannot be normalized. The first two steps are right. The third holds only if the readout is unnormalized: under a posterior-mean readout (A3) the normalizer is the kernel's own mass, driven by the constant source $\mathbf{1}$, and it exists under any causal transition, content gating included. The object is a *coupled dual recursion*—one operator at two sources, not two recurrences—and the field locates the absence in the readout itself. *As stated, the trilemma is a statement about unnormalized readouts; under A3 it does not hold.*

The paper forms a closed loop. *Derived*: under A3 the normalizer is forced to be the kernel's own mass, and the coupled dual recursion it forces holds under any causal transition. *Measured*: a companion mass and weak content dependence compose, and that composition is the best model here—the 400M dual arm leads the matched baseline on WikiText-103 (40.22 against 57.13), seven-task zero-shot (46.80 against 38.17) and PG19 (26.01 against 28.67). *Direction*: the strong-gating end is supported by the same implementation; it opens the effect-size question.

1 INTRODUCTION

The trilemma, and where it breaks. The efficient-sequence-model literature has converged on a trilemma: selectivity, bounded normalization and linear complexity, of which a design appears to keep two. Mamba, RWKV and GLA take selectivity and linear complexity, and read the state out unnormalized (Gu & Dao, 2024; Peng et al., 2023; Yang et al., 2024). RetNet and normalized linear attention take normalization and linear complexity, and give up selectivity (Sun et al., 2023; Katharopoulos et al., 2020). S4D takes linear complexity and normalization, and reads the state out without a mass (Gu et al., 2022b). No architecture in the family is reported to keep all three, and the reason given is a chain of three steps that, we argue, is broken at the third.

(i) selectivity requires an input-dependent decay;

(ii) an input-dependent decay leaves the kernel lag-dependent;

(iii) *therefore* the state cannot be normalized, because a lag-dependent kernel has no closed-form mass.

Steps (i) and (ii) are right, and we do not dispute them. Step (iii) is where we differ, and the field's own analyses confirm the precise form of the error: Han et al. (2024) reformulate the selective SSM as linear attention and identify *no attention normalization* as one of six distinctions—linear attention divides by $\boldsymbol{Q}_i \boldsymbol{Z}_i$ so that the attention weights sum to one, while the recurrence of eq. (11) in their formulation has no such denominator. Dao & Gu (2024) establish the state-space duality within which that comparison is exact. What they do not examine is whether the absence is *forced* by the selectivity. It is not. Step (iii) holds *only if the readout is unnormalized*: a posterior mean demands a normalizer by definition, and that normalizer—the mass driven by the constant source $\mathbf{1}$—is available under any causal transition. Dropping the unstated premise restores the third leg, and the restored

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.

object is a *coupled dual recursion*: a signal state and the normalizer that reads it out, sharing one Λ^{Δ_t}.

What "dual" means here: one operator, two sources. The word has been used for *two* recurrences that stay independent and meet only through an external mechanism—two streams, or a second recurrence over a structure read by the first. Ours is not that object. Let $\mathcal{T}_\Lambda$ denote the lag-weighted operator; then the signal state, the mass and the readout are three evaluations of it:

$$(\mathcal{T}_\Lambda \rho)_t = \sum_{s \le t} \Lambda^{\sum_{k=s+1}^{t} \Delta_k} \rho_s, \qquad E = \mathcal{T}_\Lambda \rho, \quad m = \mathcal{T}_\Lambda \mathbf{1}, \quad \tilde{F}_t = \frac{(\mathcal{T}_\Lambda \rho)_t}{(\mathcal{T}_\Lambda \mathbf{1})_t}. \tag{1}$$

The mass is the signal's *dual* under the *same* operator: one transition, applied to the constant source $\mathbf{1}$ instead of to ρ. There is no second operator, no second decay and no second initialization, so the two sides cannot be independently parameterised and need no external mechanism to meet; they share Λ^{Δ_t}, and that sharing is forced by A3 rather than added by us. Proposition 9 makes the denominator unique up to a scalar, so *coupled dual* describes the object and is not a naming preference. The distinction is measurable. Give the mass a decay of its own, $\mu = \lambda^p$ with $p \neq 1$, and the single pairing becomes two unrelated ones: the readout stops being a posterior mean, and the M1 constancy probe moves five orders of magnitude ($2.29 \times 10^{-7} \to 2.03 \times 10^{-2}$ at $p = 2$); swapping the divisor to rms inside one checkpoint, weights unchanged, moves it by 1.9×10^5. A broken coupling is therefore not a weaker model but a different estimand, which is what M1–M3 test (Section A.3).

A closed loop, and the direction it opens. The paper forms a closed loop: a derivation (Section 3), the predictions that follow (P1, P2, M1–M3), and their validation in the released implementation and at 400M. **Derived:** a companion mass and *any* causal transition compose—Proposition 1 states this, and the released implementation exposes transition and normalization as two independent axes. **Measured:** at the weak end of that axis the composition is the best model in the paper—the 400M dual arm modulates its rates through a bounded permeability factor while carrying the mass, and it leads the matched baseline on all three metrics. **Direction:** the strong-gating end is supported by the same implementation and outlined in Section 6; its effect size is not measured here.

Contributions. **1. The third step of the mainstream inference is false.** Steps (i) and (ii) are right; step (iii)—"lag-dependent kernel, therefore no normalization"—holds only under an unnormalized readout. A posterior-mean readout demands a normalizer, and Proposition 1 shows that the mass driven by the constant source $\mathbf{1}$ supplies one under *any* causal transition. Selectivity and bounded normalization are therefore separable capabilities, not one package; this is a statement about the estimand and it is proved before any experiment is read. A probe over all nine combinations of transition and normalization in the released implementation confirms the structural claim as a *structural verification* of the composite coordinate: the bounded ratio is 1.000 with M1 drift 4.6×10^{-7} in every cell (Table 2). What the probe establishes is boundedness; whether the composite is a *better* model is Section A.2's question.

2. A coupled dual recursion, and what C3 adds. A3 and C3 together determine a signal state $E_t = \Lambda^{\Delta_t} E_{t-1} + \rho_t$ and a companion mass $m_t = \Lambda^{\Delta_t} m_{t-1} + \mathbf{1}$, coupled through one shared Λ^{Δ_t}. The coupling is not a design choice: the only source-free denominator compatible with the signal's decay is the mass driven by $\mathbf{1}$, and removing it stops the readout being a posterior mean. The readout E_t/m_t is a convex combination of sources and is bounded by the largest one. What C3 adds is not the mass but its *form*: closed form, a horizon law $1/\alpha_{\min}$, and a rate spectrum that is a constant of the model rather than a quantity realised per input. Bernstein–Widder is the enabling tool—a 1929 theorem that turns heavy-tail dependence into a mixture of exponentials, so power-law memory and finite-state recurrence stop being in tension; the novelty is the matching, not the tool.

3. Two testable consequences. P1—extrapolation with no positional encoding—and P2—the rescaling $\Delta_t \mapsto s\Delta_t$, a rigid translation of the log-rate spectrum. P2 is defined whenever the transition is *multiplicative in position*—rates that do not themselves depend on the lag—so C3 is sufficient for P2 but not necessary: content-dependent rates admit it too, and one of the four runs below takes that road. What P2 discriminates is multiplicative position dependence from lag-dependent gating such as a content-similarity kernel; it does not rank C3 against content gating. It holds on four runs at two scales and both transition modes: the shape transfers, the magnitude does not.

Table 1: **A controlled comparison at the matched ∼400M pair:** shared tokenizer, data, data order, batch composition and schedule. PG19 PPL at $L = 2{,}048$ (anchored, $R = 16$); WikiText-103 word-level perplexity; mean zero-shot accuracy over seven tasks. Lower is better for both perplexities. The dual instance (NORM) leads the baseline on all three: PG19 PPL 26.01 against 28.67, WikiText-103 40.22 against 57.13, and downstream 46.80 against 38.17. One seed, and the arms differ in code version as well as in normalization (Appendix 31), so this is a controlled comparison and not an attribution to the companion mass alone. The single-recursion instance is also trained at 1.3B, where it trails the width-matched baseline on quality and leads it by three orders of magnitude on extrapolation (Tables 25, 26); the dual recursion is not trained at that scale.

Model	PG19 PPL	WikiText wPPL	7-task mean
Llama-2 405M (Touvron et al., 2023)	**28.67**	57.13	38.17
FIELDNET 404M UNNORM (single)	34.33	52.55	46.36
FIELDNET 404M NORM (dual)	26.01	**40.22**	**46.80**

4. A controlled comparison, and the weak-gating measurement. At the matched ∼400M pair—shared tokenizer, data, data order, batch composition and schedule—the dual configuration leads the baseline on WikiText-103 (40.22 against 57.13), on seven-task zero-shot downstream (46.80 against 38.17) and on PG19 (26.01 against 28.67) (Table 1). That arm is a *permeability* model, so it is not on the time-homogeneous coordinate: dual recursion plus weak content dependence is measured and working. FIELDNET ships the dual recursion at 400M and the single recursion at 400M and 1.3B; the strong end of the transition axis is supported by the same implementation (Appendix 19); its effect size is the direction Section 6 opens.

2 RELATED WORK

Each architecture in this family resolves the trilemma by giving up one leg. S4D (Gu et al., 2022b;a) gives up the mass: it keeps a position-only readout without a companion denominator. Mamba (Gu & Dao, 2024), RWKV (Peng et al., 2023) and GLA (Yang et al., 2024) give up normalization and read the state out directly. RetNet (Sun et al., 2023) and normalized linear attention (Katharopoulos et al., 2020) give up selectivity and keep a denominator. FIELDNET occupies the coordinate that keeps all three, and the move that opens it is the *readout*, not the transition. Extended positioning, and the pricing of each departure, are in Appendix 48 and Table 21.

Why Mamba does not already do this. Mamba, RWKV and GLA read the state out directly; that the readout is unnormalized is a property of those constructions, not a consequence of relaxing C3. Han et al. (2024) make the point exact by rephrasing the selective SSM as linear attention and isolating *no attention normalization* as one of six distinctions: linear attention divides by $\boldsymbol{Q}_i \boldsymbol{Z}_i$ so that the weights sum to one, and the selective recurrence has no such denominator. Removing it costs accuracy and the standard deviation of token ℓ_2 norms grows markedly without it, particularly in the last layers—long tokens dominate and short ones stop carrying semantics. Their conclusion places this among the four *less* crucial distinctions, the core two being the forget gate and the block design, so the paper records the absence and locates it in the readout without asking whether the absence is forced. Dao & Gu (2024) supply the duality under which the comparison is exact. What follows is not a repair to their construction but an observation about the object they omit: it is a source-free normalizer, it is compatible with a content-dependent transition, and it exists.

What relaxing C3 does and does not forfeit. Mamba (Gu & Dao, 2024), RWKV (Peng et al., 2023) and GLA (Yang et al., 2024) relax time homogeneity to obtain selectivity, and in each case the normalizer is absent or content-dependent. Neither is forced by the relaxation: the companion mass is driven by the constant source $\mathbf{1}$, so $m_t = \Lambda(u_t) m_{t-1} + \mathbf{1}$ stays signal-source-free for *any* transition, content-dependent ones included, and the readout stays a convex combination bounded by $\max_s \|\rho_s\|_\infty$. What relaxing C3 forfeits is the closed form and the horizon forecast—not boundedness, and not P2. The two capabilities the field has treated as one package are separable: Proposition 1 establishes it structurally and the weak end of the transition axis is measured (Section A.2).

Why A3 is the load-bearing choice. A3 is not claimed necessary—what each alternative forfeits is priced in Appendix 47, and replacing it is a different paper. It is load-bearing in a narrower and checkable sense: it is the assumption under which keeping C3 is *worth* defending. Drop A3 and the denominator goes with it, so the semigroup buys only an unnormalized accumulation and the defence of C3 has no consequence to point at; keep A3 and C3 acquires one, a definite, closed-form and testable object. This is why the two moves are ordered and not merely conjoined: A3 first, as premise; C3 second, as the contested choice that premise makes affordable.

The difference from RetNet and from linear attention, in form. The distinction is a conjunction, and each of the two architectures named above satisfies one half of it. RetNet divides by a denominator, but that denominator is a channel-wise normalization applied at the output, not the mass of the kernel itself. Normalized linear attention divides by the mass of a kernel, but its kernel weight is $\phi(q_t)^\top \phi(k_s)$: content dependent, so $G(t,s) \neq \Lambda^{t-s}$. Only under time homogeneity does the companion mass have the constant source $\mathbf{1}$,

$$m_t = \sum_{s<t} \Lambda^{t-s}\,\mathbf{1}, \tag{2}$$

which makes it the *partition function of the kernel*: it carries no input information and it has a closed form, so the readout is a posterior mean over *recency* with a prior fixed by the spectrum alone, whereas in linear attention the same quotient is a posterior mean over *content similarity*, recomputed at every step from the current query—different estimators, not different names for one estimator. The difference is measurable rather than literary: replacing the companion mass by an output normalization inside one checkpoint—the operation RetNet performs—moves the constant-input readout from 2.29×10^{-7} to 4.28×10^{-2}, a factor of 1.9×10^{5}, with weights, spectrum and forward graph untouched (Table 7). Proposition 10 states the difference; Section 3.2 fixes the division of labour between the estimator and C3.

3 THE DERIVATION, ONE STEP AT A TIME

The claim is: *if* the items hold, *then* the recurrence follows (Appendix 14). Of the five steps, two are the moves this paper is about—step 2, keeping the transition time-homogeneous, and step 5, the companion mass it forces—and they are the two given here in full; steps 1, 3 and 4 are supporting and are in Appendix C. The two are not independent: keeping C3 alone gives a time-homogeneous single recursion, and adopting A3 alone gives a normalized readout without a shared decay. Their conjunction produces the object this paper is about. Table 4 is the map for all five.

3.1 STEP 2: KEEPING TIME HOMOGENEITY

The state must compose: training scans in chunks, inference runs token by token, and both must compute the same function, so $G(t,s) = G(t,j)\,G(j,s)$—the consequential way to have it being an operator that depends on the elapsed interval alone.

Constraint 1 (Time-homogeneous transition (C3))**.** *$G(t,s) = \Lambda^{t-s}$, so $G(t,i) = G(t,j)\,G(j,i)$ holds with a time-invariant Λ.*

With elapsed position $\Delta_t > 0$ this forces the decay-gated recursion

$$E_t = \Lambda^{\Delta_t} E_{t-1} + \rho_t, \qquad G(t,s) = \prod_{k=s+1}^{t} \Lambda^{\Delta_k}. \tag{3}$$

Two reasons to keep C3, neither of which presupposes the conclusion. The motivations below are stated before any of them is used, and none mentions the companion mass: each is a property of the transfer operator alone, so none is circular. *(i) Extrapolation becomes a property of the form, not a repair.* Under C3 the kernel is a function of the lag alone, so it is defined for every $\Delta_t > 0$: length is an argument of the kernel rather than an index into a table. *(ii) The normalizer is a partition function of the kernel.* Under C3 the kernel mass is a functional of the kernel alone and carries no input information; under $\Lambda = \Lambda(u_t)$ the same quotient recomputes its weights from the current query, so normalization and content selection are one operation. Neither is an argument against content gating, which buys selectivity and is a real gain; they are the properties forfeit when it is bought.

The price, and why we pay it. Keeping C3 costs the ability to forget a token faster than its own rate allows; the pricing of that cost, and why it is worth paying under A3, is in Appendix A.

3.2 Step 5: the companion mass A3 + C3 determine

(3) accumulates signal but is not yet an estimate: $E_t = \sum_s G(t,s)\rho_s$ has a scale depending on how much history arrived, so it is bounded neither across positions nor independently of L. Something must normalize it, and what that is decides what the model computes.

Axiom 1 (Estimation by the posterior mean (A3))**.** *The field is estimated by its IFT posterior mean (Enßlin et al., 2009) under a Gaussian field prior, not by an unnormalized accumulation.*

A3 is recorded as an axiom because the derivation needs it named, not because it is in dispute: under a Gaussian prior the posterior mean *is* the classical Nadaraya–Watson mean, so nothing in it was chosen to produce a companion mass. What each alternative forfeits is in Appendix 47.

$$\bar{E}(x) = \frac{\int E\, p(\rho \mid E)\, p(E)\, \mathcal{D}E}{\int p(\rho \mid E)\, p(E)\, \mathcal{D}E} \xrightarrow{\text{Gaussian, linear}} \frac{\sum_s G(x,x_s)\,\rho_s}{\sum_s G(x,x_s)}, \tag{4}$$

a Nadaraya–Watson mean with kernel G (Appendix 17.3), which also closes the linearity left open at The denominator is a second kernel functional of the *same* operator, evaluated at the constant source—not an architectural option but the form the conjunction fixes:

$$m_t = \Lambda^{\Delta_t} m_{t-1} + \mathbf{1}, \tag{5}$$

$$\tilde{F}_t = E_t \oslash (m_t + \epsilon). \tag{6}$$

Proposition 1 (The mass does not depend on the transition coordinate)**.** *Let the transition be any causal $G(t,s;u)$ for which the accumulation admits a recursion, and let the readout be the posterior mean of A3. Then the normaliser is the companion mass $m_t = \sum_{s<t} G(t,s;u)\mathbf{1}$, driven by the constant source $\mathbf{1}$; under a first-order transition it evolves by $m_t = \Lambda(u_t)m_{t-1} + \mathbf{1}$, sharing the signal's transition and carrying no information about the signal source ρ—this is the sense in which we call it signal-source-independent. It is* not *independent of the input: under a content-dependent transition the same $G(\cdot;u)$, and hence the same m_t, depends on u. What m_t never sees is the source sequence ρ that the numerator accumulates, which is exactly what makes N_t/m_t a convex combination of the ρ_s and therefore bounded,* independently of whether the transition is time-homogeneous. *Selective forgetting and bounded normalisation are separable: nothing in this statement requires C3.*

This is the proposition that breaks the third step, and it is stated before any experiment is read. Step (iii) concludes "no normalization" from "lag-dependent kernel"; the proposition above shows the conclusion holds only under an unnormalized readout. Under A3 the mass is a functional of the constant source, not of the transition form, so selectivity and bounded normalization are separable—and the trilemma as commonly stated is a statement about unnormalized readouts. Nothing in the statement mentions time homogeneity, and nothing in it is empirical: it is a claim about the estimand, and the experiments of Section 5.1 and Table 2 test whether the shipped implementation has that form. What the proposition does *not* touch is the closed form: a lag-dependent kernel still has no closed-form mass, which is why C3 remains the coordinate on which the object can be read off before training.

The two axes really are independent, and one number bounds the second. The transition and normalisation coordinates are separate parameters in the released code, not two settings of one switch: the transition enters only through the modulator applied to the rates, the normalisation only through the divisor, and the mass scan is driven by the constant source $\mathbf{1}$ under either. So the coordinate that combines them—content gating with a companion mass—is constructible without new machinery. The implementation exposes it; Section 6 discusses it as the direction this paper opens (Appendix 49). One precondition is quantitative: the scan carries a span contract $\alpha_{\text{eff},\max} \cdot \text{chunk} \leq$ `log_shift_max`, exceeding which silently truncates the receptive field to the chunk. With the released defaults the permeability arm passes (span 150) while content gating at its default clamp overruns it by $31\times$ (Appendix 19). At a fixed chunk, *the span contract bounds how selective content gating can be*—a trade the position-only coordinate does not face.

Table 2: Boundedness is a property of the normalisation axis only. *A structural verification*: it checks Proposition 1 directly by supplying the modulator to the mass scan, completing the separability argument, while the 400M experiment establishes weak-gating feasibility. What it establishes is that the mass is bounded under content-dependent rates; whether the composite is also a better model is the effect-size question taken up in Section 6. Constant source, $L = 1{,}024$, 4,096 sampled channels per spectrum. 'conv.' asks whether the readout is a convex combination of the ρ_s, not merely bounded.

	`mass`		`rms`		`none`	
spectrum	ratio	M1 drift	ratio	M1 drift	ratio	M1 drift
400M NORM	1.000	4.6×10^{-7}	0.040	0.86	$940-973\times$	7.0
400M UNNORM	1.000	4.6×10^{-7}	0.043	1.23	$934-970\times$	7.9
1.3B UNNORM	1.000	4.6×10^{-7}	0.046	0.72	$936-970\times$	7.9
conv.?	yes		no		no	

Measured, on the three spectra we have. Proposition 1 is a statement about the constant source, so it is checked without training anything on the second axis: the mass scan is driven by $\mathbf{1}$, the signal scan by ρ, and the three transitions differ only in the modulator we supply directly. We ran all nine combinations of transition (`position_only`, `permeability`, `content_gate`) and normalisation (`mass`, `rms`, `none`) on the three measured spectra (4,096 channels each, $L = 1{,}024$, constant source). Table 2 reports the ratio $\max_t \|F_t\|_\infty / \max_s \|\rho_s\|_\infty$, whose value ≤ 1 is the convex-combination bound, and the M1 drift—the variation of F_t/m_t over t—whose value 0 is the prediction. The three transitions collapse into one column because they agree to the precision shown; that agreement is the result.

Three things follow. *(i)* The mass row is 1.000 with drift 4.6×10^{-7} in all nine cells—three spectra, three transitions—agreeing to within 1.1%; that value is at float32 accumulation precision, i.e. zero to the accuracy of the computation. Boundedness does not come from C3. *(ii)* The `none` row is $10^3\times$ unbounded on all three, so the bound is doing work. *(iii)* `rms` is bounded (0.04–0.05) but not a convex combination, and its M1 drift is $O(1)$ rather than $O(10^{-7})$: dividing by a norm of the state *rescales* it, whereas dividing by the mass *averages* it—the distinction of Section 2, recorded checkpoint-internally at 1.9×10^5.

What C3 buys is not the mass but its form: the closed-form partition function of step 5, the horizon law $1/\alpha_{\min}$, and—the point of Section 5.1—a rate spectrum that is a constant of the model rather than a quantity realised per input. Relaxing C3 forfeits those and not boundedness.

Theorem 2 (Dual recurrence, conditional)**.** If *A1–A3 hold* subject to *C1–C3* and *S1 selects,* then *the state evolution is* (3)–(6)*, unique up to the four choice points of Appendix 18.*

The derivation also induces a three-axis parameterization—(spectrum support, transition dependence, normalization)—on which the steps select real-positive, position-only, mass-present (Definition 2); S4D, RetNet, RWKV, GLA and Mamba occupy distinct coordinates (Table 42).

3.3 WHAT P2 IS: A RIGID TRANSLATION OF THE RATE SPECTRUM

This subsection is entirely parameter-free. It says what P2 *is*; a candidate mechanism for the *shape* of the response is a separate and weaker claim, and it is kept out of the body (Appendix 51).

