Learning Length-Extrapolatable Recurrent Models

Credit Stabilization through Time (CST)

Paper: Hanwen Jiang | Adobe Research
arXiv:2609.09157v1, submitted 8 September 2026

Central idea

Rescale backward state credit at recurrent boundaries to improve behavior beyond the training horizon.

Publication not found as of 29 October 2026: arXiv record, author page.

Public code not found as of the same date: author GitHub.

Research-group reading presentation; the presenter is not a paper author.

A fixed-size state can run longer than it can learn

Train short, deploy long

Learn a shared recurrent rule on short sequences; retain and use information over many more steps at inference.

- A recurrent architecture makes long execution feasible.
- Learning still depends on how future losses reach early computation.
- Dense local supervision can train a shared rule even when late-loss credit vanishes.

Question: when is distant credit needed, and where should it be stabilized?

Paper, Secs. 1–3, PDF pp. 1–5.

Three task families separate retention from computation

Task	Context	What a longer sequence adds
Fixed-FSA	One shared transition rule	More dependent steps: $N = D$, $M = 0$
MQAR	New key–value bindings	More time before retrieval; $D = 0$
Meta-FSA	New transition table	K segments of depth D; $N = KD$

- M: memory load; N: task-input horizon; D: transition depth.
- Meta-FSA resets the *task state* between segments, while the model state retains the context.
- Experiments hold M fixed: temporal extrapolation is tested.

Paper, Sec. 2 and Table 1, PDF pp. 2–3; serialization in App. C.1, pp. 21–22.

State credit sits upstream of the parameter gradient

$$h_t = \Phi_\theta(h_{t-1}, x_t), \qquad \hat{y}_t = G_\psi(h_t), \qquad \mathcal{L} = \sum_{t \in \mathcal{I}} \ell_t$$

$$\lambda_t^{(\tau)} = J_{t+1}^\top \cdots J_\tau^\top \nabla_{h_\tau} \ell_\tau, \qquad J_j = \frac{\partial h_j}{\partial h_{j-1}}$$

$$\lambda_t = \sum_{\tau \in \mathcal{I}:\, \tau \geq t} \lambda_t^{(\tau)}, \qquad \boxed{g = \nabla_\theta \mathcal{L} = \sum_t P_t^\top \lambda_t}$$

- λ_t: credit assigned to state h_t by future losses.
- P_t: local parameter Jacobian, holding the incoming state fixed.
- Mapping and summation can obscure the upstream credit trajectory.

Paper, Sec. 3.1, Eqs. (1)–(5), PDF p. 4. Assumes h_0 independent of θ; recurrent θ and readout ψ are disjoint.

Distant-credit necessity and norm survival disagree

$$g_{\mathrm{far}} = g_{\mathrm{full}} - g_{\mathrm{local}}, \qquad p_{\mathrm{far}} = \frac{\langle g_{\mathrm{far}}, g_{\mathrm{full}} \rangle}{\|g_{\mathrm{full}}\|^2}$$

Task	Far-gradient projection p_{far}	Late-loss retention S_1
Fixed-FSA	0.021	≈ 0
Meta-FSA	0.089	1.05
MQAR	0.259	0.82

- S_1: norm ratio for one fixed late-loss window, about 895 positions apart.
- MQAR intervention: remove $g_{\mathrm{far}} \Rightarrow$ 0.60% accuracy; full BPTT $\Rightarrow$ 96.30%.
- g_{local} keeps all losses but detaches states at each boundary.

Paper, Table 2 and Eqs. (6)–(9), PDF pp. 4–5; App. A, p. 19. $T = 1024$, $H = 128$. p_{far} is signed, not a probability; random guessing is 0.098%.

CST rescales incoming credit before it crosses a chunk

$$\underbrace{\lambda_{k-1}}_{\text{earlier boundary}} = \underbrace{q_{k-1}}_{\text{current-chunk loss}} + \overline{J}_k^{\top} \underbrace{\lambda_k}_{\text{later-chunk losses}} \qquad \text{(BPTT)}$$

CST intervention

$$\widetilde{\lambda}_k = \gamma_k \lambda_k, \qquad \lambda_{k-1} = q_{k-1} + \overline{J}_k^{\top} \widetilde{\lambda}_k, \qquad \gamma_k > 0$$

- $\gamma = 1$: exact BPTT; $\gamma = 0$: cut the cross-boundary signal.
- Positive scaling preserves the corrected component's local direction.
- Forward computation and loss are unchanged; the modified backward signal is generally not the exact objective gradient.

Paper, Sec. 4.1, Eqs. (10)–(11), PDF p. 6; App. C.2, p. 22. $\overline{J}_k$ is the chunk-state Jacobian; $q_{k-1} = \nabla_{h_{b_{k-1}}} \mathcal{L}_k$.

Event-CST waits for substantial accumulated contraction

Let $n_k = \|\lambda_k\|$, anchor A, trigger ratio $\rho > 1$.

$$\gamma_k = \begin{cases} \min\{A/\max(n_k, \epsilon),\, \gamma_{\max}\}, & n_k \le A/\rho, \\ 1, & n_k > A/\rho. \end{cases}$$

$$A \leftarrow \begin{cases} n_k, & n_k > A, \\ \gamma_k n_k, & n_k \le A/\rho, \\ A, & A/\rho < n_k \le A. \end{cases}$$

- New peak: raise the anchor. Small contraction: hold it.
- Event: restore up to the cap, then anchor at the achieved scale.
- Adjacent-CST instead corrects every adjacent norm decrease.

Paper, Sec. 5.1, Eqs. (12)–(14), PDF p. 7. Initialize at the latest valid boundary with gain 1; invalid/near-zero norms use identity gain.

EMA replay reuses locations while recomputing live gains

1. Every P steps, run a dense identity-gain backward probe.
2. Smooth signed log-contraction over training probes:

$$d_k^{(p)} = \operatorname*{median}_q \log \frac{n_{k+1}^{(q,p)}}{n_k^{(q,p)}}, \quad \overline{d}_k^{(p)} = \beta \overline{d}_k^{(p-1)} + (1-\beta) d_k^{(p)}.$$

3. Predict boundaries where accumulated drawdown reaches $\log \rho$.
4. At ordinary steps, expose those boundaries and compute current gains.

$$\text{Boundary overhead:} \quad \mathcal{O}(B\,\mathrm{size}(h)) \longrightarrow \mathcal{O}((R + B/P)\,\mathrm{size}(h))$$

Paper, Eq. (15), PDF p. 8; App. C.3, p. 22. q: state component; B: dense candidates; R: replayed boundaries. Probe steps also perform an ordinary BPTT update.

The main synthetic test separates length from depth

Training

- Gated DeltaNet: 3 layers, width 512, 16 heads.
- Meta-FSA: 32 states, 4 actions.
- Physical length 1024; depth $D = 4$.
- 100K steps; global batch 512; three seeds.