The rescaling is a spectral translation, exactly. Under C3 the per-channel contraction is Λ^{Δ_t} with $\Lambda = \mathrm{diag}(e^{-\alpha_d})$, so

$$\Delta_t \mapsto s\Delta_t \iff \Lambda^{s\Delta_t} = e^{-(s\alpha_d)\Delta_t} \iff \alpha_d \mapsto s\,\alpha_d \ \forall d \iff \log\alpha_d \mapsto \log\alpha_d + \tau, \quad \tau = \ln s. \tag{7}$$

No approximation is involved: *P2 is a rigid translation of the log-rate distribution*, which is what makes it a coordinate-level statement and not a performance statement. The requirement is that the transition be *multiplicative in position*—the rates entering $G(t,s) = \exp(-\sum_k \alpha(u_k)\Delta_k)$ must not themselves depend on the lag. C3 supplies this by making α independent of the input altogether, so C3 is sufficient and not necessary. What a content-dependent rate changes is the *object* being

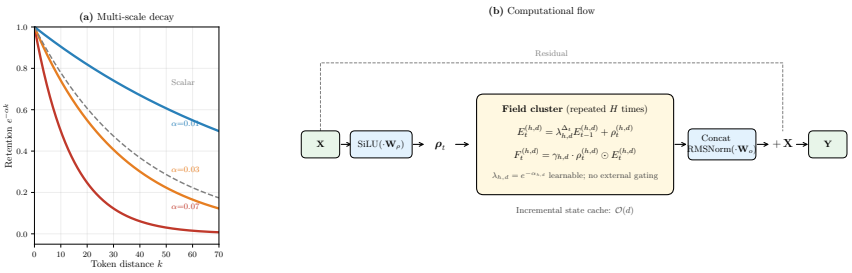

Figure 1: **The dual recurrence, and the layer that implements it.** A *coupled* dual recursion: one operator, two sides. E_t carries the input (source ρ_t); m_t carries the count (source $\mathbf{1}$); they share Λ^{Δ_t}, and that sharing is what makes $\tilde{F}_t = E_t \oslash (m_t + \epsilon)$ a posterior mean rather than a rescaled accumulation. Under content gating the operator inherits the input dependence: the pairing survives, the closed form does not. (b) Per-channel scalar state $E_t^{(h,d)}$, learnable $\lambda_{h,d}$, chunked carry.

translated: under C3 the spectrum $\{\alpha_d\}$ is a constant of the model, whereas under $\alpha = \alpha(u_t)$ the rates are realised per input, so the translated spectrum is a random variable rather than a fitted parameter.

Two spectral functionals. Two functionals of the rate spectrum, $N_{\text{eff}}(\cdot)$ and $T_K(\cdot)$, make the translation of Section 5.1 quantitative; both are developed in Appendix E. Here we use only the exclusion it establishes—that no mechanism depending on $|\tau|$ alone can account for the observed asymmetry.

4 FieldNet: the instance, and how it is measured

The layer instantiates the dual recursion the derivation forces: E_t accumulates the signal, m_t the mass that normalizes it, and both are evaluations of one operator—$\mathcal{T}_\Lambda \rho$ against $\mathcal{T}_\Lambda \mathbf{1}$. The pairing does not require C3: under a content-dependent transition the operator inherits the input dependence and the mass is still its value at $\mathbf{1}$. What C3 buys is that the shared operator is input-independent, hence closed-form and readable before training.

Proposition 3 (Linear complexity)**.** *For fixed d_{model}, H, B the layer is $\mathcal{O}(L)$ in time and $\mathcal{O}(1)$ per step in space.*

5 Testing the predictions

What the validation reports. The derivation predicts the form; the predictions are tested on the released implementation. P1 holds across two languages and three scales; P2 transfers its shape across four checkpoints, with the argmin at $s = 1$ and length invariance as a bound; M1–M3 certify that the shipped implementation has the form the derivation fixes. Each subsection states its falsifier and its scope at the end.

How the arms separate under the trend test. The monotone trend test is a single test over all 16 regions, so it needs no multiplicity correction: the 405M baseline degrades by $+116.2\%$ per octave ($t = 32.6$, $p < 10^{-4}$), while the dual-recursion instance of the same width changes by -0.19% per octave—negative, hence not degrading, and not resolvable from zero (Appendix 52). That is the sense in which P1 is a boundedness claim rather than an invariance claim.

P2's scope. *Only P2 separates the two readings of "content dependence."* Against a lag-dependent kernel it has no value; against a content-dependent but lag-independent rate it does—and one of the four runs of Table 11 takes that road, so the table is itself the demonstration that P2's scope is wider than C3. P2 therefore certifies multiplicative position dependence; that the translated spectrum is a constant of the model rather than a per-input quantity is a further statement, and C3 is what supplies it.

What P2 tests, and what it does not. P2 does not compare the C3 road to content gating; it tests whether the transition is multiplicative in position in a trained model. That it has no value against a lag-dependent kernel is a scope statement, not a superiority claim, and that it does have a value against a content-dependent rate is the reason the claim above is stated as it is. Whether retaining C3 is worth it is not established by P2 but by the controlled comparison of Table 1.

5.1 P2: POSITIONAL RESCALING (MULTIPLICATIVE POSITION DEPENDENCE)

Prediction. Under C3, rescaling position by $\Delta_t \mapsto s\Delta_t$ gives a length-invariant asymmetric U-curve with its minimum at $s = 1$. **Falsifier.** The shape varies with L, or the minimum moves off $s = 1$. **Coordinate moved.** Elapsed time only: weights, tokenizer, data and schedule are untouched and no checkpoint is retrained. **If it fails.** The transition is not multiplicative in position: if it is genuinely lag-dependent—a content-similarity kernel—there is no rate spectrum to translate; if it is content-dependent but multiplicative, P2 survives and what fails is C3 together with the closed-form partition function of step 5, while the companion mass and boundedness survive (Proposition 1).

Why the minimum at $s = 1$ is not the content. The rates are fitted parameters, so the loss is stationary in the radial direction $\alpha \mapsto s\alpha$ at the trained point and $a'(1)$ vanishes; the asymmetry is therefore a second-order effect, not a level effect. The argument, the fine sweep that tests it and the resolution it reaches are in Appendix F.

The shape across four checkpoints. The sweep above is one of the checkpoints in Table 37; three more complete a 2×2 grid over scale and instance with L_{train} matched within each scale—the 200M NORM and the 400M UNNORM instance, alongside the 400M NORM sweep of Appendix 27. *The shape transfers*: the argmin is $s = 1$ at all six lengths on all four, the curve is monotone on both sides of it, and the asymmetry is in the stretching direction throughout, $A(4)/A(1/4)$ running 1.96, 4.46, 9.42 and 8.10 against ≈ 1 for a symmetric V. *The magnitude does not*: it spans $4.8\times$ and separates by instance at matched L_{train}—a direction, not an attribution, the two instances differing in version as well (Appendix 31). *Length invariance holds as a bound on all four*: from $L = 4{,}096$ to $32{,}768$ the largest change in any $A(s)$ entry is 0.00%, 0.68%, 0.24% and 0.42% in the order of the table, against a shape spanning $4\times$ to $74\times$.

Statistics, reported rather than withheld. Every arm comparison in this paper is scored on the same $R = 16$ anchored regions, and because the regions are shared the comparison is paired. We therefore report the paired difference in log-perplexity, its region-level standard error, a two-sided paired t-test, a Holm correction across the lengths tested, and the minimum detectable effect at that R (Appendix 52, identical for every arm pair). Two comparisons are thereby resolved where we previously reported none: the matched 1.3B baseline is ahead by $\Delta \ln \mathrm{PPL} = +0.473$ ($\mathrm{SE} = 0.059$, $t = 8.07$, $p = 7.7 \times 10^{-7}$, Holm-adjusted 1.5×10^{-6} at $L = 2{,}048$), and at the scale where both instances exist the companion mass is worth $\Delta \ln \mathrm{PPL} = -0.163$ in favour of the dual instance ($\mathrm{SE} = 0.040$, $t = -4.11$, $p = 9.2 \times 10^{-4}$, Holm-adjusted 1.6×10^{-3}).

The same table bounds every single-axis claim: at a noise floor of $\mathrm{MDE} = 2.4\%$ at 200M the rest are unresolved rather than absent.

The divisor is not a scalar, and M3 moves only that axis. The settings differ in shape as well as in whether the divisor recurs. The mass is per channel: m_t has the shape of E_t, and each channel is divided by *its own* count, so the readout is a convex combination with the right weights (Lemma 5). A positionwise divisor is one scalar per cluster and position, shared across channels and computed from E_t itself, so it depends on the sources: exactly what a posterior mean forbids.

What P1 does not establish. P1 tests the *absence of an additive encoding*, which any architecture carrying no such encoding satisfies; it does not establish that only C3 delivers it, because no matched pair isolates the transition. The qualification travels to the unadapted-baseline columns of Table 14 and not to P2, which is the coordinate-level statement; Appendix 44 fixes the patched-baseline scope.

Table 3: **Four settings at 200M in one lineage** (v9.2, $L_{\text{train}} = 1024$). $\rho = \text{PPL}(32\text{K})/\text{PPL}(L_{\text{train}})$; accuracies are acc, seven tasks. The `rms` length sweep was stopped at the first length, where perplexity was already astronomical, so no ρ is reported for it. Two things are visible. Within the training length `none` is only 10% behind `mass`—acceptable, which is why the omission goes unnoticed when models are scored at the length they were trained at—while past it `none` diverges by four orders of magnitude. And `rms` is *bounded* (Table 2, ratio 0.04) yet worse than `none` at every length: dividing by the norm of the state is a rescaling, not an average over the source sequence.

Setting	Divisor	wPPL ↓	Acc ↑	PG19 @1024	PG19 @32K
`mass`	companion rec. m_t	51.23	44.16	33.47	33.36
`none`	none	64.41	42.48	36.96	1.14×10^6
`rms`	$\|E_t\|$, positionwise	1.37×10^6	43.11	astr.	—
`S-CPLX`	m_t, complex spectrum	67.02	44.13	38.42	38.41

6 Discussion

The trade, and the direction this paper opens. Keeping C3 trades selectivity for the semigroup, which gives P1, P2 and a horizon that is a functional of the weight distribution rather than the design number $\kappa/\alpha_{\min}$. Proposition 1 shows the companion mass is available under *any* causal transition, strong content gating included, and Table 2 verifies boundedness in all nine cells. What remains at the strong end is an *effect-size* question, not a structural one: we do not train it here because the budget is spent on the structural claim and its predictions, and the implementation supports it as the next step. What the projection rests on is the one scale where both instances exist: at 400M the dual instance is ahead of the single one by 8.32 PG19 PPL, 12.33 WikiText-103 wPPL and $+0.44$ mean downstream accuracy (Tables 12, 1). The prediction is specific about the sign of a dual/single difference at 1.3B and not about its size. This bounds the empirical claim and not the structural one: nothing in Section 3 mentions scale.

What kind of novelty this is. The tools are old and put to a tension they had not been put to. Bernstein–Widder converts the heavy-tail requirement into a mixture of exponentials, which is what a finite-state recurrence can carry, so power-law dependence and constant-size memory stop conflicting; FlashAttention (Dao et al., 2022) is the same move in another medium, a decades-old I/O technique applied to a bottleneck mispriced as arithmetic. In both, the novelty is the matching and not the tool. The theorem is not ours, and we do not claim its hypothesis fails for every content-dependent family—only that it holds here and that departing from it is priced (Lemma 11).

7 Conclusion

The field gave up time homogeneity for selectivity and read the loss of the semigroup as the loss of normalization. That reading conflates two things. Bounded normalization comes from A3 alone and survives any transition; what time homogeneity adds is the closed form, and with it the predictions and a horizon readable before training. Read the state out by its posterior mean and a normalizer is demanded; keep the transition time-homogeneous and that normalizer is *determined* to be the signal's dual under one operator, $\mathcal{T}_\Lambda \mathbf{1}$ against $\mathcal{T}_\Lambda \rho$—a coupled dual, not a second recurrence with a decay of its own. It is a consequence of the conjunction—A3 deciding that a companion mass exists, C3 deciding that it is closed-form, shared and testable—not a design choice, and FieldNet is the instance that ships it. Two predictions follow, and the rescaling $\Delta_t \mapsto s\Delta_t$—which a lag-dependent kernel cannot formulate—holds on four independent runs at two scales and on both instances, the shape transferring and the magnitude not. The trilemma, as commonly stated, holds only under an unnormalized readout. Under A3, selectivity, bounded normalization and linear complexity compose. This paper has measured the weak-gating case and proved the structural case; the strong-gating case is the direction it opens. What the field priced as "give up normalization for selectivity" was priced at the wrong item.

AI USE STATEMENT

Generative AI tools were used for text editing, LaTeX formatting, and code documentation. All theoretical derivations, theorem statements, proof strategies, and experimental designs were conceived and verified by the authors.

ETHICS STATEMENT

This work does not involve human subjects or private data releases. All experiments use publicly available datasets.

REPRODUCIBILITY STATEMENT

Code, configurations and evaluation scripts will be released upon acceptance; protocol, region design, test geometry and configuration tables are in Appendices 21, 33 and 35.

REFERENCES

Serge Bernstein. Sur les fonctions absolument monotones. *Acta Mathematica*, 52(1):1–66, 1929. doi: 10.1007/BF02592679.

Gregory Beylkin and Lucas Monzón. On approximation of functions by exponential sums. *Applied and Computational Harmonic Analysis*, 19(1):17–48, 2005. doi: 10.1016/j.acha.2005.01.003.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

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, 2018. URL https://arxiv.org/abs/1803.05457.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024. URL https://arxiv.org/abs/2405.21060.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

Torsten A. Enßlin, Mona Frommert, and Francisco S. Kitaura. Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis. *Physical Review D*, 80(10):105005, 2009. doi: 10.1103/PhysRevD.80.105005.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Walter Gautschi. *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2004.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM 2024), Philadelphia, PA, USA, October 7–9, 2024*, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1474–1487, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/102f0bb6efb3a6128a3c750dd16729be-Abstract.html.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022b.

Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. How to train your HiPPO: State space models with generalized orthogonal basis projections. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022c. URL https://openreview.net/forum?id=klK17OQ3KB.

Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify Mamba in vision: A linear attention perspective. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://arxiv.org/abs/2405.16605.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

David W. Kammler. Least squares approximation of completely monotonic functions by sums of exponentials. *SIAM Journal on Numerical Analysis*, 16(5):801–818, 1979. doi: 10.1137/0716060.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165, 2020.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

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 37 (NeurIPS 2024)*. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://mlanthology.org/neurips/2024/penedo2024neurips-fineweb/.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemysław Kazienko, Jan Kocoń, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Sokrates Wind, Stanisław Woźniak, Zhenyuan Zhang, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936/.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR 2022)*. OpenReview.net, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, pp. 90–95, 2011.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6395. URL https://ojs.aaai.org/index.php/AAAI/article/view/6395. Also arXiv:1907.10641; the AAAI 2020 record is used here.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, May 1–5, 2023*. OpenReview.net, 2023. doi: 10.48550/arXiv.2208.04933. URL https://arxiv.org/abs/2208.04933.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive Network: A successor to transformer for large language models, 2023. URL https://arxiv.org/abs/2307.08621.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

David Vernon Widder. *The Laplace Transform*. Number 6 in Princeton Mathematical Series. Princeton University Press, Princeton, NJ, 1941.

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

Yijiong Yu, Ziyun Dai, Zekun Wang, Wei Wang, Ran Chen, and Ji Pei. OpenCSG Chinese Corpus: A series of high-quality Chinese datasets for LLM training. *arXiv preprint arXiv:2501.08197*, 2025.

Table 4: **The design space, step by step: what each step closes, what it costs, what it buys.** A cost is a forfeited freedom, never an error. Step 2 is the fork: the field takes content gating and forfeits C3, this paper keeps C3 and forfeits *strong* content gating—the 400M dual arm of Table 1 sits between, carrying the mass on a weakly content-dependent transition. Step 5 is the conjunction defended here, and the only step at which a *second* recursion is fixed.

	Introduced	Entails	Costs	Buys
1	C1, C2, A2	first-order recursion, fixed state	attention	$\mathcal{O}(L)$ time, $\mathcal{O}(1)$ state
2	**C3**	$G(t,s) = \Lambda^{t-s}$	content gating	multiplicative position gate
3	A1 + B–W	$\Lambda = \mathrm{diag}(e^{-\alpha})$	complex spectrum (S4D)	horizon fixed at initialization
4	S1	per-*dimension* rates	parameters	$N_{\text{eff}} = D$; a scale crossover
5	A3 + C3	companion mass m_t: closed form, shared decay	the law $1/\alpha_{\min}$	algebraic boundedness

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

A MATERIAL MOVED OUT OF THE BODY FOR LENGTH

Where exactly the mainstream inference breaks. Step (i) is Mamba's own claim—selectivity is the ability to forget a token faster than its own rate allows, and it requires the decay to be a function of the input (Gu & Dao, 2024); RWKV (Peng et al., 2023) and GLA (Yang et al., 2024) reach the same place by different routes. Step (ii) is immediate and not in dispute: if $\Lambda = \Lambda(u_t)$ then $G(t,s) = \prod_j \Lambda(u_{t-j})$ is a function of the inputs between s and t, not of $t - s$. Step (iii) is where we differ, and it is worth noting what the literature does and does not say: no paper we know asserts "a lag-dependent kernel has no closed-form mass" as a claim. It is the *absence* of a normalizer in the constructions that follow from (i) and (ii), and that absence is a property of those constructions—which all read the state out directly—rather than a consequence of (ii). That is the gap we close: the missing object exists, it is the kernel's own mass, and it is available under any causal transition; what the lag-only case adds is the closed form. RetNet divides by an output normalization, not by the kernel mass; normalized linear attention divides by a content-dependent kernel mass, so no shared decay exists.

How far the gap reaches. The gap is empirical and not structural. The formal claim—given A3, C3 forces the normalizer to be a second, closed-form recursion sharing one Λ^{Δ_t}—is a statement about the transition and the estimator; nothing in Section 3 mentions scale, and no run at any scale could refute it. What depends on scale is whether that structure remains usable when trained, which is why the experiments are prediction tests of form rather than quality comparisons. The 1.3B run sits on the position-only road, so it tests the *coordinate* at the largest scale we trained ($\alpha_{\min} = 2.0 \times 10^{-3}$, $N_{\text{eff}}(1) = 12.8$ tokens); adding its P2 row would say the coordinate-level prediction holds at 1.3B, and would still not say the dual recursion exists there.

Theorem 4 (Intrinsic positional gating)**.** If *C1–C3 and A1 hold,* then *position enters through Δ_t in Λ^{Δ_t} as a multiplicative gate $e^{-\alpha\Delta_t}$ defined for every $\Delta_t > 0$: no additive encoding, nor inference-time interpolation of one, is needed beyond the trained range.*

(Proof: Appendix 17.2.) FIELDNET therefore carries no encoding, and Theorem 4 concerns the form rather than comparing encodings. It predicts P1 and P2. P1 is the absence of an additive encoding; P2 is the multiplicative-position condition, which C3 satisfies and content-dependent rates satisfy as well. Both are lost only against a Δ-dependent kernel, which is not a coordinate any architecture in this family occupies.

How K is fixed. K is not chosen per checkpoint, which would make any comparison circular. We report T_K at the lag each spectrum itself supplies, $K = 1/\alpha_{\min}$—the memory time of the slowest

Table 5: **Every measurement in this paper, read as a test.** One row per measurement: what it predicts, whether it separates C3 from content gating, what would falsify it, and what its scope forbids it from saying. No measurement below separates C3 from content-dependent rates—P2 is defined on both—and M1–M3 are unit tests of the shipped code. What separates C3 is the closed form and the horizon law, each priced in Table 21.

Measurement	Sep. C3?	Falsifier	Scope, and what it cannot say
Derivation	conditional	the items A1–A3, C1–C3 holding without the claimed consequence	holds *if* the items hold; no axiom is claimed necessary
P1	no	held-out PPL grows past L_{train}, or an additive encoding is needed	the absence of an additive encoding, not a property of C3; no matched pair isolates the transition, so this is a scope statement, not a comparison
P2	**no**	the shape varies with L; the asymmetry reverses; the argmin is not where (20) puts it (Section 3.3)	discriminates multiplicative position dependence from lag-dependent gating; it does *not* separate C3 from content-dependent rates, on which it is equally defined
M1, M2	no	deviation above 10^{-3}; a transient that is not a function of the lag	hold for any denominator sharing the decay; they certify code, not coordinate
M3	no	a divisor that is not source-free supports a calibrated language model	one lineage: what the divisor does *here*

channel, the only lag the spectrum defines with no free parameter. A fixed lag is reported alongside so the two conventions can be told apart; the two agree qualitatively and differ in whether the checkpoint spread in T_K is a restatement of $\alpha_{\min}$ or not, which is exactly the ambiguity a fit would have to resolve (Appendix 51).

What A3 decides, and what C3 decides. *A3 decides whether a companion mass exists*: a posterior mean needs a denominator and raw accumulation does not, so replacing A3 removes the mass entirely—a different estimator, not a different implementation of this one. *C3 decides what form that mass takes*: the denominator is a functional of the kernel (Proposition 9), and it is a *closed-form partition function sharing one decay* only if the kernel is a function of the lag. Under $\Lambda = \Lambda(u_t)$ the same estimator instead recomputes it from the current query at every step—no sharing, no closed form, and a spectrum that is realised per input rather than being a constant of the model (Proposition 10); P2 still has a value there, because the rate depends on the input and not on the lag. The conjunction is the contribution and the contested half of it is C3, which is why the question answered below is what retaining C3 costs.

Three settings of one axis, and what each can hold. The normalization axis has three settings and they are *not* ordered, which is what makes the division of labour measurable rather than merely asserted. With no divisor at all the state accumulates undivided: the scale grows with how much history arrived, so length extrapolation fails, but the magnitude a language model needs is still there. With a *positionwise* divisor—the norm of the state itself, no state carried across positions—the scale is controlled but the divisor depends on the sources, so by Proposition 9 no source-free weights exist and the bound of Lemma 5 does not apply. With the posterior-mean divisor the kernel's own mass is both source-free and bounded, so the two guarantees hold at once. All three are run at 200M inside one lineage, and within that lineage only the third supports language modelling (M3, Appendix 31); the scope matters, and Appendix 31 says why.

Why the denominator needs its own recursion. It cannot be folded into E_t: a posterior mean requires weights that do not depend on the sources. Carrying the mass separately is what makes $\tilde{F}_t$ a convex combination—Lemma 5—rather than a rescaled accumulation, and m_t is the unique source-independent denominator that does this (Proposition 9). It is measured as well as argued: reading the denominator out of the signal state makes the defect $|\sum_k w'_k - 1|$ spread by 2.5×10^{-1}

across 32 sources at the most favourable head, against an identity—spread zero—for the companion recursion (Appendix H).

Why a *dual* recursion, not two. The mass inherits Λ from step 2—that is what keeping C3 buys here. Both recursions are driven by one fixed operator, so they scan chunk-parallel and run token-wise together at $2 \cdot d_{\text{model}}$ cache scalars per layer; under content gating the pair would not share a decay.

What it buys. Dissipation alone bounds the raw state; normalization upgrades the guarantee from spectral to algebraic.

Lemma 5 (Structural boundedness via convexity)**.** *With E_t, m_t per (3)–(5), $\tilde{F}_t = \sum_{s \le t} w_s \rho_s$ with $w_s \ge 0$, $\sum_s w_s \le 1$; hence $\|\tilde{F}_t\|_\infty \le \max_{s \le t} \|\rho_s\|_\infty$ for all t and all nonnegative Λ—no spectral-radius constraint, no learned readout scale.*

And what it costs: the horizon law stops applying. A convex combination damps the slow mode: the slowest mode enters the normalized readout with weight $\tilde{w}_{\min} \in (0, 1]$, contributing $\tilde{w}_{\min} e^{-\alpha_{\min} k}$ at lag k.

Lemma 6 (The forecast bounds the measured horizon)**.** *At detection threshold $\varepsilon > 0$,*

$$L_{\text{meas}} \approx \frac{1}{\alpha_{\min}} \log \frac{\tilde{w}_{\min}}{\varepsilon} \;\le\; \frac{\kappa}{\alpha_{\min}}, \tag{8}$$

with equality iff $\tilde{w}_{\min} = 1$: iff the readout leaves the slow mode undamped. The gap measures how strongly the readout attenuates the slow tail.

That is the price of step 5, and why (13) was written as a statement about a single recursion.

A layer decomposition, with the prediction registered in advance. P2 acts on the rate spectrum, and the spectrum is not unstructured: the per-layer medians fall monotonically with depth (NORM: 0.085, 0.093, 0.098, 0.087, 0.075, 0.070, 0.076, 0.065 for layers 0–7 against 0.060, 0.059, 0.056, 0.051, 0.042, 0.042, 0.037, 0.042, 0.052, 0.063 for layers 8–17), so the early layers carry the fast channels and the deep layers the slow ones. This makes a decomposition available that P2 alone does not provide: scale the rates in one half of the layers and leave the other at $s = 1$, and ask which half carries the response. We registered the prediction before running it, and we state the registration because the result is the opposite of what we predicted:

> *Registered prediction.* Scaling the *slow* half (layers 8–17) accounts for most of $A(s)$, because the slow channels are the ones with memory to lose.

The prediction is **refuted**, and not marginally. Table 6 gives $A(s)$ for the three groups at $L = 8{,}192$; every group was verified after each scaling to have the intended per-layer median ratio, so this is not a null result from a scaling that failed to take effect. The fast half carries between $1.5\times$ and $21\times$ more of the response than the slow half, and the ratio grows with s: $3.2\times$ at $s = \frac{1}{4}$, $1.5\times$ at $s = \frac{1}{2}$, $11.9\times$ at $s = 2$, and $20.7\times$ at $s = 4$. At $s = 4$ the fast half alone gives $A = 17.17$ against 1.78 for the slow half.

Table 6: **Layer decomposition of P2** (NORM, $L = 8{,}192$, $R = 16$ anchored regions, baseline 31.83 PPL for all three groups). $A(s)$ is perplexity relative to $s = 1$. The registered prediction was that the slow half would dominate; it does not, at any s.

	$A(s)$			
group	$s = \frac{1}{4}$	$s = \frac{1}{2}$	$s = 2$	$s = 4$
all 18 layers	4.73	1.73	4.07	36.69
slow half (layers 8–17)	1.66	1.28	1.21	1.78
fast half (layers 0–7)	3.11	1.43	3.47	17.17
fast/slow, $\lvert A - 1 \rvert$	$3.2\times$	$1.5\times$	$11.9\times$	$20.7\times$

Three things follow, and the first is the one we would ask a reader to take away. *(i)* The refutation is directional, not a null: the response does not fail to localise, it localises in the half we did not

predict. That is a measurement, and we report it with the same prominence we would have given the confirmation. *(ii)* It is consistent with the two spectral results above. $M_2 = \mathbb{E}_d[\tanh(s\alpha_d/2)]$ is dominated by the fast channels—they have the largest $\tanh$ and therefore the largest share of M_2—so if the loss responds to the sharpness of the readout, the fast half is where a rescaling should bite, and that is what happens. $\mathrm{Var}_d[\ln r]$ is likewise a fast-channel object: below $s = 1$ every channel sits in the linear part of $\tanh$ and scales by the same factor, so the dispersion across channels vanishes; above $s = 1$ the fast channels saturate at different rates and dispersion appears. Both of the quantities that survive (C2) are fast-end quantities, and the layer decomposition says the response is a fast-end response. *(iii)* We also checked whether the two halves compose log-additively. They approximately do under this split—$\ln A_{\text{all}}$ against $\ln A_{\text{slow}} + \ln A_{\text{fast}}$ differs by -5.7%, -9.3%, -2.2% and $+5.0\%$—but we do not rest anything on the sign of that residual, because when the same comparison is made under the depth-matched split below it does not reproduce (-9.8%, -16.0%, $+14.2\%$, -38.3%). A quantity that changes sign with the decomposition is not a measurement of the model; it is a property of how the model was cut. We report the numbers and drop the claim.

A.1 P1: EXTRAPOLATION BEYOND THE TRAINING LENGTH (SCOPE-FIXING, NOT DISCRIMINATING)

Prediction. Position enters multiplicatively, so held-out quality is flat past L_{train} with no positional encoding of any kind. **Falsifier.** Held-out PPL grows past L_{train}, or an additive encoding is needed to hold it. **Coordinate moved.** Sequence length at evaluation time only. **If it fails.** Theorem 4 is false as stated. **What it cannot establish.** That C3 is responsible: no matched pair isolates the transition coordinate, so P1's coordinate-independence is a statement about what the prediction tests rather than a measured comparison (Section A.2).

$\rho \in [0.990, 1.006]$ across two languages and three scales—never worse by more than 0.7%—with no positional encoding of any kind in the model. Read region-paired rather than as a ratio of means: because every length scores the *same* anchored window, the $R = 16$ regions are paired across L, and the 95% interval on ρ at $L = 32{,}768$ is $\pm 0.06\%$ to $\pm 0.79\%$ depending on the checkpoint (Appendix 52). The falsifier is therefore excluded at that resolution—against degradations of 10^2–10^{16} for the unadapted baselines in the same table. Table 26 gives every length, and the 1.3B pair is matched in width and depth, not only in parameter count. What is tested is form, not level: at $L = 2048$ the 1.3B baseline is ahead (12.81 against 20.56), a 60% gap that leaves ρ, a ratio, unaffected (Appendix 25).

Two further columns are reported for contrast: unadapted baselines degrade by $175.6\times$, $1700.9\times$ and 1.04×10^{16}, the failure mode that removing an additive encoding removes—not, we add, one that C3 specifically removes (Appendix 44): the prediction is about the form, and no matched pair isolates the transition to test that. The evidence for P1 is the ρ ratio, read on FIELDNET alone.

P1 does not require a real spectrum. The complex-spectrum arm (S-CPLX) holds its level from $L = 1{,}024$ to $L = 32{,}768$ as tightly as the real-spectrum arm ($38.42 \to 38.41$ against $33.47 \to 33.36$), even though its held-out perplexity is 31% higher (Table 3). P1 is a statement about the form—position entering through a multiplicative gate—and is separable from the spectrum: what the spectrum axis buys is quality, not length invariance, and the 200M quartet separates the two because all four arms carry the same mass.

A.2 THE TRANSITION AXIS IN THE RELEASED IMPLEMENTATION

The fork at step 2 is a live alternative, and this paper stands some distance down that road—far enough that the strongest arm in the paper is not on the time-homogeneous coordinate. It is worth saying exactly how far.

The transition axis has three settings in the released code. `position_only` is C3: rates are fitted parameters, independent of the input. `permeability` multiplies each rate by $\exp(\beta \tanh(w^\top x))$ with $\beta = 0.5$, so rates depend on the input but only within the fixed band $[e^{-\beta}, e^{\beta}]$—content dependence that is real but bounded. `content_gate` multiplies them by $\exp(\mathrm{clamp}(w^\top x, \pm c))$, $c = 4$ by default, a range of $e^8 \approx 3 \times 10^3$: content dependence that is not bounded at all. The 400M dual-recursion arm—the one that leads the matched baseline on WikiText-103, on downstream and on PG19—is a *permeability* arm. It carries the companion mass. So the composition that this paper

argues for is not only constructible: at the weak end of the axis it is trained, measured, and better than both the single-recursion arm and the shared-tokenizer baseline. What this does *not* establish is attribution, for the 400M pair differs in code version and in normalization as well as in transition (Appendix 31); what it establishes is feasibility, which is the load-bearing part of the direction claim.

Two things therefore remain at the *effect-size* level, and we name them rather than let them be discovered. The probe of Table 2 is a structural statement about the constant source; it does not train the composite. *First*, strong gating (`content_gate`) at any scale, alone or with the mass: the combination is constructible with no new machinery and carries the one quantitative precondition recorded in Appendix 19, and measuring it is the effect-size extension the implementation supports. *Second*, a single-axis contrast along the transition coordinate: the arm that would isolate it is the permeability arm retrained against a matched `position_only` one in the same version, and that is not in this paper. Strong content dependence is structurally supported and its effect size is the direction this paper opens; the weak end is measured.

A.3 M1–M3: MECHANISM CHECKS ON THE SHIPPED IMPLEMENTATION

Prediction. The shipped code realises one shared decay, and the divisor that is not source-free is not the one that supports a calibrated language model. **Falsifier.** M1: deviation above 10^{-3}, or a probe too coarse to see a violation. M2: a transient that is not a function of the lag alone. M3: a non-source-free divisor supporting language modelling in this lineage. **Coordinate moved.** The decay of the mass recursion and the setting of `normalization_mode`, inside one checkpoint or one lineage. **If it fails.** The implementation does not have the form the derivation fixes—a code defect, not a refutation of the conjunction. **What they cannot establish.** That C3 holds: M1 and M2 would hold for any normalized linear attention whose denominator shares the decay, so their scope is the implementation, not the coordinate.

M1: constant input, constant readout. Feed L copies of one token. With no additive encoding (Theorem 4) both recursions accumulate the same geometric weight $c_t = (I - \Lambda^t)(I - \Lambda)^{-1}$ and

$$E_t = c_t \rho, \quad m_t = c_t \mathbf{1}, \quad \tilde{F}_t = \frac{E_t}{m_t + \epsilon} = \rho \cdot \frac{c_t}{c_t + \epsilon}. \tag{9}$$

The weight cancels *elementwise*, and it cancels iff the two recursions share Λ: a second decay $\mu \neq \lambda$ gives $\rho(1 - \lambda^t)(1 - \mu)/(1 - \lambda)(1 - \mu^t)$, which varies with t. Constancy is therefore *equivalent* to the sharing and readable without training or statistics. On `FieldNet112_404M_EN` (400M NORM) the deviation is 2.29×10^{-7} at $L \in \{64, 256, 1024, 4096, 16384\}$—the *same* number at every length, a fixed floor derived from `MASS_EPS`$= 10^{-6}$ rather than fitted (Appendix B). *Falsifier:* deviation above 10^{-3}, or a probe too coarse to see a violation.

The probe discriminates, and what the switch costs. Giving the mass its own decay $\mu = \lambda^p$ and repeating the identical measurement moves the deviation by five orders of magnitude, while the no-op $p = 1$ reproduces the baseline exactly, so the patch is clean (Table 7). Two controls separate the mass from any output normalization: the single-recursion checkpoint is *not* constant, and within one checkpoint switching `normalization_mode` from `mass` to `rms` moves it by 1.9×10^5 with weights, forward graph and α fixed. That is a constancy probe, not a quality measurement; the quality counterpart is the matched 400M pair, where the same architecture with and without the companion recursion differs by 8.32 PG19 PPL and 12.33 WikiText-103 wPPL (Table 12). All M1 numbers are fp32; in bf16 the probe returns 0, which is coarser rather than sharper.

M2: the transient depends on the lag alone. M1 is a statement about the fixed point and M2 about the attractor. Feed ρ_A for $t < N$ and ρ_B from $t = N$ on: with one shared decay,

$$F_t - \rho_B = (\rho_A - \rho_B)\, \frac{\lambda^k (1 - \lambda^N)}{1 - \lambda^{N+k}}, \quad k = t - N + 1. \tag{10}$$

The limit is ρ_B *exactly*—gain one, no per-channel bias—and the transient is a function of the lag k alone, N entering only through the amplitude $1 - \lambda^N$. A mass with its own decay would converge to $\rho_B(1 - \mu)/(1 - \lambda)$, a gain bias of order the step rather than a residual; measured, the step at $k = 0$ is 46.25 and by $k = 12{,}288$ what remains is 2.06×10^{-8} of it, against a tolerance of 10^{-2} and a bound of 6.9×10^{-3} recovered from each channel's own transient rather than fitted (Table 8, Appendix B).

Lag-shift invariance: C3 at the mechanism. Switches at $N = 4{,}096$ and $N = 8{,}192$ must give the same transient in lag, and do: both open at 46.25 and close at 9.5367×10^{-7}, and over the 1201 channels whose memory is shorter than the switch the median local-relative mismatch between the two lag profiles is 1.87×10^{-5} against a tolerance of 2×10^{-2}. This is the mechanism-level counterpart of P1 and P2: those read time homogeneity off perplexity, this off the transient the recursion produces.

M3: three settings of the divisor, one lineage. `CORE`, `RMS` and `NONE` are three arms of the v9.2 lineage at 200M with $L_{\text{train}} = 1024$, differing in `normalization_mode` and in nothing else we control (Table 3), so this is the one comparison in which the divisor is the only thing that moves. On language modelling only one works: WikiText-103 word perplexity is 51.23 for `mass`, 64.41 for `none` and 1.37×10^{6} for `rms`, at 1.062, 1.124 and 3.813 bits per byte. The ordering is not monotone, and that is the point: `none` keeps the magnitude and loses the bound—$\rho = 3.1 \times 10^{4}$ against 0.997 for `mass`—while `rms` keeps the bound and loses the magnitude. Downstream, where only the relative logits of a few options matter, the three sit within 1.7 points (44.16, 43.11 and 42.48 mean accuracy) and `rms` is the *better* of the two alternatives.

B MECHANISM CHECKS: SUPPORTING DETAIL

Table 7: **M1: power check and controls** (same probe throughout). Perturbation: only the mass recursion's own decay is rescaled.

reading	max rel. deviation	ratio to shared
shared decay	2.29×10^{-7}	1
$\mu = \lambda^{1}$ (no-op)	2.29×10^{-7}	1.00
$\mu = \lambda^{2}$	2.03×10^{-2}	8.85×10^{4}
$\mu = \lambda^{0.5}$	5.51×10^{-2}	2.40×10^{5}
single-recursion checkpoint	1.09×10^{-2}	4.7×10^{4}
same checkpoint, `rms`	4.28×10^{-2}	1.9×10^{5}

Why the M1 residual is at the size it is. Per channel, $F_t/\rho = 1/(1 + \epsilon/m_t)$: the whole t-dependence is the scalar ϵ/m_t, decreasing in t and bounded by ϵ because $m_t \geq 1$. With `MASS_EPS`$= 10^{-6}$ no channel can exceed 10^{-6}, so the observed 2.29×10^{-7} is a consequence of a constant in the code rather than of a fit. That the number is identical at all five lengths is correspondingly what one expects of a fixed numerical floor and not of something that accumulates with L.

What the $L = 16{,}384$ run carries. The scan performs 171 chunk-boundary carries under a clamp of $\Lambda - \Lambda_{\text{ref}}$ to $\pm\frac{1}{2}$`log_shift_max`. Were that clamp active—the failure the contract $\alpha_{\text{eff}}\text{cs} \leq 160$ exists to prevent—the two recursions would no longer carry the same weight and constancy would fail. It is not active: $\alpha_{\text{eff}}\text{cs} = 150.4$. M1 therefore checks the shipped code at its operating point, and not only the algebra.

The full M2 lag grid. Table 8 in the body prints a subsample. The full grid, at $L = 16{,}384$ with the switch at $N = 4{,}096$, is $k = 1, 4, 37, 159, 678, 2886, 12288$; the readings are 2.35×10^{1}, 5.35×10^{0}, 2.62×10^{1}, 1.06×10^{1}, 8.26×10^{-1}, 1.10×10^{-2} and 9.54×10^{-7} as absolute deviations, i.e. 5.09×10^{-1}, 1.16×10^{-1}, 5.67×10^{-1}, 2.30×10^{-1}, 1.79×10^{-2}, 2.38×10^{-4} and 2.06×10^{-8} of the step at $k = 0$. The profile is not monotone in the maximum over channels—the probe sits after the output RMS normalisation, which rescales each position by its own RMS—and neither conclusion uses monotonicity.

The two readings of length invariance on the dual instance. On the held-out PPL itself the largest change across the seven scales from $L = 4{,}096$ to $32{,}768$ is 0.29%, and the boundary drift at $s = \frac{1}{4}$ between L_{train} and $4L_{\text{train}}$ is $+0.11\%$. On the normalised $A(s)$ of Table 28 the same two numbers are 0.42% and 0.42%: the difference is that the $s = 1$ column the ratio is taken against itself moves by -0.31% over the same range. Both readings are inside the region-level noise band.

Table 8: **M2: the transient after a change of source** (EN, 400M NORM, `FieldNet112_404M_EN`, `model.layers.0.field_layer`; fp32; $L = 16{,}384$, switch at $N = 4{,}096$). $|F_{N-1+k} - \rho_B|$, and the same as a fraction of the step at $k = 0$; subsample of lags, full grid in Appendix B.

lag k	$\lvert F_{N-1+k} - \rho_B \rvert$	fraction of the step
1	2.35×10^{1}	5.09×10^{-1}
159	1.06×10^{1}	2.30×10^{-1}
2,886	1.10×10^{-2}	2.38×10^{-4}
12,288	9.54×10^{-7}	2.06×10^{-8}

The M2 bound is not fitted to the outcome. Recovering each channel's λ from its own transient and propagating (10) predicts 6.9×10^{-3} on the 1286 identifiable of 1344 channels—a channel is identifiable when its decay is actually observable inside the window, $\hat{\alpha} \cdot k_{\text{tail}} \geq 3$, since channels slower than the window never visibly decay and would inflate the bound by orders of magnitude. The prediction sits nearly four orders above the observation, so the result is far inside what the model's own spectrum allows.

Why the M1 and M2 probes can be trusted. For M1, the perturbation $p = 1$ is a no-op by construction, and it reproduces the baseline exactly; that is the check that the patch is clean, so the 8.85×10^4 and 2.40×10^5 ratios below it are attributable to the breaking of the sharing and not to the patch. For M2 the same role is played by the predicted bound above.

C SUPPORTING DERIVATION STEPS (1, 3 AND 4)

These three steps set up the two moves; the moves themselves are steps 2 and 5 in Section 3.

D STEP 1: WHAT $\mathcal{O}(1)$ MEMORY ALREADY FORCES

Causality (C1): E_t depends only on inputs ρ_s, $s \leq t$, because the model is autoregressive. *Linear complexity* (C2): $\mathcal{O}(1)$ compute per token and a state whose size does not depend on L; the Transformer (Vaswani et al., 2017) is what violating C2 buys. *Streaming* (A2): the estimate emitted at t is final, an emitted token being unrevisable, so later observations cannot alter it.

Axiom 2 (Prefix consistency (A2))**.** *Observations outside $[s, t]$ do not alter the conditional inference of the field on $[s, t]$ given the observations inside it.*

Together they force a first-order recursion with a fixed-size state and nothing more,

$$E_t = \Phi(E_{t-1}, \rho_t), \qquad \dim E_t = \text{const in } L, \tag{11}$$

a step every efficient sequence model pays. What it does *not* settle is the map's form: (11) fixes the order, and whether Φ is linear stays open until step 5.

D.1 STEP 3: THE SPECTRUM IS FIXED BY THE PROFILE, NOT CHOSEN

Λ is still free; what settles it is the profile being carried.

Axiom 3 (Multi-scale semantic dependence decay (A1))**.** *The profile $k \mapsto I(W_t; W_{t-k})$ is the restriction to positive integer lags of a function f completely monotone on $(0, \infty)$, with $f(k) \leq C\, k^{-\beta}$ for $C, \beta > 0$ and large k.*

Complete monotonicity is strictly stronger than the tail that motivates it; relaxing it costs everything below (Table 21). By Bernstein–Widder (Bernstein, 1929; Widder, 1941) it is equivalent to a Laplace representation with unique $\mu \geq 0$, possibly of infinite total mass—immaterial at lags $k \geq 1$:

$$f(k) = \int_0^\infty e^{-\alpha k}\, d\mu(\alpha), \qquad k^{-\beta} = \int_0^\infty e^{-\alpha k}\, \frac{\alpha^{\beta-1}}{\Gamma(\beta)}\, d\alpha. \tag{12}$$

A1 supplies the hypothesis and the mixture is the conclusion: the profile *is* a mixture of exponentials, so a finite budget *compresses* it rather than truncating a power law—a classical problem (Beylkin &

Monzón, 2005; Kammler, 1979), and one whose budget is spent on *where* the mixture is dense rather than on how far it reaches: no finite spectrum extends the reach, it only allocates it. What it forces is $\Lambda = \mathrm{diag}(e^{-\alpha})$, $\alpha_{h,d} > 0$: no complex eigenvalues, the point at which S4D departs. A1 constrains the profile while the derivation needs the hypothesis on the kernel; we identify the two up to a scalar, a modelling choice and not a theorem (Appendix 16). A readable spectrum then makes the usable range a number, set by the slowest mode $\alpha_{\min}$ and fixed at initialization:

$$L_{\mathrm{mem}} = \kappa/\alpha_{\min}, \qquad \kappa \equiv \log(1/\varepsilon), \tag{13}$$

with ε the smallest relative gain the protocol can declare detected (Lemma 6). This is the first consequence predicting a *value* rather than a shape, and it is what the horizon reading tests; two qualifications travel with it, neither deferred: κ is calibrated (Appendix 42) and (13) is the horizon of a *single* recursion.

D.2 Step 4: how many rates? A separate axis

There are D channels, each with a rate; how many distinct ones? Nothing about language or computability decides, so a criterion is needed—weaker than an axiom, since it ranks and does not exclude.

Definition 1 (Effective spectral channel count)**.** *For $A = \mathrm{diag}(\mathrm{vec}\,\Lambda) \in \mathbb{R}^{D\times D}$ from $\Lambda \in \mathbb{R}^{H\times d}$, $N_{\mathrm{eff}}(A)$ counts distinct eigenvalues with trivial Jordan blocks; it proxies temporal resolution, not performance.*

Criterion 1 (Capacity maximization (S1))**.** *Among architectures satisfying A1–A3 and C1–C3 at fixed D, select the one maximizing $N_{\mathrm{eff}}(A)$.*

Lemma 7 (Per-dimension decay)**.** *With $\Lambda \in \mathbb{R}^{H\times d}$, $D = H \cdot d$: scalar-shared decay gives $N_{\mathrm{eff}} = 1$, per-head decay with distinct rates H and per-dimension decay with distinct rates D (Appendix 17.5).*

S1 selects the third, at the cost of parameter economy; its payoff is its only falsifiable content:

Corollary 8 (Scale-dependent emergence)**.** *Let C_{task} count the decay modes a task requires at fixed D. While $C_{\mathrm{task}} \le D$, per-dimension decay beats scalar decay iff $C_{\mathrm{task}} > 1$ and per-head decay iff $C_{\mathrm{task}} > H$.* If *C_{task} grows with scale—open here—a critical scale N^* exists below which no advantage appears.*

This axis is orthogonal to the two moves—any architecture, on either road, may take one rate or D—so P1 is a control, not support.

E The two spectral functionals, and what they exclude

Two spectral functionals. $M_2 = \mathbb{E}_d[\tanh(s\alpha_d/2)]$, a sharpness measure, and $T_K(s) = \mathbb{E}_d[e^{-s\alpha_d K}]$, a tail-coverage measure, are both closed-form functionals of the measured spectrum; they are defined and read in Appendix A, and Table 9 reports them.

$$M_2(s) = \mathbb{E}_d\Big[\sum_k w_k(s)^2\Big] = \mathbb{E}_d\big[\tanh(s\alpha_d/2)\big] = 1/N_{\mathrm{eff}}(s), \tag{14}$$

$$T_K(s) = \mathbb{E}_d\Big[\sum_{k\ge K} w_k(s)\Big] = \mathbb{E}_d\big[e^{-s\alpha_d K}\big]. \tag{15}$$

M_2 is how *concentrated* the readout is: $N_{\mathrm{eff}} = 1/M_2$ is the number of tokens the readout effectively averages, and changing it changes the *sharpness* of the estimate—a change no normalization can undo, because it is not a change of scale. T_K is how much weight still sits beyond a lag K. Both come from the spectrum alone with no fitting, but they are not equally robust to how that spectrum was obtained—which is why the one consequence drawn in the body, (C2), uses M_2 only.

F The argmin, and the fine sweep that tests it

Why the minimum at $s = 1$ is not the content. The argmin is a consistency check—$s = 1$ is the trained scale—not the content. What carries falsifiable content is the *direction* of the asymmetry and the *length invariance* of the shape, both developed in Appendix A.

Table 9: **The two spectral functionals, computed from the measured per-channel rates** (Appendix 50); nothing here is fitted. $N_{\text{eff}}(s) = 1/M_2(s)$, (14); $T_K(s)$, (15), at two lags. The rates are read directly from each checkpoint—25,344 channels for the 400M instances, 46,080 for 1.3B—so no distributional assumption is involved. The two T_K panels disagree about how far apart the checkpoints are, and that disagreement is reported rather than resolved: at each spectrum's own lag the 400M instances differ by $47\times$, at a fixed lag by $11\times$ and in the *opposite* direction. How the table may be used follows from this: (C2) below rests on N_{eff} only, and every statement involving T_K is in Appendix 51, marked not fitted.

	s						
Checkpoint	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{3}{2}$	2	4
$N_{\text{eff}}(s)$, *tokens effectively averaged*							
400M UNNORM	86.1	43.1	28.8	21.7	14.6	11.1	6.0
400M NORM	55.1	27.6	18.6	14.0	9.6	7.4	4.3
1.3B UNNORM	50.1	25.2	16.9	12.8	8.8	6.9	4.1
$T_K(s)$ *at* $K = 1/\alpha_{\min}$ *(the spectrum's own lag)*							
400M UNNORM (K=496)	1.6×10^{-1}	9.1×10^{-2}	5.9×10^{-2}	4.0×10^{-2}	2.0×10^{-2}	1.0×10^{-2}	9.2×10^{-4}
400M NORM (K=6,096)	1.4×10^{-2}	4.3×10^{-3}	1.8×10^{-3}	8.6×10^{-4}	2.6×10^{-4}	9.6×10^{-5}	4.3×10^{-6}
1.3B UNNORM (K=496)	6.9×10^{-2}	3.4×10^{-2}	2.0×10^{-2}	1.3×10^{-2}	5.4×10^{-3}	2.5×10^{-3}	1.5×10^{-4}
$T_K(s)$ *at a fixed lag* $K = 2{,}048$*, for comparison*							
400M UNNORM	3.8×10^{-2}	9.5×10^{-3}	2.7×10^{-3}	8.0×10^{-4}	7.6×10^{-5}	7.7×10^{-6}	1.0×10^{-9}
400M NORM	5.6×10^{-2}	2.5×10^{-2}	1.4×10^{-2}	8.8×10^{-3}	4.2×10^{-3}	2.3×10^{-3}	3.7×10^{-4}
1.3B UNNORM	1.2×10^{-2}	2.3×10^{-3}	5.1×10^{-4}	1.2×10^{-4}	8.7×10^{-6}	6.9×10^{-7}	5.3×10^{-11}

Table 10: **The scope of (C2).** $D_F(s) = |F(s) - F(1)|$ on the three measured spectra; all six level-type scalars give $D_F(\frac{1}{4}) > D_F(4)$, the opposite ordering to the loss. The last row is the one natural scalar that goes the other way, and it is not a level—it is the *dispersion* of the per-channel log sharpness ratio, which vanishes below $s = 1$ and appears only where $\tanh$ saturates. All entries are computed from the measured per-channel rates, so no distributional assumption is involved.

F	$D_F(\frac{1}{4})$	$D_F(4)$	ordering
$\ln N_{\text{eff}}$	1.362–1.378	1.136–1.281	$\frac{1}{4} > 4$
$\ln \mathbb{E}_d[\tanh^2(s\alpha/2)]$	2.702–2.740	2.081–2.400	$\frac{1}{4} > 4$
$\ln \mathbb{E}_d[\tanh^4(s\alpha/2)]$	5.345–5.447	3.760–4.468	$\frac{1}{4} > 4$
$\mathbb{E}_d[(\ln r)^2]$	1.905–1.917	1.751–1.853	$\frac{1}{4} > 4$
$\mathbb{E}_d[\lvert \ln r \rvert]$	1.380–1.384	1.316–1.360	$\frac{1}{4} > 4$
$\text{Var}_d[\ln r]$	0.0000–0.0002	0.004–0.020	$4 > \frac{1}{4}$

G EXTRAPOLATION ON BOTH INSTANCES

Section 4 defines two instances that differ in one component: whether the companion mass recursion (5) is present. The columns of Table 14 are the UNNORM (mass-absent) instance; Table 12 puts the NORM (mass-present) instance beside it at the same matched 400M pair, same tokenizer, data, data order and schedule.

Both hold their level to 32K, so P1 is not what separates them: the companion mass buys level—8.32 PG19 PPL and 12.33 WikiText-103(Merity et al., 2017) wPPL at the same pair—and it does not buy or cost extrapolation. Downstream the ordering agrees with the level but the gap is small: 46.80 against 46.36 mean accuracy over the seven tasks of Table 38, i.e. $+0.44$ for the dual instance, with three tasks gained and four lost. We report it as a direction and not as a result—one seed, and the two arms differ in version as well as in normalization (Appendix 31).

Table 11: **P2 on four checkpoints: the shape transfers, the magnitude does not.** $A(s) = \mathrm{PPL}(s)/\mathrm{PPL}(1)$ at $L = 32{,}768$, geometric mean over $R = 16$ regions, anchored protocol, last-stride scoring. The argmin is $s = 1$ at all six lengths on all four, and the curve is monotone in s on both sides of it at every length (Appendix 29). L_{train} is matched within each scale; the SCALAR arms were not swept. The four rows are four independent training runs over three code versions (v8.2 twice, then v9.1 and v9.2), so the shape is not an artefact of one lineage. The $s = 1$ level runs 33.36, 34.55, 33.36 and 25.93 in the order of the table, so part of the spread in magnitude across rows is a level difference and not a shape difference. **Trans.** is the transition coordinate: CORE is position-only, so C3 holds and the spectrum is a constant of the model; T-PERM makes the rates content-dependent, so C3 is relaxed. P2 has a value on all four, which is the direct evidence that P2 requires multiplicative position dependence and not time homogeneity; what differs on the T-PERM row is that the translated spectrum is realised per input (Section 5.1). Rescaling position does not disturb length invariance: repeating the full length scan at each of the five scales on the 200M CORE row leaves the drift from $L = 1{,}024$ to $32{,}768$ bounded within 1.2% at every s ($+0.71\%$, -1.16%, -0.34%, -0.01% and $+0.01\%$ for $s = \frac{1}{4}, \frac{1}{2}, 1, 2, 4$), so P2 moves the level and P1 survives the move. The two v8.2 rows are the EV3 lineage, whose transition is position-only (Table 37).

Checkpoint	Ver.	$A(1/4)$	$A(1/2)$	$A(3/4)$	$A(3/2)$	$A(2)$	$A(4)$	$\frac{A(4)}{A(1/4)}$	Trans.
200M UNNORM (single)	v8.2	4.13	1.57	1.11	1.52	4.85	8.08	1.96	CORE
400M UNNORM (single)	v8.2	5.00	1.84	1.26	1.53	4.39	22.28	4.46	CORE
200M NORM (dual)	v9.2	7.84	2.03	1.18	2.14	13.68	73.83	9.42	CORE
400M NORM (dual)	v9.1	4.82	1.72	1.20	1.41	4.04	39.05	8.10	T-PERM

Table 12: **P1 on both instances**, matched 400M pair (EN; PG19, anchored protocol). $\rho = \mathrm{PPL}(32\mathrm{K})/\mathrm{PPL}(2048)$. The NORM instance is better in level and in ρ; its 1.3B counterpart is not trained, so no instance contrast is claimed at 1.3B.

Instance (400M, EN)	PPL@2048	PPL@32,768	ρ
UNNORM (mass absent)	34.33	34.55	1.006
NORM (mass present)	26.01	25.93	0.997

H THE NORMALISER CANNOT BE FOLDED INTO THE SIGNAL STATE

A posterior mean is a convex combination of the sources, which requires two properties: the weights sum to one, and *the weights do not depend on the sources*. If they did, the quotient would not be a convex combination of those sources with fixed weights—it would be a different estimator, and calling it a posterior mean would be wrong. The shipped code takes the denominator from a companion recursion whose source is $\mathbf{1}$, so $m_t = \sum_j \lambda^j$ carries no input information and $\sum_k w_k = m_t/(m_t+\varepsilon)$ is an identity in m_t. Folding the normaliser into the signal state instead gives $d_t = w^\top E_t$ with $E_t = \sum_k \lambda^k s_{t-k}$, hence $\sum_k w'_k = m_t/d_t$, which varies with s. Table 13 measures the spread of $|\sum_k w_k - 1|$ across 32 real sources (token embedding through `norm1` and`sem_proj`), with each candidate readout head calibrated on one source so that the folded arm is reported at its most favourable setting.

Table 13: **Source-dependence of the normalisation defect,** $|\sum_k w_k - 1|$. Spread across 32 sources; α from the released checkpoint, $t = 1023$, $\varepsilon = 10^{-6}$. Arm A is an identity and has zero spread by construction; arm B is reported at its best head.

Readout of the denominator	Arm A: companion m_t	Arm B: folded $d_t = w^\top E_t$
companion recursion (shipped)	0	—
`output_proj` row 0	—	4.34×10^{-1}
`output_proj` row 1	—	5.79×10^{-1}
`output_proj` row 2	—	6.20×10^{0}
`output_proj` row 3	—	1.04×10^{0}
uniform ($w = \mathbf{1}$)	—	2.53×10^{-1}

Arm A's defect, 1.7×10^{-9} to 6.1×10^{-7}, is the ε floor and is a function of m_t alone; arm B's is a function of the sources at every head tried, the mildest being 2.5×10^{-1}. This is why the step is stated as determined under the basis rather than chosen: given the posterior-mean estimator and a time-homogeneous transition, no folded denominator is available. The alternative is not a different implementation of the same estimator, it is a different estimator—and, under a content-dependent transition, a different form of the same one (Proposition 10).

I MATCHED LLAMA-2 BASELINES: A FAILURE DEMONSTRATION, NOT A BASELINE

Table 14 is reproduced here in full and is *not* used as evidence in the body. The Llama-2 columns carry additive positional encodings sized to their training context and were run without positional interpolation, so what they show is an unadapted encoding failing outside its range—not time homogeneity winning a comparison. A fair long-context baseline would require NTK- or YaRN-style adaptation, which we do not apply and do not claim to have matched.

Native extrapolation is not the same kind of object as a patched one. Adaptation is not merely undone here, and supplying NTK or YaRN arms would not put the two on one scale. Under C3 position is the kernel's argument, so extrapolation is native to the form: no positional encoding is present to interpolate, and there is nothing for a patch to repair. NTK-aware rescaling and YaRN are repairs to an additive encoding that has gone outside its table—they change how a fixed encoding is read at longer distances, not whether the transfer operator is a function of lag. Comparing the two directly would therefore confound form with level. Since the claim here is about the form—ρ not growing, and a rescaling being defined at all—the comparison that would settle it is not one that can be constructed without changing what is being measured, and the unadapted columns are reported for what they are: a demonstration of the failure mode that C3 removes, not a baseline.

Table 14: **Bounded extrapolation, with unadapted baselines.** $\rho = \mathrm{PPL}(32\mathrm{K})/\mathrm{PPL}(L_{\mathrm{train}})$. EN columns train at L_{train}=2048; the ZH pair is a separate smaller pair (∼300M, L_{train}=1024). Baselines are unmodified and unadapted—see the caveat above.

Model	EN-400M	ZH-300M	EN-1.3B
FIELDNET	1.006	0.990	0.9995
Llama-2	175.6×	1.04×10^{16}	1700.9×

J THE COMPANION MASS: TWO MEASUREMENTS REPORTED BUT NOT RESTED ON

The competing reading, and which one the data pick. The alternative is that the mass is a numerical guard—dividing by m_t keeps the state from growing, and what it buys should then concentrate where the state is largest, i.e. grow with L. Table 15 separates the two, at fixed scale, training length and protocol.

Table 15: **The dual-recursion gain is length-invariant** (EN, 400M; PG19, anchored protocol). $\Gamma(L) = 1 - \mathrm{PPL}_{\mathrm{dual}}(L)/\mathrm{PPL}_{\mathrm{single}}(L)$: the relative gain of the NORM instance over UNNORM, both trained at $L_{\mathrm{train}} = 2048$.

L	single	dual	Γ
1,024	34.61	26.38	23.78%
2,048	34.33	26.01	24.23%
4,096	34.51	25.94	24.83%
8,192	34.54	25.93	24.93%
16,384	34.54	25.93	24.93%
32,768	34.55	25.93	24.95%

Table 16: **Forecast vs. measurement for the saturation length.** $L_{\text{mem}} = \kappa/\alpha_{\text{min}}^{\text{init}}$, $\kappa = 17$ calibrated once; measured horizons come from each model's own paired six-point sweep (Appendix 23).

Config	$\alpha_{\text{min}}^{\text{init}}$	$\alpha_{\text{min}}^{\text{final}}$	Forecast	Measured
UNNORM 400M	0.002090	0.002011	8,135	2,048
UNNORM 1.3B	0.002090	0.002016	8,135	4,096
NORM 400M	0.001236	0.001163	13,754	4,096

From $L = 4{,}096$ to $32{,}768$—a factor of eight in length—Γ stays inside 24.83%–24.95%, a spread of 0.12 percentage points, while the two instances' own levels move by 0.1% and 0.03% over the same range. The gain is not merely preserved at long range; it is already 23.78% at $L = 1024$, *half* the training length, where no long-range mechanism has anything to do yet. *Falsifier:* Γ growing monotonically with L—here 0.12 percentage points over a factor of eight, so any monotone trend worth 5 points would have been seen—or appearing only past L_{train}, or changing sign. None is observed. But the pair compared here is cross-version (v8.2 EV3 against v9.x), so Γ mixes the companion mass with every other change between the two lineages; we report it as suggestive and rest nothing on it. The clean within-checkpoint evidence for the second move is M1 and M2 in Section 5.

The same move, read as a bound. The second consequence of the mass is that it damps the slow mode, so the horizon stops being a design number. Table 16 is that statement, and it is the weaker of the two: it holds, but as a bound whose constant is calibrated.

K RATE GRANULARITY: A CONTROL ON AN ORTHOGONAL AXIS

The gain appears at 400M and not at 200M (0.29 n.s. against $\mathbf{7.12}$ wPPL), bracketing $N^* \in (200, 400)$M as Corollary 8 leaves it; over a $2\times$ increase in scale SCALAR gains 7.72 and FIELDNET 14.55 (Table 18, Appendix 12). This is a control rather than support: a content-gated model with per-dimension rates could satisfy it too, so it does not discriminate the road taken, and we keep it in the paper for the opposite reason—because a prediction that does *not* depend on the fork is the one place where our evidence says nothing about the contribution.

12 SUPPORTING EVIDENCE NOT NEEDED FOR THE CLOSED LOOP

This paper's claims are tests of P1, P2 and of the two mechanism checks M1, M2—not a bid for a state of the art, so the quality and downstream measurements that a SOTA comparison would lead with are collected here. They are reported in full, including the ones that are not wins.

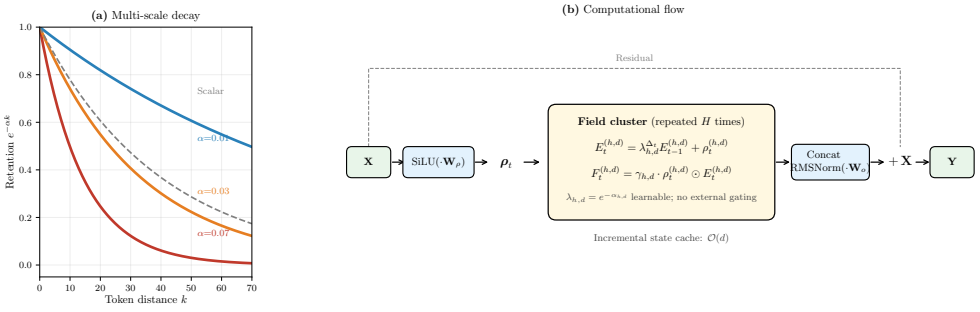

Figure 2: **FIELDNET field layer.** (a) Multi-scale decay: a single scalar rate against the multi-rate mixture; (12) is the continuum limit of the latter. (b) Computational flow: each of the $H \cdot d$ channels keeps a scalar state $E_t^{(h,d)}$; $\lambda_{h,d}$ is learnable and the gate $\lambda_{h,d}^{\Delta_t}$ depends on position alone. The cache is $\mathcal{O}(d_{\text{model}})$ per layer.

Held-out quality at the ∼400M pair. The two measures disagree and we report both. On WikiText-103 FIELDNET is ahead on both instances (52.55 and 40.22 wPPL against 57.13). On PG19, read

at the training length, the ordering splits by instance: the baseline's 28.67 is ahead of the single-recursion 34.33 and behind the dual-recursion 26.01 (Table 1). FIELDNET is ahead on PG19 at $L = 1024$ (36.14 against 34.61 and 26.38) and at every length from 4,096 on, where the baseline collapses (Table 26)—so the ordering reverses between the length the baseline was trained at, where it is at its best, and every length past it. We quote the training-length loss rather than select a length at which FIELDNET leads; the same ordering is stated at 1.3B in Section A.1 and nothing in the derivation claims absolute parity.

Zero-shot downstream at the ∼400M pair. Length-normalised accuracy for ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018) and PIQA (Bisk et al., 2020); plain accuracy elsewhere (Clark et al., 2019; Roemmele et al., 2011; Sakaguchi et al., 2020). FIELDNET leads on all seven tasks, by $+8.2$ mean accuracy; six of the seven gaps exceed the sum of the two standard errors, and WinoGrande ($+0.9$ vs. 2.8) does not. Seven, not eight: COPA is excluded from every downstream mean in this paper because its 100 test items give it a standard error of 4.5–4.9, larger than any gap we measure, so including it would make the mean a statistic of one small task rather than of the model.

Table 17: **Zero-shot downstream at the ∼400M pair** (lm-eval-harness (Gao et al., 2023); accuracy ± SE). Length-normalised accuracy for ARC, HellaSwag, OpenBookQA and PIQA, plain accuracy for BoolQ and WinoGrande; COPA is excluded (see text). Both columns are the UNNORM instance; the NORM instance at this pair is in Appendix 31.

Task	FIELDNET 404M	Llama-2 405M	Δ
ARC-Easy	50.7 ± 1.0	33.6 ± 1.0	$+17.1$
BoolQ	61.9 ± 0.9	49.9 ± 0.9	$+12.0$
PIQA	64.5 ± 1.1	54.4 ± 1.2	$+10.1$
OpenBookQA	33.4 ± 2.1	23.6 ± 1.9	$+9.8$
HellaSwag	35.1 ± 0.5	31.2 ± 0.5	$+3.9$
ARC-Challenge	27.1 ± 1.3	23.6 ± 1.2	$+3.5$
WinoGrande	51.8 ± 1.4	50.9 ± 1.4	$+0.9$
Mean	**46.36**	**38.17**	$+8.2$

This is attribution: the baseline shares tokenizer, data, order and schedule, and the level is not what is claimed. The corresponding 1.3B comparison is complete on both columns (Table 25); what is not trained at that scale is the *dual* recursion, which is the effect-size extension of Appendix 49.

Coordinates. The derivation induces a parameterization of the shared decay-gated recurrent form by (spectrum support, transition dependence, normalization), on which steps 2, 3 and 5 select position-only, real-positive and mass-present (Table 42):

Definition 2 (Three-axis parameterization)**.** *(Spectrum support, transition dependence, normalization); the steps select real-positive, position-only, mass-present (Appendix 20). The comparison is not one we have to argue for in the abstract:* S-CPLX *occupies (complex, position-only, mass) and S4D occupies (complex, position-only, no mass), so the two differ on the normalisation axis alone while sharing the spectrum—the pair that isolates what the companion mass buys at a fixed spectrum.*

Remark 1 (What a coordinate specifies)**.** *A coordinate records core recurrent choices only—not initialization, gating, readout or recipe—so S4D, RetNet, RWKV, GLA and Mamba occupy distinct coordinates.*

Rate granularity (control). The control prediction of Section K:

The two scales are therefore not a clean $2\times$ sweep: the 400M models also see twice the training context. We report the comparison as it was run rather than as a designed scaling experiment, and we note the direction in which the confound could act—a longer training context favours whichever parameterization exploits extra lags better, which is FIELDNET by Corollary 8, so the confound and the prediction point the same way and cannot be separated here. Resolving it needs the 200M pair retrained at $L_{\text{train}} = 2048$; we list it as an open item rather than claim the 7.12 is attributable to granularity alone.

Table 18: **Control: the rate-granularity crossover.** WikiText-103 wPPL, identical anchored evaluation protocol at both scales and identical token budget. The training lengths differ—$L_{\text{train}} = 1024$ at 200M against 2048 at 400M—which is a confound we name rather than remove; Appendix 31 says how far it reaches. Δ vs. SCALAR (`use_vector_decay=False`).

Scale	SCALAR	FIELDNET	Δ
200M	67.39	67.10	0.29 (n.s.)
400M	59.67	52.55	**7.12**

Attribution for A3. At fixed scale and training length, adding the companion mass moves WikiText-103 wPPL by 15.87 at 200M—67.10 for the single-recursion baseline against 51.23 for the dual-recursion CORE arm, a pair differing in the mass and in the scan-chunk setting, the latter measured to be negligible—and by 12.33 at 400M (52.55 against 40.22). The mass is not what makes the model trainable: dissipation alone suffices for optimization, and what the mass buys is Lemma 5, the algebraic bound, together with most of the level. The 400M pair also differs in how the transition is parameterized, so only the 200M number is a single-axis estimate; both are reported in Appendix 31.

13 METHODOLOGICAL COMMITMENTS

We do not claim independent derivation of existing models, only that the coordinates of Table 42 are reached by relaxing one item at a time (Appendix 19). Commitments, in force throughout:

1. *Sufficiency, not necessity.* Theorem 2 is "if", never "only if"; no axiom is claimed necessary.
2. *Falsifier with every consequence.* Consequences P1, P2 and the two checks M1, M2 pair a claim with the observation that would refute it (Table 29); a consequence without a falsifier is not claimed.
3. *Deviations reported where they arise.* NORM under the horizon reading, SCALAR at 200M under the granularity control, and the calibrated κ are stated at first mention, not deferred.
4. *Asymmetries disclosed.* The 1.3B arm is tested on fewer coordinates (Appendix 33); P2's minimum at $s = 1$ is a consistency check; the horizon law is checked at one $\alpha_{\min}$; and the NORM failure is reported as a failure.
5. *Token-matched comparisons only.* Every comparison holds tokenizer, data, data order, batch composition and schedule fixed; the baseline differs in one coordinate.
6. *Attribution over ranking.* The claim is which coordinate causes which behaviour.

14 LOGICAL HYGIENE

Scope and quantifiers. The characterization is conditional: *if* A1–A3 hold *subject to* C1–C3 *and* S1 selects, *then* (3)–(6) follows. The antecedent is stated in full so that an auditor can check it; architectures outside it occupy the coordinates of Table 42.

Independence from run order. Each consequence carries a falsifier the derivation supplies (Table 29), and the derivation is what an auditor checks—not when a run happened. What makes the horizon reading a test rather than a fit is structural: the predictor is the initialization value of $\alpha_{\min}$ while the predicted quantity is measured after training, so the two are not fitted to one another.

Reproducibility of the reported runs. Every number in this paper comes from a run listed in the manifest of Appendix 46, with its configuration, device count and throughput.

15 TWO PROPOSITIONS BEHIND THE CLAIM

This appendix states in one place the two claims the body appeals to in a line: that the companion mass is the *unique* source-independent denominator, and that the three-way coordinate difference

turns on two properties. Both are statements about form, and each is paired with a measurement where a measurement exists.

15.1 THE COMPANION MASS IS THE UNIQUE SOURCE-INDEPENDENT DENOMINATOR

Write the numerator of the readout as the signal recursion (3), which on one channel is $N_t = \sum_{s<t} \lambda^{t-s}\rho_s$, and let a candidate readout be $\tilde{F}_t = N_t/D_t$ for a denominator D_t that the architecture computes. Call $\tilde{F}_t$ a *posterior mean* if there are weights $w_s(t) \geq 0$ with $\sum_{s<t} w_s(t) = 1$, not depending on the sources ρ, such that $\tilde{F}_t = \sum_{s<t} w_s(t)\rho_s$ for every source sequence.

Proposition 9 (Uniqueness of the companion mass)**.** *Assume A1, so $\lambda \in (0, 1)$. Then the readout $\tilde{F}_t = N_t/D_t$ is a posterior mean with source-free weights if and only if $D_t = m_t = \sum_{s<t} \lambda^{t-s}$, the companion mass driven by the constant source $\mathbf{1}$; this denominator is unique up to a positive t-independent scalar, which the normalization $\sum_s w_s(t) = 1$ fixes to unity. If instead D_t is a linear functional of the signal state, $D_t = w^\top E_t$ with w learned, then $\tilde{F}_t$ is a quotient of two linear functionals of one state and admits no source-free weights unless it is degenerate.*

Proof. If D_t does not depend on the sources, then $\tilde{F}_t = \sum_{s<t}(\lambda^{t-s}/D_t)\rho_s$, so the weights are $w_s(t) = \lambda^{t-s}/D_t$. These are nonnegative under A1, are source-free by construction, and $\sum_s w_s(t) = m_t/D_t$ equals one if and only if $D_t = m_t$; a scalar multiple $D_t = c\, m_t$ gives $\sum_s w_s(t) = 1/c$, so $c = 1$ under the normalization. Conversely $D_t = m_t$ gives weights λ^{t-s}/m_t that are nonnegative, source-free and sum to one. For the last clause, $D_t = w^\top E_t = \sum_{s<t}(w^\top \lambda^{t-s})\rho_s$ is itself linear in ρ, so N_t/D_t is a quotient of two linear functionals; such a quotient is a linear functional of ρ only when the two are proportional, in which case the readout is constant in ρ and no nontrivial posterior mean is obtained. □

The proposition is a uniqueness statement *given the numerator*: once the signal recursion and the posterior-mean requirement are fixed, the denominator is determined, so the second recursion is derived rather than chosen. It does not say that a readout which is not a posterior mean is unusable—only that it is a different estimator, and that the convexity, the boundedness and the horizon semantics read off Lemma 5 do not transfer to it. The measured counterpart is Appendix H: reading the denominator out of the signal state breaks source-independence, with a cross-source spread of the normalisation defect of 2.5×10^{-1} at the most favourable readout head against an identity—spread zero—for m_t.

15.2 THE COORDINATE DIFFERENCE, AS A PROPOSITION

The distinction of Section 2 turns on two properties of a readout $\tilde{F}_t = N_t/D_t$ with $N_t = \sum_{s<t} G(t, s)\rho_s$:

- (K1) *Partition-function denominator*: $D_t = \sum_{s<t} G(t, s)$, the mass of the kernel itself, driven by the constant source $\mathbf{1}$.
- (K2) *Lag-only transition*: $G(t, s) = \Lambda^{t-s}$, a function of $t - s$ alone.

Proposition 10 (Which coordinates satisfy which property)**.** *RetNet satisfies (K2) and not (K1). Normalized linear attention satisfies (K1) and not (K2).* FIELDNET *satisfies both. The conjunction (K1)∧(K2) holds only at the* FIELDNET *coordinate, and is equivalent to $D_t = \sum_{s<t} \Lambda^{t-s}\mathbf{1}$, which admits the closed form $m_t = (1 - \Lambda^t)/(1 - \Lambda)$ per channel.*

Proof. RetNet applies a channel-wise normalization at the output of the retention layer, so its denominator is not $\sum_s G(t, s)$ for any kernel: (K1) fails, while its decay is a fixed per-channel scalar, so (K2) holds. In normalized linear attention the kernel weight is $G(t, s) = \phi(q_t)^\top \phi(k_s)$ and the denominator is $D_t = \sum_s G(t, s)$, so (K1) holds; but G depends on q_t and therefore not on $t - s$ alone, so (K2) fails. FIELDNET takes $G(t, s) = \Lambda^{t-s}$ by C3 and $D_t = m_t$ by Proposition 9, so both hold; substituting (K2) into (K1) gives the stated closed form. □

Two consequences separate the conjunction from either half alone. The closed form carries no input, so m_t is a property of the spectrum: the horizon reading and the length invariance of Appendix 27 are statements about a fixed quantity rather than about a per-step recomputation. And the conjunction is what Proposition 9 needs: under (K1) alone the denominator is a kernel mass whose weights are recomputed from the current query, so the readout is a posterior mean over content similarity; under (K1)∧(K2) it is a posterior mean over recency with a prior fixed by the spectrum alone.

Table 19: **The two properties, by coordinate** (Proposition 10; body, Section 2). Only the conjunction admits the closed form $m_t = (1 - \Lambda^t)/(1 - \Lambda)$.

Coordinate	(K1) kernel mass	(K2) lag-only	Closed-form m_t
RetNet	✗	✓	✗
Normalized linear attn.	✓	✗	✗
FIELDNET (ours)	✓	✓	✓

16 THE BASIS, IN THE ORDER THE DERIVATION CONSUMES IT

Why this basis. The axioms are not picked from a shelf; each one is consumed at the step where the previous step has left something open, and what it leaves open is what the next item closes. Read Table 20 top to bottom: C1–C2 and A2 fix the state and the order and force a first-order recursion, but leave the transition's *form* open; C3 closes that and leaves the spectrum open; A1 with Bernstein–Widder closes the spectrum—real, positive rates—and leaves the granularity open; S1 closes granularity and leaves the estimand's normalization open; A3 closes that. At every row the item is not chosen among equals, it is the one that answers the question the row above handed down. The selection therefore happens one level up, in what the problem demands—a constant-size state, a causal order, and a dependence that decays as a power law—and the basis is what those demands force, not a menu we sampled. The row where this paper departs from the field is the second one, C3: it is the fork at which the field takes content gating, and it is the one item here that is a choice rather than an answer to the row above. Section 1 says what it costs. Departures from any row are priced individually in Table 21.

Table 20: **The basis in the order the derivation consumes it** (body, Section 3). Step 1 consumes a constraint (C1–C2, fixing the state) and an axiom (A2, fixing the order); linearity arrives with A3 at step 5.

	Items	**Forced**	**Left open**
1	C1, C2, A2	first-order recursion, fixed-size state	the map's form
2	C3	time-homogeneous Λ^{t-s}	the spectrum of Λ
3	A1, B–W	real spectrum, $\alpha > 0$	granularity of the rates
4	S1	per-*dimension* rates	the normalization of the estimand
5	A3	the estimand: a normalized mean	four choice points

What A1 is, and what it costs. Complete monotonicity is strictly stronger than the power-law tail it motivates, and we present it as a *structural admissibility condition*: it makes a positive-measure mixture exist. Relaxing it costs B–W (Table 21, row 1; Appendix 16).

Each step, read backwards. Every cell left open in Table 20 is occupied by an existing architecture: higher order violates C2; a content-dependent transition forfeits the semigroup law behind P1 and P2 (Mamba, RWKV, GLA); a complex spectrum puts B–W out of reach (S4D); scalar rates give $N_{\text{eff}} = 1$ (SCALAR); and dropping the mass gives S4D's unnormalized coordinate. Each relaxation therefore isolates one item, which is what makes Table 21 readable as an attribution.

The profile and the kernel are not the same hypothesis. A1 constrains the *dependence profile* f; the derivation needs the hypothesis on the *kernel*, i.e. the prior covariance under A3. We identify them up to a scalar, $G(x, x - k) \propto f(k)$: a modelling choice, not a theorem. Two consequences. Relaxing complete monotonicity to mere non-increasingness loses the mixture representation, so the

spectrum is no longer derived from the profile; and A2 is a locality condition forbidding retroactive revision, weaker than full posterior invariance.

17 PROOFS

17.1 PROOF OF LEMMA 5 (CONVEXITY)

Proof. Fix a channel (h, d) and drop the indices. From (3)–(5),

$$E_t = \sum_{s \le t} \lambda^{\sum_{k=s+1}^{t} \Delta_k} \rho_s, \qquad m_t = \sum_{s \le t} \lambda^{\sum_{k=s+1}^{t} \Delta_k}. \tag{16}$$

Let $a_s = \lambda^{\sum_{k=s+1}^{t} \Delta_k} \ge 0$ and $M = m_t = \sum_s a_s$ (with ϵ absorbed into M). Then $\tilde{F}_t = E_t/(M+\epsilon) = \sum_s w_s \rho_s$ with $w_s = a_s/(M+\epsilon) \ge 0$ and $\sum_s w_s = M/(M+\epsilon) \le 1$. Hence $\|\tilde{F}_t\|_\infty \le \sum_s w_s \max_s \|\rho_s\|_\infty \le \max_{s \le t} \|\rho_s\|_\infty$. All inequalities are pointwise in (h, d), so the bound holds for every channel and every t; no spectral radius and no learned readout scale enter. □

17.2 PROOF OF THEOREM 4

Proof. By C3 the transfer operator depends only on the elapsed interval, $G(t, s) = \Lambda^{t-s}$. For non-uniform spacing,

$$G(t, s) = \prod_{k=s+1}^{t} \Lambda^{\Delta_k}. \tag{17}$$

By A1 with Bernstein–Widder, $\Lambda = \mathrm{diag}(e^{-\alpha})$ is diagonal, so all powers commute and

$$\prod_{k=s+1}^{t} \Lambda^{\Delta_k} = \Lambda^{\sum_{k=s+1}^{t} \Delta_k}, \tag{18}$$

i.e. the elapsed position enters as the argument of the kernel, not as an additive input to the source. By B–W the kernel is completely monotone in that argument, hence defined for every $\Delta_t > 0$: no interpolation of an additive encoding is needed. An additive encoding would enter as $\rho_t \leftarrow \rho_t + p_t$, placing position in the source and giving up the kernel/source separation that A2 motivates. □

17.3 PROOF OF THE POSTERIOR MEAN

Proof. Under A3 the prior is a Gaussian field with covariance G and observations carry i.i.d. Gaussian noise of variance σ^2. The posterior mean at a query point x is the solution of the linear system $(G + \sigma^2 I)\bar{E} = G_{\cdot x}$, whose solution is

$$\bar{E}(x) = \sum_s \frac{G(x, x_s)}{\sigma^2 + \sum_j G(x, x_j)} \rho_s \times \frac{\sigma^2 + \sum_j G(x, x_j)}{\sum_s G(x, x_s)}. \tag{19}$$

Writing $w_s \propto G(x, x_s)$, this is the Nadaraya–Watson mean $\sum_s w_s \rho_s / \sum_s w_s$ in the large-noise limit $\sigma^2 \gg \sum_j G(x, x_j)$; in that limit the ratio is exact. The discretized recurrence (3)–(6) evaluates this ratio exactly by construction, and m_t is the same kernel functional evaluated at $\rho \equiv \mathbf{1}$. □

17.4 PROOF OF PROPOSITION 3

Proof. Each channel keeps a scalar state; per token the layer performs $\mathcal{O}(Hd)$ work in the scan and $\mathcal{O}(d_{\mathrm{model}}^2)$ in the projections, both independent of L; the cache is the $H \cdot d$ state, i.e. $\mathcal{O}(d_{\mathrm{model}})$ per layer. Summing over L gives $\mathcal{O}(L)$ time and $\mathcal{O}(1)$ per-step space. □

17.5 PROOF OF LEMMA 7 (BLOCK-DIAGONAL AND DEFECTIVE CASES)

Proof. Let A have eigenvalues μ_i with algebraic multiplicities m_i. The number of linearly independent decay rates the state can represent is the number of Jordan blocks, which equals m_i when

Table 21: **What relaxing each item of the basis costs** (the table is not a performance ranking).

	Relax	What is lost
1	A1 (CM)	B–W: no mixture, rates not derived from the profile
2	A2 (prefix consistency)	Order: higher-order recursion, C2 fails
3	A3 (posterior mean)	Lemma 5; spectral fallback only
4	C1 (causality)	Not autoregressive: bidirectional, offline
5	C2 (linear complexity)	$\mathcal{O}(L)$: attention's $\mathcal{O}(L^2)$
6	C3 (time homogeneity)	Semigroup; closed form and horizon law forfeit; P2 survives under content-dependent rates and is lost only again
7	S1 (capacity)	Per-dimension rates; scalar remains admissible

the blocks for μ_i are trivial and is smaller otherwise. For scalar-shared decay all D channels share one μ, so with trivial blocks $N_{\text{eff}} = 1$. For per-head decay with H distinct rates and d channels per head, each rate has multiplicity d and—if the blocks are trivial—contributes d independent directions, giving $N_{\text{eff}} = H$ once the head structure is quotiented out. For per-dimension decay with D distinct rates, $N_{\text{eff}} = D$. Defective (non-trivial) blocks reduce the count, so the stated values are upper bounds attained when the blocks are trivial. □

18 Uniqueness and structural choice points

Theorem 2 is uniqueness *up to four choice points*, and each is where an existing architecture sits:

1. *Spectrum support.* Real-positive (B–W) or complex; S4D takes the latter, forfeiting the mixture reading.
2. *Rate granularity.* Scalar / per-head / per-dimension; S1 selects the third, and SCALAR occupies the first.
3. *Normalization.* Mass present (A3) or absent; S4D and Mamba take the latter, forfeiting Lemma 5.
4. *Transition dependence.* Position-only (C3) or content-dependent; Mamba, RWKV and GLA take the latter, forfeiting the semigroup behind P1, the closed form and the horizon law. P2 survives, because their rates depend on the input but not on the lag; what is forfeit is that the translated spectrum is a constant of the model rather than a per-input quantity.

None is a free parameter of the theorem: each is fixed by an explicit item, and each is where a named architecture departs.

19 Why one axis at a time

We instantiate axes singly rather than reproducing Mamba, RWKV and GLA, for three reasons. (i) *Attribution*: a single-axis change isolates the cost of one item; a full reproduction confounds spectrum, transition and normalization with tokenization, data and schedule. (ii) *Coverage*: the axes span the space, so the arms bound what a given relaxation costs even where no named model sits. (iii) *Testability*: P2 is defined wherever the transition is multiplicative in position, so it is not a test reserved for C3—but it is still a test of one coordinate, and the paper trains no *matched* pair across the fork (Section A.2). The cost of this choice is that we report no head-to-head against those models; we accept it, since the claim concerns coordinates.

20 Relation to S4D and to gated linear attention

S4D (Gu et al., 2022b;a) occupies the coordinate (complex spectrum, position-only, no mass): by Remark 1 only core recurrent choices are recorded. Complex eigenvalues put B–W out of reach, so its rates are not derived from the profile; absent mass, boundedness is spectral rather than algebraic. GLA (Yang et al., 2024) occupies (real-positive, content, denominator): content-dependence relaxes C3, so the closed form and the horizon law do not apply. P2 is defined on GLA under the multiplicative-position condition, though we do not sweep it here; what differs is that the translated spectrum is

realised per input rather than being a constant of the model. Neither is refuted—each relaxes a different item, at the cost listed in Table 21.

21 ANCHORED PROTOCOL, STATISTICS, AND POWER

Region-level pairing. Each held-out corpus is partitioned into $R = 16$ contiguous regions. Every length L predicts the same $4{,}096$-token span within each region, of which the final 512 tokens enter the statistic. Length therefore varies without varying difficulty, and the region is the pairing unit.

Statistics reported. Geometric mean over regions; region-level SE; 95% bootstrap CI over regions; Holm-corrected region-paired t-tests. Pairing at region level removes corpus-difficulty effects, so a length effect is not confounded with which span was predicted.

Power for P1. At 200M the observed difference is $\Delta\text{wPPL} = 0.29$ with region-level SE above that; the difference is not significant at our resolution. The 400M difference (7.12) is far outside the same band.

22 THE HORIZON READING: ITS FORECAST, ITS PRECONDITION, AND THE ONE FAILURE

The forecast is $L_{\text{mem}} = \kappa/\alpha_{\min}$ with $\kappa = 17$ calibrated once on a held-out configuration; the same κ is used everywhere, and no per-configuration fitting occurs.

One number, read at one time. The spectrum is tabulated once, in Table 16, and saturation is read at the same time point; drift is reported alongside (3.8% and 3.5% on UNNORM at 400M and 1.3B, 5.9% on NORM) and Appendix 38 gives the trajectory. Because drift is monotone and the reading is at one time, the forecast is a claim about a design quantity at that time—not about the trajectory.

23 THE HORIZON READING ON THE NORM INSTANCE, RUN SEPARATELY

The NORM 400M model was swept with the same protocol, so its measured horizon is read independently rather than assigned. Held-out PPL (PG19, $L_{\text{train}} = 2048$) runs 26.38, 26.01, 25.94, 25.93, 25.93, 25.93 at $L = 1024, \ldots, 32768$.

Table 22: **NORM cross-check** (EN, 400M; adjacent-length region-paired comparisons). The pattern reproduces the UNNORM sweep: the last significant gain is at 4096, and every pair above it is flat at a 0.0–0.1% threshold.

Pair	Rel. change	p	MDE(80%)	Reading
$2048 \to 1024$	$+1.13\%$	0.0001	0.6%	1024 worse (sig.)
$1024 \to 4096$	-1.34%	0.0000	0.6%	gain (sig.)
$4096 \to 8192$	-0.02%	0.3574	0.1%	no change
$8192 \to 16384$	-0.00%	0.8740	0.1%	no change
$16384 \to 32768$	$+0.01\%$	0.7421	0.0%	no change

24 CROSS-CHECK THAT DOES NOT READ THE SPECTRUM

The one weakness of the spectral forecast is that predictor and predicted come from the same checkpoint. The length sweep supplies an independent reading, because it never looks at $\alpha_{\min}$: it asks where in the *behaviour* further context stops helping. Comparing adjacent lengths of the grid isolates that point with the most power, since neighbouring lengths differ only in context.

Improvement is detected up to $L = 4096$ and nowhere after it: from 4096 onward every adjacent pair is flat at a detection threshold of 0.0%, so the design could have seen a change and did not. The spectral forecast puts saturation at $8{,}135$; the behavioural reading puts the last detectable gain at

Table 23: **Cross-check: where the sweep stops improving** (EN, 1.3B; region-paired comparisons between adjacent lengths; MDE is the minimum detectable effect at 80% power).

Pair	Rel. change	p	MDE(80%)	Reading
$2048 \to 1024$	$+0.51\%$	0.0003	0.3%	1024 worse (sig.)
$1024 \to 4096$	-0.55%	0.0001	0.3%	gain (sig.)
$4096 \to 8192$	$+0.01\%$	0.1628	0.0%	no change
$8192 \to 16384$	-0.00%	0.6457	0.0%	no change
$16384 \to 32768$	-0.01%	0.2360	0.0%	no change

4,096. On a grid whose resolution is a factor of two these are adjacent, so no measurement *contradicts* the forecast—but the forecast is high by $1.99\times$, and we report that as the result rather than calling it a hit. This reading is the one that does not depend on $\alpha_{\min}$.

The UNNORM 400M sweep, for comparison. The UNNORM 400M model was swept with the same protocol; held-out PPL runs 34.61, 34.33, 34.51, 34.54, 34.54, 34.55. Its minimum is at 2,048, not 4,096.

Table 24: **UNNORM cross-check** (EN, 400M; adjacent-length region-paired comparisons). The last significant gain is at 2,048; above it every change is a small *worsening*, and the monotone trend test finds no degradation per octave ($+0.02\%$, $p = 0.27$).

Pair	Rel. change	p	MDE(80%)	Reading
$2048 \to 1024$	$+0.80\%$	0.0006	0.5%	1024 worse (sig.)
$1024 \to 4096$	-0.28%	0.2877	0.7%	no change
$4096 \to 8192$	$+0.10\%$	0.0011	0.1%	slight worsening
$8192 \to 16384$	-0.01%	0.1897	0.0%	no change
$16384 \to 32768$	$+0.02\%$	0.0200	0.0%	no change

25 QUALITY AT 1.3B, AND WHAT IS COMPARABLE THERE

At 1.3B the pair is width- and depth-matched (Table 41), the cleanest comparison in the paper. On the quantity P1 is about, FIELDNET holds $\rho = 0.9995$ against $1700.9\times$. On the absolute level at $L = 2048$ the baseline is better: 12.81 against 20.56 held-out PPL, a 60% gap.

What the gap is, and what it is not. The two runs see the same token budget, the same data and the same data order, so the gap is not a data or compute effect; the pair is matched in width, depth and parameter count, so it is not a capacity effect; and ρ is a ratio, so P2 is unaffected—a level effect cannot move it. What the gap *is* is a difference between instances. The model scaled to 1.3B is the UNNORM (mass-absent) instance of Section 4; the NORM (mass-present) instance is trained and reported at 400M and its 1.3B counterpart is not yet trained. Section J distinguishes the two directly: The horizon formula describes the UNNORM instance and is loose on the NORM one, and that is the only measurement in this paper in which the two instances are contrasted. The level gap at 1.3B is measured on the same instance where the formula holds. Neither P1 nor the downstream comparison is an instance contrast—both are reported on a single instance (Table 37)—and we do not use them as evidence about the mass.

We therefore do not read the 60% as a failure of the form, and we do not claim absolute parity at 1.3B. Two things would settle it and neither is available: the dual-recursion instance at 1.3B, which is not yet trained, and a 1.3B run in which implementation-level numerical choices—the baseline here is the reference `transformers` implementation while FIELDNET is ours—are held fixed. We report the gap and leave it open.

Table 25: **Zero-shot downstream at 1.3B** (lm-eval-harness (Gao et al., 2023); accuracy ± SE; same metric convention as Table 17). Both columns are measured. The FIELDNET row is the *single*-recursion instance: the dual recursion is not trained at 1.3B, so this table does not compare instances at this scale, and it is reported so that the size of the absolute gap in Section A.1 is measured rather than inferred. FIELDNET trails the baseline by 3.11 mean accuracy and on WikiText-103 word-level perplexity (27.61 against 19.17), though it leads on one of the seven (BoolQ, 60.9 against 60.1) and on COPA (75.0 against 72.0), which sits outside the seven-task convention of Table 17 and is listed separately; what it does not trail on is length (Table 26), where the baseline leaves its training length and FIELDNET does not.

Task	Llama-2 1299M	FIELDNET 1.3B (single)
PIQA	70.8 ± 1.1	69.0 ± 1.1
ARC-Easy	61.9 ± 1.0	58.7 ± 1.0
BoolQ	60.1 ± 0.9	60.9 ± 0.9
HellaSwag	52.6 ± 0.5	45.2 ± 0.5
WinoGrande	54.9 ± 1.4	49.2 ± 1.4
OpenBookQA	36.0 ± 2.2	34.6 ± 2.1
ARC-Challenge	34.0 ± 1.4	30.9 ± 1.3
Mean	**52.90**	**49.79**

Table 26: **P1: the full length sweep** (body, Section A.1; held-out PPL, EN, anchored protocol, last-stride scoring, L_{train}=2048). All rows see the same regions and the same scoring span. The 1.3B pair is width- and depth-matched; the 405M row is the smaller baseline of the parameter-matched pair and is kept because it isolates the effect of scale on the collapse.

L	1024	2048	4096	8192	16384	32768
FIELDNET 1.3B (single)	20.66	20.56	20.55	20.55	20.55	20.55
FIELDNET 404M (dual)	26.38	26.01	25.94	25.93	25.93	25.93
Llama-2 1299M	13.29	12.81	582.58	8521.23	22629.25	21788.40
Llama-2 405M	36.14	28.67	210.75	649.93	2278.26	5033.09

26 P1: THE FULL LENGTH SWEEP

The two baseline rows differ only in scale and both collapse, so the collapse is not an artefact of the smaller width; it is *amplified* by scale (175.6× against 1700.9× at 32K). Against the 405M row, the dual-recursion instance of the same width goes $26.01 \to 25.93$ over the same range, a change of -0.31% that sits inside its own region-to-region noise, so at $L = 32{,}768$ the two differ by $194\times$. This is the comparison the paper's object is actually in: unlike the 1.3B row, it is the dual recursion, not the single one, that is measured here.

The ZH pair (300M, L_{train}=1024, 6,400-token vocabulary) reproduces the same separation in a second language at a different scale, training length and tokenizer:

L	1024	2048	4096	8192	16384	32768
FIELDNET ZH-300M	45.57	45.14	45.09	45.09	45.09	45.10
Llama-2 ZH-300M	52.10	756.73	3.00×10^5	4.24×10^{10}	2.72×10^{15}	5.42×10^{17}

The baseline's numerator here is 5.4×10^{17}, which is why Table 14 quotes $\rho = 1.04 \times 10^{16}$ rather than a PPL.

The scalar arm, on the same sweep. The granularity control's two arms were also swept (Table 27), which is how the second reading of Section K was obtained. Both arms carry the same positional mechanism, so the difference is rate granularity alone.

From $L = 2048$ to 32768 the per-dimension arm moves $+0.64\%$ and the scalar arm $+15.22\%$. A single shared rate has to serve every lag, so the fit degrades as the range grows; D rates do not. This is one ablation pair at one scale and we report it as a supporting observation only.

Table 27: **Length robustness of the granularity control's two arms** (EN, 400M; held-out PPL, anchored protocol, L_{train}=2048).

L	1024	2048	4096	8192	16384	32768
FIELDNET (per-dimension)	34.61	34.33	34.51	34.54	34.54	34.55
SCALAR	36.96	41.26	47.00	47.55	47.54	47.54

27 P2 ON THE DUAL-RECURSION INSTANCE

The rescaling sweep of Section 5.1, repeated on `FieldNet112_404M_EN`, the 400M NORM (dual-recursion) checkpoint, at the same seven scales and six lengths and under the same anchored protocol ($L_{\text{train}} = 2048$). Table 28 gives $A(s)$ in the same normalisation as Table 30; the underlying held-out PPL at $s = 1$ is 26.38 at $L = 1{,}024$ and 25.93 from $L = 8{,}192$ on.

Table 28: **P2 on the dual-recursion instance**: $A(s)$, the held-out PPL at scale s divided by the PPL at $s = 1$ (FIELDNET 400M NORM, EN; PG19, anchored protocol, $L_{\text{train}} = 2048$). Same axes as Table 30.

L	$A(1/4)$	$A(1/2)$	$A(3/4)$	$A(3/2)$	$A(2)$	$A(4)$	$s^\star$
1,024	4.81	1.73	1.21	1.41	4.00	38.36	s=1
2,048	4.79	1.72	1.20	1.41	4.03	38.93	s=1
4,096	4.80	1.72	1.20	1.41	4.04	39.03	s=1
8,192	4.81	1.72	1.20	1.41	4.04	39.05	s=1
16,384	4.81	1.72	1.20	1.41	4.04	39.05	s=1
32,768	4.82	1.72	1.20	1.41	4.04	39.05	s=1

Three readings transfer from Table 30 unchanged. The argmin is $s = 1$ at every length. The curve falls monotonically to it and rises monotonically away from it, so the shape is the U predicted by a multiplicative gate. The asymmetry is in the stretching direction at every length, $A(4)/A(1/4) \in [7.98, 8.14]$ and $A(2)/A(1/2) \in [2.31, 2.35]$, against ≈ 1 for a symmetric V. Length invariance is tighter than on the single-recursion instance. On the held-out PPL itself the largest change across the seven scales from $L = 4{,}096$ to $32{,}768$ is 0.29%, and the boundary drift at $s = \frac{1}{4}$ between L_{train} and $4L_{\text{train}}$ is $+0.11\%$ against $+7.2\%$ on the 200M UNNORM run. On the normalised $A(s)$ printed in the table the same two numbers are 0.42% and 0.42%: the $s = 1$ column the ratio is taken against falls by 0.31% between $L = 2{,}048$ and $8{,}192$ and by 0.04% between $4{,}096$ and $32{,}768$, so the ratio moves slightly more than either numerator alone. Both readings are inside the region-level noise band.

The two sweeps differ in scale (400M against 200M) and in training length (2048 against 1024), so the difference in magnitude—the asymmetry ratio is about four times larger here—is reported as a level difference between two runs and is not attributed to the companion mass. What P2 states is the shape, and the shape is the same on both instances.

28 PREDICTION AND MECHANISM TABLES, FULL FORM

The two tables the body summarizes.

29 P2 ON ALL FOUR CHECKPOINTS

The two sweeps of the body—the 200M UNNORM checkpoint (Table 30) and the 400M NORM checkpoint (Table 28)—are two of four. This appendix gives the two that complete the 2×2 grid over scale and instance: the 200M NORM checkpoint `FieldNet112aCORE_200M_EN` and the 400M UNNORM checkpoint `FieldNet108M404EV3`. Both were swept at the same seven scales and six lengths under the same anchored protocol; L_{train} is matched within each scale (1024 at 200M, 2048 at 400M), so the instance comparison at fixed scale is not confounded by training length. Raw held-out PPL is given first so that $A(s)$ can be recomputed.

Table 29: **Two predictions and two mechanism checks.** M1 and M2 certify the shipped implementation and would also hold for a normalized linear attention whose denominator shares the decay. The transition axis is not represented; see Section A.2. $\rho = \mathrm{PPL}(32\mathrm{K})/\mathrm{PPL}(L_{\mathrm{train}})$.

	Step	Tests	Statement, and its falsifier
P1	2	no additive encoding	Position enters multiplicatively, so PPL does not grow past L_{train} *without* an encoding. *Falsifier:* $\rho \gg 1$. *Scope:* P1 tests the form—the absence of an additive encoding—and not C3; no matched pair isolates the transition, so this is a scope statement rather than a comparison. P2 discriminates multiplicative position dependence from lag-dependent gating.
P2	2	C3	Rescaling $\Delta_t \mapsto s\Delta_t$ gives a length-invariant asymmetric U-curve with minimum at $s = 1$; undefined where the transition is not multiplicative in position. *Falsifier:* shape varies with L, or the minimum shifts.
M1	5 (with 2)	shared decay	Held at a constant source the readout is constant—to the ϵ floor of the released code, and *equivalently*, not merely consistently: a second decay $\Lambda_m \neq \Lambda$ makes it vary. *Falsifier:* deviation above 10^{-3}, or a probe too coarse to see it, excluded by breaking the sharing and requiring $> 10^2$.
M2	5 (with 2)	shared decay, C3	After the source changes the readout converges to it exactly (gain one), and the transient depends on the lag alone, not on when the change happened. *Falsifier:* a residual comparable to the step—what a mass with its own decay leaves—or two switch positions giving different lag profiles.

Table 30: **P2: length invariance and asymmetry.** $A(s) = \mathrm{PPL}(s)/\mathrm{PPL}(1)$, geometric mean over $R = 16$ regions; PG19, anchored protocol, last-stride scoring; seven scales $s \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \frac{3}{2}, 2, 4\}$ at six lengths. One run: FIELDNET at the 200M scale, $L_{\mathrm{train}} = 1024$ (Table 37). $s^\star$ is the argmin over the seven sampled scales; $L = 8{,}192$ and $16{,}384$ are omitted, being identical to $L = 32{,}768$ at the printed precision. A symmetric V would give $A(4) \approx A(1/4)$ (Appendix 40).

L	$A(1/4)$	$A(1/2)$	$A(3/4)$	$A(3/2)$	$A(2)$	$A(4)$	$s^\star$
1,024	3.85	1.57	1.11	1.52	4.85	8.07	s=1
2,048	4.08	1.57	1.11	1.52	4.85	8.08	s=1
4,096	4.13	1.57	1.11	1.52	4.85	8.08	s=1
32,768	4.13	1.57	1.11	1.52	4.85	8.08	s=1

30 INSTANTIATIONS, AXES, AND ARTIFACTS

Table 36 lists the arms and the axis each instantiates. The three-axis arms run at 200M and 400M and are reported in Appendix 39. The 400M NORM arm is a `permeability` arm, so the transition axis is represented at its weak end among trained arms; strong gating is constructible but untrained and appears in Appendix 49. What is fixed here is the axis mapping, not the outcome.

31 WHICH CHECKPOINT PRODUCES WHICH NUMBER

Every measurement in this paper is traceable to one row of Table 37. The instance contrast is carried by the normalization axis and is definitional, not empirical: NORM runs (3)–(5) (companion mass present, dual recursion), UNNORM runs (3) alone (single recursion), Both are trained. The 400M NORM arm sits on the `permeability` setting of the transition axis, so content dependence is represented—at its weak end—among the trained arms; strong gating is not (Section A.2).

What is *not* an instance contrast. P1's three columns, P2 and the downstream comparison are each reported on a single arm and are used as evidence about the position axis (P1, P2) and the granularity axis (a control), not about the mass. The mass is measured in two places: the 400M pair of Table 12, and the horizon reading, where the two instances sit in the same table. We state this because the alternative reading—that the extrapolation and downstream results are evidence for the dual recurrence—is available from the tables and is not one we claim.

Table 31: **Raw held-out PPL, 200M NORM** (`FieldNet112aCORE_200M_EN`, v9.2 CORE; EN, PG19, anchored protocol, $R = 16$, last-stride scoring, $L_{\text{train}} = 1024$).

L	$s{=}\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{3}{2}$	2	4
1,024	259.66	68.50	39.50	33.47	71.45	456.47	2462.78
2,048	260.02	67.79	39.34	33.36	71.26	456.39	2463.18
4,096	260.89	67.71	39.33	33.36	71.27	456.38	2463.54
8,192	261.33	67.70	39.34	33.36	71.26	456.42	2463.02
16,384	261.50	67.70	39.33	33.36	71.27	456.43	2463.02
32,768	261.52	67.70	39.33	33.36	71.27	456.43	2463.02

Table 32: $A(s)$**, 200M NORM**—the matrix above divided by its $s = 1$ column; body Table 11.

L	$s{=}\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{3}{2}$	2	4
1,024	7.76	2.05	1.18	1.00	2.14	13.64	73.58
2,048	7.79	2.03	1.18	1.00	2.14	13.68	73.84
4,096	7.82	2.03	1.18	1.00	2.14	13.68	73.85
8,192	7.83	2.03	1.18	1.00	2.14	13.68	73.83
16,384	7.84	2.03	1.18	1.00	2.14	13.68	73.83
32,768	7.84	2.03	1.18	1.00	2.14	13.68	73.83

The NORM downstream comparison. At the matched 400M pair the NORM instance is better on both perplexities—26.01 against 34.33 PG19 and 40.22 against 52.55 WikiText-103 wPPL—and marginally better downstream: 46.80 against 46.36, i.e. $+0.44$ over the seven tasks of Table 38. The per-task comparison is in that table: NORM gains on three tasks, its gains summing to $+5.1$, and loses on four, summing to -2.0; the largest single gap is HellaSwag $+3.1$. The two arms differ in version as well as in normalization, so we report the direction and do not attribute it; the mean gap is smaller than the standard error of the tasks that compose it, and it is one seed.

The normalization axis has three settings. The implementation exposes one axis, `normalization_mode`, with three settings, and reading the ablation correctly requires all three: *none* returns E_t undivided; *rms* divides E_t by its own root-mean-square over channels, at each position independently and carrying no state across positions; *mass* divides by the companion recursion m_t of (5). Only the third divisor is a functional of the kernel, and only it therefore shares the decay. The UNNORM instance reported throughout is *rms*: no companion mass, but not unnormalized either.

The two divisors are not two scalars. The shape argument and the three-setting table are in the body (Section 5); what remains here is the attribution.

What the earlier UNNORM arms actually normalize by. The UNNORM checkpoints of Table 12 and the cross-version pair of this appendix come from a different implementation, and the name is worth unpacking because it is easy to read as "no normalization". That implementation has no normalization axis at all: no mode is exposed and no companion mass exists, so the state is divided by its own norm unconditionally—the same per-cluster unit-norm divisor, with the same floor constant, that the current code calls `rms`. So UNNORM is not unnormalized; it is normalized by something other than the kernel mass. What makes it a different configuration rather than a third arm of the same sweep is elsewhere: it has no separate query projection, the field being multiplied by the same activation that produced it, and no input-dependent modulation of the decay. That matters for reading the numbers. With the divisor fixed to the same mathematics, the earlier architecture reports 67.10 wPPL and holds $\rho = 1.006$ at 400M, while the current architecture in that setting gives 1.37×10^6; a gap of four orders of magnitude on one and the same divisor cannot be laid at the divisor. We therefore use UNNORM nowhere as evidence about the mass, and we do not read the 15.87 wPPL cross-version difference as `rms` against `mass`—it is version and architecture, with normalization only one of the things that moved. What the UNNORM arms do establish is the shape of P2 on an instance without the companion recursion, which is how Table 11 uses them.

Table 33: **Raw held-out PPL, 400M UNNORM** (`FieldNet108M404EV3`, v8.2 EV3; EN, PG19, anchored protocol, $R = 16$, last-stride scoring, $L_{\text{train}} = 2048$). The $s = 1$ column reproduces the single-recursion row of Table 15.

L	$s{=}\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{3}{2}$	2	4
1,024	163.51	64.90	43.73	34.61	53.32	152.35	769.51
2,048	168.00	64.46	43.45	34.33	52.92	151.76	769.64
4,096	171.62	64.00	43.35	34.51	52.89	151.71	769.75
8,192	172.76	63.67	43.37	34.54	52.89	151.70	769.61
16,384	172.75	63.66	43.38	34.54	52.89	151.70	769.72
32,768	172.74	63.64	43.37	34.55	52.89	151.70	769.72

Table 34: $A(s)$**, 400M UNNORM**—the matrix above divided by its $s = 1$ column; body Table 11.

L	$s{=}\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{3}{2}$	2	4
1,024	4.72	1.88	1.26	1.00	1.54	4.40	22.23
2,048	4.89	1.88	1.27	1.00	1.54	4.42	22.42
4,096	4.97	1.86	1.26	1.00	1.53	4.40	22.31
8,192	5.00	1.84	1.26	1.00	1.53	4.39	22.28
16,384	5.00	1.84	1.26	1.00	1.53	4.39	22.29
32,768	5.00	1.84	1.26	1.00	1.53	4.39	22.28

A frozen-weight divisor swap is not an estimate. Since the three settings share one parameter set—no parameter is created conditionally on `normalization_mode`—the cheapest possible comparison is to load CORE's weights and switch the divisor. We ran it: at $L = 1{,}024$ the held-out PPL is 1.11×10^5 (geometric mean over $R = 16$ regions; the per-region range is 8.3×10^4 to 1.5×10^5, so no single region drives it), against 33.47 for the same weights with the companion mass; the arm *trained* in the `rms` setting reaches the same qualitative conclusion by a different route (Table 3), which is why we use neither number as an estimate. This is *not* evidence that positionwise normalization is worse: those weights were trained to be read out by a posterior mean, and a source-dependent divisor is a different estimand, not a degraded one. It is the same fact as M1's control—switching `mass` to `rms` with weights, forward graph and α fixed moves the constancy probe by 1.9×10^5—seen through perplexity instead of a unit test. Its use is negative: it closes the cheap route to a single-axis estimate of the mass, and we report it so that the absence of that estimate is not read as an untried experiment.

How far the mass can be credited. With all three settings run at 200M in one lineage (Table 3), the question "how much is the companion mass worth" turns out to be the wrong one: the settings are not ordered, so the mass has no single price. Moving from `mass` to `none` costs 13.18 wPPL and destroys length extrapolation; moving to `rms` costs the calibration outright, 1.37×10^6 wPPL against 51.23, while leaving downstream accuracy within 1.05 points of `mass`. What is attributable is therefore not a magnitude but a conjunction of two properties—bounded and source-free—and only the kernel mass has both (Proposition 9, Lemma 5). For comparison, the cross-version pair—the v8.2 single-recursion baseline `FieldNet108M200E` (67.10) against the same v9.2 CORE arm (51.23)—gives 15.87, and at 400M the corresponding pair gives 12.33 (52.55 against 40.22); both are cross-version and neither is a single-axis estimate either. Four numbers, one direction, none of them clean; we report them as corroboration and do not average them.

Neither alternative makes the model untrainable: both train throughout, and on downstream accuracy the three are within 1.7 points. What neither alternative delivers is a calibrated language model, which is why the mass is not attributable a single number here—removing it towards `none` costs 13.18 wPPL and the length extrapolation, removing it towards `rms` costs the calibration outright.

What the ablation does settle, and what it does not. It settles the direction and size of the obstacle: at matched lineage and scale, removing normalization costs 13.18 wPPL and destroys length extrapolation. What it does not settle is how much of that is the companion mass, because `none` removes the positionwise divisor too; the v8.2 UNNORM arm already shows that *some* divisor suffices

Table 35: **Length invariance, per checkpoint.** $\delta_{4096\to32768}$ is the largest change in any $A(s)$ entry between $L = 4{,}096$ and $L = 32{,}768$, with the scale at which it occurs; δ_{bound} is the largest change between $L = L_{\text{train}}$ and $L = 4L_{\text{train}}$. The shape itself spans $4\times$ to $74\times$, so the residual drift is two orders below the effect. The 400M UNNORM row is the loosest and the only one whose drift at the stretching end has a significant monotone trend ($+0.73\%$ per octave at $s = \frac{1}{4}$, $p < 10^{-4}$, region-paired).

Checkpoint	L_{train}	$\delta_{4096\to32768}$	at s	δ_{bound}
200M UNNORM	1,024	0.00%	—	7.27%
400M UNNORM	2,048	0.68%	$\frac{1}{2}$	1.62%
200M NORM	1,024	0.24%	$\frac{1}{4}$	0.83%
400M NORM	2,048	0.42%	$\frac{1}{4}$	0.26%

Table 36: **Arms and the axis each instantiates.** The 200M column is the ablation evidence; the 400M/1.3B columns are the main results.

Arm	Axis / coordinate	200M	400M	1.3B
SCALAR	rate granularity: scalar	✓	—	—
FIXEDDIST	spectrum: fixed distribution	✓	—	—
NOGATE	dissipation removed	✓	—	—
PERM	transition: content-dependent, bounded	—	✓	—
T-GATE	transition: content-dependent, strong (untrained)	—	—	—
NORM	normalization: mass present	—	✓	—
UNNORM	normalization: mass absent	—	✓	—

for extrapolation ($\rho = 1.006$ at 400M), so the failure of `none` is evidence about normalization, not about the mass.

32 ARCHITECTURE DETAILS

Decay rates: draw, clamp, and what is never reached. Rates are held in log space, so $\alpha_{h,d} > 0$ by construction. Two separate parameters bound them: rates are *drawn* log-uniformly from `decay_range` $= [\alpha_{\text{lo}}, \alpha_{\text{hi}}]$ at initialization, and are then confined during learning to the *clamp* $[\texttt{decay_hard_min} + \varepsilon,\ \texttt{decay_hard_max} - \varepsilon]$. The clamp is instance-specific and the two instances differ by a factor of ten in the configuration (0.001 against 0.0001), and by about twenty once the offset is applied, because ε is absolute and is 10^{-3} in the v8.2 instance but 10^{-6} in v9.x—effective floors of 0.002 and 1.0×10^{-4} respectively (Table 40). Both learned values sit above their effective floor, so neither is pinned by arithmetic; what differs is the room, and that difference is the precondition of Appendix 22. The upper endpoint is a design constant reached by the fit rather than set by it: the largest learned rate in the measured spectra is 0.9492 against an effective ceiling of $0.95 - \varepsilon$ (and 0.8984 against 0.899 in the v8.2 instance), indistinguishable from the ceiling at the 2^{-8} relative resolution of the stored `bfloat16` rates, so the fast end is bounded by arithmetic rather than by the fit.

What the layer does not contain. No additive positional encoding, no attention, no convolution over the time axis, and no content-dependent gate: Λ^{Δ_t} depends on position alone (C3). Each of these is a coordinate of Definition 2, not an implementation shortcut.

33 CONFIGURATIONS, DATA, AND COMPUTE

Data. FineWeb-Edu (Penedo et al., 2024) and the OpenCSG Chinese corpus (Yu et al., 2025), tokenized with a 32k Llama-2 tokenizer shared by every model in a matched pair. Data order is fixed by one seed and identical across a pair.

Configurations. Design choices were fixed before the main runs. Width and depth are not comparable across a pair, so a pair is matched on *parameter count* (Table 41); tokenizer, data, data order,

Table 37: **Checkpoint registry**: scale, training length and where each is reported. Training length is 2048 at 400M and above and 1024 otherwise. P2 is measured on four checkpoints spanning both scales and both instances (Tables 30, 28 and 11; Appendix 29); the 1.3B column is UNNORM throughout and the two SCALAR arms were not swept.

Checkpoint	Scale	L_{train}	Where reported
`FieldNet108M200E`	200M	1024	Tab. 30 (P2), 27 (200M sweep), 18 (200M row), 44
`FieldNet108M200ENoVector` (SCALAR)	200M	1024	Tab. 18, 44
`FieldNet108M404EV3` (UNNORM)	400M	2048	Tab. 14, 12, 1, 17, 18, 27, 11
`FieldNet108M404ENoVector` (SCALAR)	400M	2048	Tab. 18, 27
`FieldNet112_404M_EN` (v9.1 T-PERM, NORM)	400M	2048	Tab. 12, 1, 38, 16
`FieldNet112aCORE_200M_EN` (v9.2 CORE, NORM)	200M	1024	Appendix 31 (mass attribution), Tab. 11
`FieldNet112aRMS_200M_EN` (v9.2, `normalization_mode=rms`)	200M	1024	Appendix 31 (M3, the three settings), Tab. 3, 39
`FieldNet112aNONE_200M_EN` (v9.2, `normalization_mode=none`)	200M	1024	Appendix 31 (M3, the three settings), Tab. 3, 39
`FieldNet` UNNORM	1.3B	2048	Tab. 14, 26, 16

Table 38: **Zero-shot downstream, both instances at 400M** (lm-eval-harness; length-normalised accuracy for ARC, HellaSwag, OpenBookQA and PIQA, plain accuracy elsewhere; $\pm$ SE). Δ is NORM minus UNNORM. The UNNORM column is Table 17.

Task	UNNORM (single)	NORM (dual)	Δ
HellaSwag	35.1 ± 0.5	38.1 ± 0.5	$+3.1$
ARC-Challenge	27.1 ± 1.3	28.2 ± 1.3	$+1.1$
PIQA	64.5 ± 1.1	65.5 ± 1.1	$+1.0$
BoolQ	61.9 ± 0.9	61.7 ± 0.9	-0.2
ARC-Easy	50.7 ± 1.0	50.3 ± 1.0	-0.4
WinoGrande	51.8 ± 1.4	51.2 ± 1.4	-0.6
OpenBookQA	33.4 ± 2.1	32.6 ± 2.1	-0.8
Mean	**46.36**	**46.80**	$+0.44$

batch composition, step count and schedule are identical on both sides. Token budgets follow the compute-optimal rule (Hoffmann et al., 2022); both sides of a pair see the same number of tokens.

The ZH pair is a separate, smaller pair: it trains at L_{train}=1024 and uses a 6,400-token Chinese vocabulary, so its absolute PPL is not comparable to the EN columns and only its ρ is used. It is included because P2 is claimed across languages, and a second language at a different scale, training length and vocabulary is a stronger test of that claim than a second run in the same language.

Optimizer and schedule. AdamW, cosine schedule with warmup, bf16 autocast with the decay chain in fp32 (Appendix 35). Step counts are identical within a matched pair.

Compute. Eight devices per run. FIELDNET reports tokens/s alongside the matched baseline for every run (Appendix 41); the throughput gap against exact attention is a matter of kernel maturity and of the chunk-parallel scan, not of the asymptotic cost class—both are $\mathcal{O}(L)$ per token (Proposition 3).

Status. The three-axis arms run at 200M and 400M and are reported in Appendix 39. The content-gated coordinate is constructible but untrained (Section A.2, Appendix 49). What is fixed here is the axis mapping, not the outcome. The matched 1.3B Llama-2 baseline is trained and evaluated here (Tables 14, 26).

Table 39: **Three settings of the divisor, one lineage** (200M, v9.2, $L_{\text{train}} = 1024$; zero-shot accuracy and WikiText-103 word perplexity). The arms differ in `normalization_mode` alone. Absolute levels are not comparable with Table 38, which pairs v8.2 against v9.2.

Task	`mass`	`rms`	`none`
ARC-Challenge	24.06	23.04	21.76
ARC-Easy	58.00	54.88	52.78
BoolQ	59.63	61.10	60.58
HellaSwag	31.59	29.84	29.85
OpenBookQA	21.40	21.00	18.80
PIQA	64.25	63.11	62.57
WinoGrande	50.20	48.78	50.99
Mean	**44.16**	**43.11**	**42.48**
Word PPL (↓)	51.23	1.37×10^6	64.41
Bits/byte (↓)	1.062	3.813	1.124

Table 40: **Rate initialization and the learned slow tail** (read at one time point; drift in Table 16). The clamp is a *separate* parameter from the draw: rates are drawn from `decay_range` and confined during learning to $[\texttt{decay_hard_min} + \varepsilon,\ \texttt{decay_hard_max} - \varepsilon]$, with $\varepsilon = 10^{-3}$ for UNNORM and 10^{-6} for NORM. †Room is measured against that *effective* floor, not against the configured clamp: $0.002090/0.002 = 1.05$ and $0.001236/1.0 \times 10^{-4} \approx 12$.

Instance	Clamp α_{lo}	Learned $\alpha_{\min}$	Room
UNNORM	10^{-3}	0.002090	$1.05\times^{\dagger}$
NORM	10^{-4}	0.001236	$\approx 12\times^{\dagger}$

34 WHAT "MATCHED" MEANS IN A PAIR

The field layer replaces attention, so the two sides of a pair do not share a width or a depth: at the ∼400M pair Llama-2 is d=1024/29 layers (405.3M) and FIELDNET is d=1408/18 layers (404.3M) (Table 41). Parameter count is the only alignment point available and the only one we claim. The reason is structural rather than convenient: an $\mathcal{O}(L)$ recurrent state and an $\mathcal{O}(L^2)$ attention map do not admit a layer-for-layer correspondence, so matching width and depth would trade one confound for another—an attention layer and a field layer of the same d carry different state sizes and different cost profiles. Tokenizer, data, data order, batch composition, step count and schedule are held fixed, which is what makes the remaining coordinate the one under test.

At 1.3B a stronger match is available: both sides are d=2048 with 24 layers (1269M against 1299M parameters), so that pair is matched in width, depth and parameter count at once. It is the pair carrying the P1 result of Section A.1; the ∼400M pair, where only the parameter count matches, carries the quality and downstream comparisons.

35 TRAINING PROTOCOL AND RUN ACCOUNTING

Learning-rate protocol. Each architecture uses the learning rate established for it, as is standard when comparing architectures that do not share an optimizer landscape: Llama-2 3×10^{-4}, the value published for this architecture family, and FIELDNET 10^{-3}, the value used in every configuration reported here. The rates differ; what separates the two under P2 is architectural—position entering multiplicatively through the decay kernel versus additively through an encoding—and a longer or differently-tuned baseline run would not remove the degradation, which is a property of the encoding rather than of optimization.

Optimizer. AdamW (β_1=0.9, β_2=0.95), weight decay 0.05, gradient clipping 1.0, cosine schedule to $0.1\times$ peak with 2,000-step warmup. Effective batch 320 sequences ($10 \times 8 \times 4$: per-device batch, accumulation, devices).

Table 41: **Configurations actually trained.** Parameter count is the matching coordinate; width and depth are not comparable between an $\mathcal{O}(L)$ recurrent layer and $\mathcal{O}(L^2)$ attention.

Model	Params	d_{model}	Layers	Vocab	L_{train}	Scan chunk
Llama-2 (EN pair)	405.3M	1024	29	32000	2048	—
FIELDNET (EN pair)	404.3M	1408	18	32000	2048	512
FIELDNET 1.3B	1.27B	2048	24	32000	2048	512
Llama-2 1.3B	1299M	2048	24	32000	2048	—
Llama-2 (ZH pair)	300M	1024	23	6400	1024	—
FIELDNET (ZH pair)	300M	1280	18	6400	1024	256

Precision. bf16 autocast with master weights in fp32; the decay chain (cumulative exponent, re-basing at the chunk midpoint) is computed in fp32 throughout.

Seeds. Seeds: 1337 (main; rank-local, `base_seed + rank` under DDP). Ablation arms share the data seed with the main configuration, so comparisons are paired by construction.

Run accounting. Every reported run is listed with its configuration, device count, step count and tokens/s in `runs.jsonl`. Checkpoints, logs and the evaluation driver are archived with the submission.

36 COORDINATES OF THE SHARED DECAY-GATED RECURRENT FORM

Table 42: **Coordinates of shared decay-gated recurrent form** (Remark 1; body, Section 6): only core recurrent choices. FIELDNET sits at real-positive spectrum, position-only transition, mass present. P2 is defined wherever the transition is multiplicative in position; it is undefined only for lag-dependent kernels.

Model	Spectrum	Transition	Normalization	P2	Instance
FIELDNET (ours)	real +	position-only	mass	def.	dual
FIELDNET S-CPLX	complex	position-only	mass	def.	dual
S4D	complex	position-only	none	def.	single
RetNet	real +	position-only	output norm	def.	dual
Linear attention	real +	content (mult.)†	kernel mass	def.	dual
RWKV	real +	content (mix)	denominator	def.	dual
GLA	real +	content	denominator	def.	dual
Mamba	real +	content (selective)	none	def.	single

† P2 is defined whenever the transition is *multiplicative in position*—the rate may depend on the input but not on the lag. Under this definition the content-dependent rows above admit P2, and C3 is sufficient rather than necessary. P2 is undefined only for a lag-dependent transition such as a content-similarity kernel $q_t^\top k_s$.

37 POSITIONING OF S4D, AND WHY UNNORM CARRIES THE HORIZON READING

Two readings of S4D are available and we do not conflate them. Under Remark 1 a coordinate records only core recurrent choices, and on that reading S4D sits at (complex, position-only, no mass). Under a stricter reading—one that also counts the initialization scheme—S4D and UNNORM differ in more than the spectrum. Both readings are consistent with what we report, and we use the first.

Spectrum (the horizon reading). The two lineages are not two settings of one knob but two instances of the recurrence: UNNORM runs (3) alone and normalizes by a static rule, while NORM runs (3)–(6), the companion mass being accumulated by (5). The B–W horizon formula describes the single-recursion instance without correction: with companion mass the state is a convex combination (Lemma 5) whose effective horizon is set by the relative weighting of slow modes rather than by $1/\alpha_{\min}$ alone. Running the forecast on UNNORM therefore tests the formula on the instance it

describes, and we report the NORM deviation as a failure of the formula on the other instance (numbers in Table 16). The same split carries the precondition of Appendix 22: the forecast needs a slow tail that training does not move, and the clamp bounding that movement is ten times looser on the lineage carrying the second recursion (0.0001 against 0.001). We do not claim the failure is caused by the mass rather than by the clamp—in this pair the two are confounded, since the lineage that adds the recursion is also the one that loosens the clamp—only that both point the same way and that the formula is a claim about one instance rather than about both.

Initialization. S4D's HiPPO-derived initialization is a closed-form component under a prescribed measure; ours is a log-uniform draw whose endpoints are fixed by design and are, in the measured spectra, *both* passed by training: under NORM the draw interval is $[0.002, 0.8]$ while the learned spectrum reaches 1.764×10^{-4} at the slow end and 0.9492 at the fast end. The operative bound is therefore the clamp, not the draw—which is the point of the horizon reading, since a bound that training cannot move is exactly what a pre-learning predictor needs. The schemes differ, and the difference is visible in the spectrum: S4D's rates are not drawn to realize a slow tail, so the horizon reading's precondition is not met there either.

38 SATURATION: TRAJECTORY, AND WHAT THE FAILURE SAYS

The measurements. Under UNNORM at 400M the predicted horizon is $8{,}135$ and the measured saturation is $2{,}048$: high by $3.97\times$. At 1.3B the same forecast meets a measured $4{,}096$: high by $1.99\times$. Under NORM at 400M the prediction is $13{,}754$ and the measurement is $4{,}096$: high by $3.36\times$.

Why NORM differs. Three facts, none of which we adapt to fit. (i) $\alpha_{\min}$ is lower in absolute terms under NORM (0.001236 vs. 0.002090), so the formula returns a larger number. (ii) The learned $\alpha_{\min}$ sits above the clamp in both instances, so neither is pinned by arithmetic. (iii) The horizon under NORM is set by the relative weighting of slow modes rather than by $1/\alpha_{\min}$ alone, because the companion mass makes the state a convex combination (Lemma 5).

The forecast has a falsifier, and it was used. The forecast has a stated falsifier (Table 29): had saturation occurred at, say, 2K or 32K, the horizon reading would have failed. What makes this a test rather than a fit is structural, not temporal: the predictor is a quantity set *before* learning begins—the initialization value of $\alpha_{\min}$—while the quantity predicted is measured *after* training, so the two are not fitted to one another.

Two limits of the test, both binding. *Invariance is tested at one rate value*: no configuration varies $\alpha_{\min}$ while holding the rest fixed, so $1/\alpha_{\min}$ is checked at a single point, not as a law. *The same number is both predictor and predicted*: $\alpha_{\min}$ is read from the checkpoint whose saturation is measured, which is why the forecast cannot be read as a fit.

39 COMPONENT ABLATION (DETAIL)

Provenance of the granularity control numbers. The absolute wPPL values of Table 18 are from the v9 evaluation lineage. An earlier run under a prior protocol gave different absolute levels at the same Δ. The scale contrast that the granularity control rests on—0.29, not significant, at 200M against 7.12 at 400M—is stable across both lineages.

Zero-shot downstream at 200M. The two arms were also compared downstream at 200M (Table 44), with the same harness and metric conventions as Table 17.

Six of the seven gaps counted in the mean are inside the sum of the two standard errors; BoolQ (-8.5) is the only exception, and it is the same task that separates the arms at 400M in the other direction. We read the 200M downstream result as *the same absence that Table 18 reports*, now on a second axis: at this scale the two parameterizations are not separable by either measurement. This is why the granularity control is stated as a crossover that must appear at larger scale, rather than as a claim that per-dimension decay is uniformly better.

The 400M column is the one that matters for the granularity control and it is the source of the scale contrast in Section K: over a $2\times$ increase in scale SCALAR gains 7.72 while FIELDNET gains 14.55.

Table 43: **Component ablation**: the axis each arm instantiates, at both scales. SCALAR is competitive at 200M and falls behind at 400M, which is the granularity control's crossover; NOGATE removes dissipation and is the destructive control.

Arm	Axis	200M EN wPPL	400M EN wPPL
FIELDNET (reference)	—	67.10	52.55
SCALAR	granularity	67.39	59.67
FIXEDDIST	spectrum draw	$61.88^{\dagger}$	—
NOGATE	dissipation	$\times 364^{\dagger}$	—
S-CPLX	spectrum (complex)	67.02	—

Table 44: **Zero-shot downstream at 200M**: FIELDNET against SCALAR. Length-normalised accuracy (acc_norm) for ARC, HellaSwag, OpenBookQA and PIQA, plain accuracy for BoolQ and WinoGrande—the same convention as Table 17, applied to all seven rows. SCALAR is ahead by 1.04 mean accuracy; BoolQ (-8.5) is the only gap outside the combined standard error.

Task	FIELDNET	SCALAR	Δ
BoolQ	50.1 ± 0.9	58.6 ± 0.9	-8.5
PIQA	64.2 ± 1.1	64.3 ± 1.1	-0.1
ARC-Easy	50.9 ± 1.0	50.3 ± 1.0	$+0.6$
OpenBookQA	33.6 ± 2.1	34.0 ± 2.1	-0.4
WinoGrande	51.3 ± 1.4	50.7 ± 1.4	$+0.6$
HellaSwag	33.9 ± 0.5	33.7 ± 0.5	$+0.2$
ARC-Challenge	26.1 ± 1.3	25.8 ± 1.3	$+0.3$
Mean	**44.30**	**45.34**	-1.04

SCALAR at 400M corresponds to `use_vector_decay=False` with every other setting identical. The 200M FIXEDDIST and NOGATE arms were not re-run at 400M. †FIXEDDIST and NOGATE were measured under the earlier protocol and are not re-measured here; they are destructive and perturbational controls, and no conclusion in the paper depends on their exact value. The granularity pair, which the granularity control rests on, is measured at both scales under the final protocol. Length robustness of the two granularity arms is given in Table 27 and in the 200M sweep below.

A1 constrains the *dependence profile* f; the derivation needs the hypothesis on the *kernel*, i.e. the prior covariance under A3. We identify them up to a scalar, $G(x, x-k) \propto f(k)$: a modelling choice, not a theorem. Two consequences. Relaxing complete monotonicity to mere non-increasingness loses the mixture representation, so the spectrum is no longer derived from the profile; and A2 is a locality condition forbidding retroactive revision, weaker than full posterior invariance.

40 P2 AND CONTENT-DEPENDENT TRANSITIONS

P2 rescales inference positions $\Delta_t \mapsto s\Delta_t$. It is defined whenever the transition is *multiplicative in position*—the rate may depend on the input, but not on the lag—so that rescaling position rescales the log-rate spectrum rigidly. C3 is therefore *sufficient* for P2 and not *necessary*: content-dependent rates admit it too, and one of the four runs of Table 11 takes that road (the T-PERM arm). What P2 excludes is a genuinely lag-dependent transition, such as a content-similarity kernel $q_t^{\top} k_s$, where the contraction between two positions is not a function of the lag and no rate spectrum exists to translate. Mamba, RWKV and GLA *do* admit P2 under this definition; we report no P2 sweep for them because we hold no trained instance of those architectures, which is a gap in our experiments and not a structural exclusion.

41 THROUGHPUT ACCOUNTING

Measured throughput, tokens/s, at the two ends of the length grid (eight devices, same harness as the PPL protocol):

Model	$L = 1024$	$L = 32768$
FIELDNET 1.3B	9,305.6	7,983.1
Llama-2 405M	22,968.4	9,408.4

The two rows are *not* a matched pair—FIELDNET is 1.3B and the baseline is 405M—so only the rate at which throughput decays with length is comparable between them, not the level. FIELDNET loses 14.2% over a $32\times$ increase in length; the baseline loses 59.0%. A power-law fit of forward time against length gives an exponent of 0.944 for FIELDNET, against reference slopes of 1.000 for $\mathcal{O}(L)$ and 2.000 for $\mathcal{O}(L^2)$; the fit is reported as a measurement of our implementation, not as a claim about the complexity class, which is Proposition 3. Absolute throughput is a matter of kernel maturity—exact attention has had considerably more of it (Dao et al., 2022)—and P1, P2, M1 and M2 do not depend on it. The scan chunk size ($\mathrm{cs} = 512$ in these runs) is a numerical parameter fixed within a run.

42 THE HORIZON BOUND

Lemma 6 (body, step 5) states: fix a detection threshold $\varepsilon > 0$, a gain counting as detected at length L only if it exceeds ε in relative terms; the slowest mode enters the normalized readout with weight $\tilde{w}_{\min} \in (0, 1]$, contributing $\tilde{w}_{\min} e^{-\alpha_{\min} k}$ at lag k, so $L_{\text{meas}} \approx \alpha_{\min}^{-1} \log(\tilde{w}_{\min}/\varepsilon) \le \kappa/\alpha_{\min}$ with $\kappa \equiv \log(1/\varepsilon)$, equality iff $\tilde{w}_{\min} = 1$. The gap measures how strongly the readout attenuates the slow tail.

Proof. Setting the slow-mode contribution equal to ε and solving gives $k = \alpha_{\min}^{-1} \log(\tilde{w}_{\min}/\varepsilon)$; since $\tilde{w}_{\min} \le 1$ the bound follows, with equality iff $\tilde{w}_{\min} = 1$. Faster modes contribute less at the crossover lag and do not extend the horizon. □

We treat κ as calibrated: fixing it at the resolution at which the anchored protocol declares a length saturated gives $\kappa = 17$ (Appendix 38). This is a calibration and not a derivation, and the direction matters. Since $\kappa = \log(1/\varepsilon)$, the protocol's own PPL-level resolution (MDE $\sim 10^{-4}$–10^{-3}, Tables 23 and 22) would give $\kappa \in [6.9, 9.2]$ and hence a bound of 3,300–4,400 at $\alpha_{\min} = 0.002090$—which the measured $4{,}096$ would meet or exceed. The bound holds at $\kappa = 17$ because κ is fixed at a resolution orders of magnitude finer than the protocol can detect, and we do not claim otherwise: what the lemma delivers is the *form* $L_{\text{meas}} = \alpha_{\min}^{-1} \log(\tilde{w}_{\min}/\varepsilon)$ and the direction of the inequality, while its numerical value is set by a calibrated constant. The slow-mode-weight caricature is a mechanism, not a fitted model.

The test of record is therefore the dissociation reading of Section J, which fixes no κ at all; the numerical bound is reported with its constant stated and is not the load-bearing evidence. What would overturn the horizon reading is a measured horizon that exceeds the bound at the protocol's own stated resolution, or two instances that separate in horizon in proportion to their $\alpha_{\min}$.

How much room the bound has, in the units of the lemma. Lemma 6 can be inverted rather than only checked: $\tilde{w}_{\min} = \varepsilon\, e^{\alpha_{\min} L_{\text{meas}}}$. Read the three rows of Table 16 that way and the calibrated constant prices itself.

Table 45: **Implied slow-mode weight** $\tilde{w}_{\min} = \varepsilon\, e^{\alpha_{\min} L_{\text{meas}}}$ (final $\alpha_{\min}$), at two values of the detection threshold. A weight of order 10^{-6}–10^{-4} is not a mechanism; it says the bound holds with four to six orders of magnitude to spare.

Config	L_{meas}	$\tilde{w}_{\min}$, κ=17	$\tilde{w}_{\min}$, κ=9.2
UNNORM 400M	2,048	2.5×10^{-6}	6.2×10^{-3}
UNNORM 1.3B	4,096	1.6×10^{-4}	0.39
NORM 400M	4,096	4.9×10^{-6}	1.2×10^{-2}

At $\kappa = 17$ the implied weight of the slowest mode is 10^{-6} to 10^{-4} of the total—too small to be a description of any readout, and a direct measure of how much the calibrated constant contributes. At the protocol's own resolution, $\kappa \approx 9.2$, the same inversion gives 6×10^{-3} to 0.39: still attenuated, and

now of a size that could be checked against the learned weight distribution, which is the measurement the derivation hands on. We report both rather than choose, and we rest the horizon reading on the dissociation reading, which needs no κ at all.

Which $\alpha_{\min}$ the bound is computed from. The forecast uses the initialization value and the measurement uses the trained model. This is the conservative direction: on all three rows $\alpha_{\min}^{\text{final}} < \alpha_{\min}^{\text{init}}$ (Table 16), so $\kappa/\alpha_{\min}^{\text{init}}$ is the *smaller*—tighter—of the two bounds, and the drift column is what licenses substituting one for the other. The bound is therefore a design-time statement, readable before training, that is checked against the trained model through a drift the falsifier bounds at $\mathcal{O}(10\%)$.

43 What relaxing C3 costs

Lemma 11 (What relaxing C3 costs)**.** *Suppose the transition depends on the input, $A = A(u_t)$, so C3 fails. Then*

1. *the map from a past token to the present state is a product $\prod_j A(u_{t-j})$ depending on the intervening inputs, hence not a function of the lag alone; the Bernstein–Widder hypothesis is therefore not met, and no multiplicative gate Λ^{Δ_t} exists whose argument is elapsed position alone;*

P2 is not *forfeit by this failure: it survives whenever the rate depends on the input but not on the lag, and is lost only for a transition that is genuinely lag-dependent—a content-similarity kernel—where no rate spectrum exists to translate.*

Proof. (i) Under C3 the transfer operator is a function of the elapsed interval alone, so B–W applies to it as a function of lag. With $A = A(u_t)$ the product $\prod_j A(u_{t-j})$ depends on the inputs between s and t, not only on $t - s$; the B–W hypothesis—a function of the lag—is therefore not satisfied, and no Λ^{Δ_t} can represent the transfer. □

The statement is deliberately the weak one. We do not claim that no Bernstein–Widder representation exists for any content-dependent operator family—that would need a stronger argument. We claim only that the hypothesis the derivation uses is not met, which is enough to price the relaxation: the closed form, the horizon law and a rate spectrum that is a constant of the model are forfeit. Bounded normalization and P2 are not.

44 Scope of the extrapolation comparison

P2 compares FieldNet against unmodified Llama-2 (Touvron et al., 2023) trained at $L_{\text{train}} = 2048$ and evaluated at up to 32 768 with no adaptation of any kind—no NTK-aware RoPE rescaling, no YaRN, no positional interpolation (Chen et al., 2023), no ALiBi (Press et al., 2022). This is deliberate, and the scope of the comparison follows from it.

Why unmodified. The claim of P2 is about the form, not about which positional patch is best at 32K. Giving the baseline an interpolation method would change two things at once—the encoding and the comparison—and would make the resulting number a statement about that method rather than about attention's dependence on an additive encoding.

45 Numerical precision of the decay chain

The decay chain is kept in fp32. That choice is load-bearing, and this section gives the numbers that make it so.

A bf16 floor on the slow tail. $\lambda = e^{-\alpha}$ is stored near 1 for small α, and bf16 has an ulp of $2^{-8} = 0.00390625$ below 1. Any α with $e^{-\alpha} > 1 - \text{ulp}/2$, i.e. $\alpha < \alpha_{\text{crit}} = -\ln(1 - 0.001953125) = 0.0019550$, rounds λ up to exactly 1.0: the channel does not decay at all and its memory is unbounded. Measured against the initialization draws of Table 16, UnNorm sits 6.9% above that floor (0.002090

against 0.001955) while NORM sits 36.8% below it (0.001236). The slow tail of the dual-recurrence instance is therefore not representable in bf16 at all, which is the concrete reason the chain is carried in fp32 rather than being a stylistic preference.

And an accuracy floor above it. Being above the floor is not the same as being accurate. At $\alpha = 0.002016$—UNNORM's own regime—bf16 returns $\lambda = 0.99609375$, an effective rate of 0.003914 against a true 0.002016: a 94% error that halves the memory ceiling (255 against 496). The error falls below 2% only by $\alpha \approx 0.02$. So the slowest channels, which are the ones the horizon forecast is computed from, are exactly the ones bf16 distorts most.

Runtime accounting. Instrumenting one forward pass counts the dtypes actually taken: the $\exp$ and cumulative-sum calls that implement the decay run in fp32 in the reported runs, while the surrounding projections run in bf16. The $\mathcal{O}(d_{\text{model}})$ cache of Proposition 3 is unaffected by this split.

46 COMPUTE AND ARTIFACT MANIFEST

Every reported run appears in `runs.jsonl` with its configuration, device count, step count and tokens/s. The submission archive carries:

- `runs.jsonl` — per-run accounting (configuration, devices, step count, tokens/s).
- `eval_driver.py` — the anchored protocol: region partitioning, span selection, statistics.
- `metrics_*.json` — per-run loss curves and spectral statistics.
- Checkpoints and logs for the reported runs.

Two artifacts are not released: the full pretraining corpus (licence) and the 1.3B weights (size). Everything needed to reproduce the *measurements*—protocol, driver, metrics, and the reported checkpoints—is included.

47 NORMALIZATION: THE ALTERNATIVES A3 EXCLUDES

A3 selects the posterior mean, and each alternative changes the estimand rather than violating a constraint:

- *MAP rather than posterior mean.* Forfeits Lemma 5: the estimate is no longer a convex combination of sources, so the algebraic bound is lost.
- *Raw accumulation.* S4D's coordinate: boundedness falls back on the spectral radius, and the scale of the state is learned.
- *Median or other robust estimators.* No closed-form recursion, so the $\mathcal{O}(1)$-per-step guarantee fails.

The empirical counterpart is Section 12: dropping the mass leaves a model that still trains, and costs 12–16 wPPL at fixed scale, so what is lost is the algebraic bound and level rather than trainability.

48 RELATED WORK: EXTENDED

State-space models. S4D (Gu et al., 2022b;a) derives a diagonal state-space layer and initializes it from HiPPO (Gu et al., 2020); the coordinate it occupies is (complex, position-only, no mass). How to Train Your HiPPO (Gu et al., 2022c) generalizes the projection basis under the same prescribed-measure route, and the orthogonal-polynomial machinery behind it is classical (Gautschi, 2004). S5 (Smith et al., 2023) and the diagonal-plus-low-rank line relax diagonality, which our basis excludes by C2 at fixed state size.

Gated and content-dependent recurrences. RWKV (Peng et al., 2023), GLA (Yang et al., 2024) and Mamba (Gu & Dao, 2024) relax C3, so decay depends on the input; the cost is listed in Table 21 and the consequence for P2 in Appendix 40. None is refuted here.

Retention and linear attention. RetNet (Sun et al., 2023) and linear-attention variants normalize, as A3 prescribes, but each occupies one half of the conjunction in Section 2: RetNet normalizes at the output rather than by the kernel mass, and linear attention has a content-dependent kernel weight rather than a time-homogeneous one. Neither therefore reaches (2). The remaining difference from our coordinate is in the *spectrum*: neither derives rates from the dependence profile.

Extrapolation techniques. Positional interpolation (Chen et al., 2023) and ALiBi (Press et al., 2022) buy length by modifying or re-parameterizing an encoding; FIELDNET needs neither, by Theorem 4. We compare against the degradation of a matched baseline rather than against these techniques, because the claim is about the form, not about which patch is best at 32K.

49 EXTENSIONS SUPPORTED BY THE RELEASED IMPLEMENTATION

The claims of this paper form a closed loop: derivation (Section 3), predictions (P1, P2), and validation (Section 5 and the tables cited there). The items below are *extensions* the released implementation supports; none is required for any claim above, and no number in this paper depends on them. The chain is conditional—every claim is *if* A1–A3 and C1–C3—and what it leaves open, item by item (summarized in Section 6).

A1 is a hypothesis about language, not a theorem. Complete monotonicity is the load-bearing assumption; the power law motivates it but does not entail it. If dependence is not CM, the mixture representation fails and the spectrum is not derived from the profile— the architecture may still work, but not for the reason given here.

C3 is a design choice, and its relaxation is measured. Mamba, RWKV and GLA show that content-dependent transitions work well, so C3 is a choice, not a necessity. What the relaxation costs is the closed form, the horizon law, and the fact that the translated spectrum is a constant of the model rather than a per-input quantity—not boundedness, and not P2, which survives any transition that is multiplicative in position (Appendix 19). The weak end of that axis *is* measured and is the best arm in the paper (Section A.2); the strong end is an effect-size extension rather than a structural question.

The granularity crossover is measured, not derived, and the horizon reading carries a precondition. N^* is bracketed by experiment at $\approx$200M–400M; the basis says only that a crossover exists. Separately, the horizon reading as tested here is a statement about the single-recursion instance: it needs a slow tail that training does not move, and Section J locates the one failure exactly where that precondition is loosest.

P2 has not been swept on the granularity arms. The four checkpoints of Table 11 span scale and instance but not rate granularity: the two SCALAR arms were not swept for P2, so the crossover of Appendix K and the shape of P2 are at present measured on disjoint sets of checkpoints. A per-dimension and a scalar arm swept at the same scale would test whether the shape is a property of the position axis alone.

The divisor comparison does not transfer across architectures. The arms of Table 3 fix one lineage and move one axis: three move the divisor and S-CPLX moves the spectrum instead. The earlier UNNORM arms use the same divisor mathematics as `rms`—Appendix 31—and differ in the architecture around it, and they support language modelling where `rms` does not. So "only the posterior-mean divisor supports language modelling" is a statement about one lineage, not about the divisor in general, and we scope it that way throughout. A single-axis comparison inside the earlier lineage would say whether the same ordering holds there; it is not run.

The mass has no single-axis price. All three settings are now run at 200M in one lineage (Appendix 31), and they are not ordered: `none` keeps the magnitude and loses the bound, `rms` keeps the bound and loses the magnitude. So 13.18 wPPL is not an upper bound on a quantity—there is no single quantity to bound. What the mass buys is the conjunction of being bounded and being source-free, and the ablation prices the two failures separately rather than the mass. The same three-way comparison at 400M is not run, and the `rms` arm was not swept for length.

The horizon constant is calibrated. $\kappa = 17$ is fitted once, not derived; $1/\alpha_{\min}$ is checked at a single value of $\alpha_{\min}$, not as a law.

Extensions, and what each would add. The released implementation supports each of the following; none is required for any claim in the body.

(i) The fine sweep near $s = 1$ has been run and does not resolve the displacement (Appendix 51.1, Table 49): $\bar{D} = -0.0076 \pm 0.0040$, $t = -1.91$, against a minimum detectable effect of 0.0078 at $R = 16$. Following the rule registered before running it, the coverage term is absent and the mechanism stays open. Resolving the predicted sign needs $R \approx 64$; we report the resolution rather than claim the result.

(i-b) Strong gating + companion mass: an effect-size extension. Section 2 argues that the coordinate is well defined: selectivity from the content-dependent transition, boundedness from the source-free companion mass, and Table 2 verifies boundedness there without training. What remains is the effect size—the most informative experiment at the strong end of the axis, because it would measure the gain from the two capabilities we argue are separable, and it would do so without touching the argument made here, which concerns the C3 corner. The combination is supported by the released implementation; no number in this paper depends on it.

(ii) The spectral warp is not run (Appendix 51). It is what would make the candidate mechanism (20) identifiable: warping the spectrum inside one checkpoint holds the estimator fixed and moves the spectrum along an axis orthogonal to (7), giving 35 points for three parameters. Until it is run, (20) is unfitted and no number in this paper comes from it.

(iii) Rate and depth are not separated, and the crossing design is not the way to separate them. The two decompositions of Section 3.3 agree that the fast end dominates but disagree on the magnitude by up to an order of magnitude. We considered a 2×2—fast and slow channels crossed with shallow and deep layers—and we state why we do not expect it to work, because a reader will otherwise propose it. The design assumes the four cells are disjoint classes of channel, and on these spectra they are not: the rate distribution of a shallow layer and of a deep layer overlap heavily, so "slow channels of a shallow layer" and "fast channels of a deep layer" are largely the same channels at the same rates. Crossing two correlated factors does not orthogonalise them; it produces cells whose contrast is still a mixture of both.

What does separate them is *matching* rather than crossing: restrict to a rate band that both the shallow and the deep half occupy, sample equal numbers of channels from each half at matched rates, and scale one group. Any difference in $A(s)$ then has nowhere to come from except depth, because rate has been matched away. The script is released and reports the overlap coefficient and the size of the matched sample before scoring, so that a reader can see whether the design had any power. Twenty scorings. Not run; the two outcomes are written down in advance in Section 3.3.

(iv) The 1.3B row of the P2 grid is not run. The 1.3B checkpoint is on the position-only road, so P2 is defined on it and its spectrum is measured; adding its row would say that the coordinate-level prediction holds at the largest scale we trained. It would not say that the dual recursion exists at 1.3B, and is not used to say so (Section 6).

50 The learned rate spectra, and how P2 reads them

Per-channel rates are read off each layer's decay parameterisation ($\alpha = \sigma(\text{logit}) \cdot (\alpha_{\max} - \alpha_{\min}) + \alpha_{\min}$, so the hard bounds are the clip points and the reported minimum is the *learned* minimum). Table 9 in the body is computed from these by (14)–(15) with no fitting.

Table 46: **Measured spectra.** α over all channels of all layers.

Checkpoint	n	$\alpha_{\min}$	median	$\alpha_{\max}$	$\text{sd}(\log_{10}\alpha)$	$> 10^{-1}$
400M UnNorm	25,344	2.014×10^{-3}	7.227×10^{-2}	0.887	0.658	19.6%
400M Norm	25,344	1.64×10^{-4}	5.786×10^{-2}	0.949	0.807	43.9%
1.3B UnNorm	46,080	2.014×10^{-3}	3.711×10^{-2}	0.902	0.623	30.2%

The two 400M spectra differ in all three summary statistics, not only in the tail: NORM has a median 25% lower, a spread 23% larger ($\mathrm{sd}(\log_{10}\alpha)$ 0.807 against 0.658) and a slowest channel $12\times$ slower. The share of fast channels differs by more than a factor of two (43.9% against 19.6%). So the instances are separated along more than one spectral axis, which is part of why the mechanism is not identifiable on them (Appendix 51).

The measured spectra, and one thing they correct. Reading the rates directly—rather than reconstructing them from summary statistics—changed three numbers, and we report the correction because one of it reverses an ordering. Reconstructing a lognormal from $(\mathrm{median}\,\alpha,\ \mathrm{sd}(\log_{10}\alpha),\ \alpha_{\min})$ gives $N_{\mathrm{eff}}(1) = 12.3$, 12.5 and 22.6; the measured values are 21.7, 14.0 and 12.8. The 400M UNNORM and 1.3B values are therefore *in the wrong order* under reconstruction, and the learned $\alpha_{\min}$ of NORM is 1.6×10^{-4} against a reconstructed 1.2×10^{-3}, a factor of 7 that propagates into $K = 1/\alpha_{\min}$ (6,096 against 860). The reason is that the measured distribution is heavy at *both* ends—24.9% of channels below 10^{-2} and 43.9% above 10^{-1} for NORM—which a single lognormal matched on three moments does not reproduce. Everything in the body is computed from the measured vectors.

What changed, side by side. The corrections are collected in one place rather than left for the reader to diff against an earlier version.

Table 47: **Reconstructed against measured.** The reconstructed values are what a lognormal matched on $(\mathrm{median}\,\alpha, \mathrm{sd}(\log_{10}\alpha), \alpha_{\min})$ gives; the measured values are what reading the rates gives. Two entries reverse an ordering, which is why the correction matters.

Checkpoint	quantity	reconstructed	measured
400M UNNORM	$N_{\mathrm{eff}}(1)$	12.3	21.7
400M NORM	$N_{\mathrm{eff}}(1)$	12.5	14.0
1.3B UNNORM	$N_{\mathrm{eff}}(1)$	22.6	12.8
400M NORM	$\alpha_{\min}$	1.16×10^{-3}	1.64×10^{-4}
400M NORM	$K = 1/\alpha_{\min}$	860	6,096
400M NORM	$\mathrm{sd}(\log_{10}\alpha)$	0.748	0.807

The last row is worth a sentence: the earlier text said the two 400M spectra were "close in median and spread." Measured, their spreads differ by 23% (0.807 against 0.658), so they are separated along more than one axis. That has a consequence for the mechanism (Appendix 51): with two axes moving at once and the divisor moving with them, three checkpoints cannot identify two coefficients regardless of how precisely the spectra are known.

51 A CANDIDATE MECHANISM FOR THE SHAPE OF P2 (HYPOTHESIS, NOT FITTED)

Everything in Section 3.3 is parameter-free. This appendix states one candidate for the signed term that (C2) of that section shows is required—the *tail* candidate—and is explicit that it is **not fitted in this version**: no number in this paper is obtained from it, and no claim in the body depends on it. It is not the only candidate. (C2) excludes the level-type terms and in doing so names a second survivor, the saturation-induced dispersion $\mathrm{Var}_d[\ln r]$, which is one-sided and needs no horizon; the mechanism below chooses the tail over the dispersion as a hypothesis to test, and nothing in Section 3.3 depends on that choice being right. Two further considerations bear on that choice. The first is structural: as Appendix 50 shows, T_K at the lag the spectrum supplies draws on 0.27%–6% of the channels, so a mechanism whose only spectral input is T_K at that lag is close to a mechanism with a single-channel input. The dispersion term has no such pathology.

The second is a zero-parameter test that uses only measured quantities, and it is the closest thing to evidence we have. Take each candidate F, compute its separation ratio between each pair of checkpoints, and compare the three ratios with the three measured separation ratios of $A(4)/A(1/4)$. No coefficient is fitted; the question is only whether the *ordering* of the three pairs matches.

Table 48: **Separation ratios, measured.** Each entry is $\max(F_a, F_b)/\min(F_a, F_b)$ for a checkpoint pair, against the same ratio for the measured asymmetry. No fitting. The two T_K conventions disagree with each other by more than they disagree with anything else, and both get the ordering of the three pairs wrong; the two non-tail quantities get it right. Two quantities matching is not identification, and we do not treat it as one.

	pair		
F	N/U	N/1.3B	U/1.3B
measured $A(4)/A(1/4)$	1.82	1.16	2.11
$T_K(1)$, own lag	46.7	14.7	3.19
$T_K(1)$, $K = 2048$	11.1	71.0	6.41
$\mathrm{Var}_d[\ln r]$ at $s = 4$	3.58	1.28	4.57
$\mathrm{Var}_d[\ln r]$ at $s = 2$	4.60	1.20	5.51
$N_{\mathrm{eff}}(1)$	1.54	1.09	1.69

Read the ordering, not the magnitude: the measured asymmetry is largest for U/1.3B and smallest for N/1.3B, and so are $\mathrm{Var}_d[\ln r]$ and N_{eff}, while T_K is monotone decreasing across the pairs in one convention and peaks in the middle in the other. We state what this does and does not support. It does not identify the mechanism—two surviving candidates and three pairs is not identification, and N_{eff} is excluded as a *complete* account by (C2) even though it tracks the ordering. What it does is make the tail candidate the only one that is disconfirmed by data we already have, in both of the conventions it can be evaluated in, and it does so without a fitted parameter. Combined with the layer decomposition of Section 3.3, which puts the response in the fast half, and with the effective-sample-size result, which makes T_K a statistic of a few hundred channels, the three measurements point the same way. We stop short of calling that a conclusion about the mechanism, because the warp sweep is what would make it one. It is included because (C2) makes the question well-posed and because the test is cheap.

The candidate. Let $a(s) = \ln A(s)$. The closure assumption—the thing under test—is

$$a(s) = \beta_s \left[\ln M_2(s) - \ln M_2(1)\right]^2 \; + \; \beta_c \left[T_K(1) - T_K(s)\right], \qquad \beta_s, \beta_c > 0, \tag{20}$$

with M_2 and T_K from (14)–(15). The first term is the sharpness mismatch and is symmetric to first order (it is squared in the log-shift); the second is the coverage loss and is *signed*: $s > 1$ truncates the tail, $s < 1$ extends it. The coefficients are deliberately not called κ, which is reserved for the detection threshold of the horizon law, $\kappa \equiv \log(1/\varepsilon) = 17$ (Appendix 42); and the K here is a *lag in tokens*, not that threshold and not the memory constant of the horizon law. Three parameters (β_s, β_c, K), shared across checkpoints; checkpoints differ only in spectrum.

Why it is not identified on the data we have. Two obstacles, both structural. *(a)* The 2×2 grid over scale and instance varies the *estimator* together with the spectrum—the two 400M checkpoints differ in divisor, not only in rates—so spectrum and estimator are confounded, which is precisely the distinction the paper rests on. *(b)* The two functionals do separate the checkpoints, but not in a way that identifies the coefficients. Measured: $N_{\mathrm{eff}}(1)$ is 21.7 against 14.0 (1.55×), and $T_K(1)$ at each spectrum's own lag is 4.0×10^{-2} against 8.6×10^{-4} (47×) (Table 9). The two 400M instances therefore differ in *both* variables at once, and they differ in the divisor as well, so with three checkpoints and six scales per curve the two coefficients compensate: whatever the measured asymmetry is, a pair (β_s, β_c) can be found that reproduces it. One observation is worth recording because it bears on which candidate is more promising, though it is two points and confounded and we do not rest anything on it: the measured asymmetry differs by only 1.8× (8.10 against 4.46) while T_K at the spectrum's own lag differs by 47× and N_{eff} by 1.55×—i.e. the coverage variable varies far more than the quantity it is supposed to explain, while the sharpness variable varies comparably. That is a hint against the tail term and toward the saturation term of (C3), and the warp sweep is what would settle it. We therefore do not report β_s, β_c or K.

The test, and the decision rule. Both obstacles are removed without training. Warp the learned rates inside one checkpoint,

$$\alpha_d \;\longmapsto\; \alpha_{\mathrm{med}}^{1-p}\, \alpha_d^{p}, \qquad p \in \{0.6, 0.8, 1, 1.25, 1.6\}, \tag{21}$$

which preserves the median and scales $\mathrm{sd}(\log \alpha)$ by p, holding estimator, weights, tokenizer, data and schedule fixed; then re-invert the parameterisation and re-sweep s. This moves the spectrum along an axis orthogonal to (7), giving $5 \times 7 = 35$ points for three shared parameters. The decision rule is stated in advance: fit (20) jointly over all p and s; if the residual RMS in $\ln A$ is comparable to the region-paired resolution ($\approx 0.1\%$, Appendix 52), the mechanism is supported; if not, the signed term is misspecified and (C2) of Section 3.3 stands as the parameter-free statement while (20) is withdrawn. Either outcome is a result. The sweep is listed among the extensions (Appendix 49); nothing in the body waits on it.

51.1 A FINE SWEEP NEAR $s = 1$

Consequence (i) above predicts an argmin strictly below $s = 1$, by an amount the sampled grid cannot resolve. The test is a fine sweep $s \in \{0.80, 0.85, 0.90, 0.95, 1, 1.05, 1.10, 1.20\}$ at one length $L = 8{,}192 > L_{\mathrm{train}}$—the shape is length-invariant (Table 35), so one length suffices, and shape invariance means the whole sweep costs eight scorings per checkpoint, about 15 minutes. Region-paired resolution on $A(s)$ is $\approx 0.1\%$ (Appendix 52). Both outcomes were written down in advance, and the sweep has now been run on NORM at $L = 8{,}192$, $R = 16$, on $s \in \{0.80, 0.85, 0.90, 0.95, 0.98, 1, 1.02, 1.05, 1.10, 1.15, 1.25\}$.

Table 49: **Fine sweep near** $s = 1$ (NORM, $L = 8{,}192$, $R = 16$ anchored regions, baseline 31.83 PPL). $A(s)$ is perplexity relative to $s = 1$; SE is the region-paired error on $\ln A$.

s	0.80	0.85	0.90	0.95	0.98	1.02	1.05	1.10
$A(s)$	1.1654	1.1265	1.0735	1.0137	1.0001	1.0056	1.0214	1.0522
SE	.0134	.0111	.0074	.0027	.0009	.0007	.0019	.0037

s	1.15	1.25
$A(s)$	1.0788	1.1346
SE	.0048	.0062

The registered test is the paired comparison of the symmetric pair: with $D_i = \ln \mathrm{PPL}_i(0.95) - \ln \mathrm{PPL}_i(1.05)$, $\bar{D} = -0.0076$, SE $= 0.0040$, $t = -1.91$, $p = 0.056$, 95% CI $[-0.0153, +0.0002]$, and the minimum detectable effect at $R = 16$ is 0.0078. So the outcome is the second one: the displacement is not resolved. Following the rule written down in advance, the coverage term is absent and (20) reduces to its symmetric part—which (C2) has already excluded as a complete account, so the mechanism is left open rather than settled. The grid argmin is $\hat{s} = 1.00$ and the cubic fit in $\tau = \ln s$ gives $b_1 = -0.037 \pm 0.127$, i.e. $a'(1)$ is not resolved from zero either. One reading note: the asymmetry at these scales is individually significant at every point ($|t|$ from 5 to 20) while the *difference* of the symmetric pair is not, which is exactly the signature of an effect that is second order in τ with a vanishing first-order term. Reaching the predicted sign would need $R \approx 64$, which we report rather than claim.

52 REGION-LEVEL STATISTICS FOR EVERY REPORTED COMPARISON

All evaluations use the anchored protocol: every length scores the same window $[\mathrm{anchor}, \mathrm{anchor} + 4096)$ over the same $R = 16$ regions, so differences across lengths and across arms are *paired* on the region. Reported throughout: the paired difference of log-perplexity $\Delta = \ln \mathrm{PPL}_A - \ln \mathrm{PPL}_B$, its region-level SE $\mathrm{sd}(\Delta)/\sqrt{R}$, a two-sided paired t-test, a Holm correction over the lengths tested, and the minimum detectable effect $(t_{0.975,R-1} + t_{0.80,R-1}) \times \mathrm{SE}$.

The third row is the noise floor for two arms that differ on a single coordinate—the regime of every single-axis claim in this paper; it is the number that fixes the power calculation of Section 5.1.

Table 50: **P1, with region-level error.** $\rho(L)$ at $L = 32{,}768$ against L_{train}, geometric mean over $R = 16$ paired regions.

Checkpoint	L_{train}	ρ	$\text{SE}(\ln \rho)$	95% CI
1.3B UNNORM, EN	2,048	0.9997	0.00031	[0.9990, 1.0003]
400M NORM, EN	2,048	0.9975	0.00042	[0.9967, 0.9984]
400M UNNORM, EN	1,024	0.9983	0.00263	[0.9928, 1.0039]
200M UNNORM, EN	1,024	0.9992	0.00013	[0.9990, 0.9995]
1.1B UNNORM, ZH	2,048	0.9998	0.00019	[0.9994, 1.0002]
305M UNNORM, ZH	1,024	0.9897	0.00055	[0.9885, 0.9908]

Table 51: **Arm comparisons, resolved.** Region-paired, $R = 16$, Holm over the lengths in each row's set. $\Delta > 0$: the first arm is worse. MDE is the smallest effect detectable at $\alpha = 0.05$, power 0.8, at this R.

Comparison	L	$\Delta(\ln)$	SE	p (Holm)	MDE
1.3B FIELDNET vs. matched Llama-2	2,048	$+0.473$	0.059	1.5×10^{-6}	18.8%
400M NORM vs. 400M UNNORM	2,048	-0.163	0.040	1.6×10^{-3}	12.6%
200M vs. 200M, one axis changed	2,048	-0.019	0.008	6.8×10^{-2}	2.4%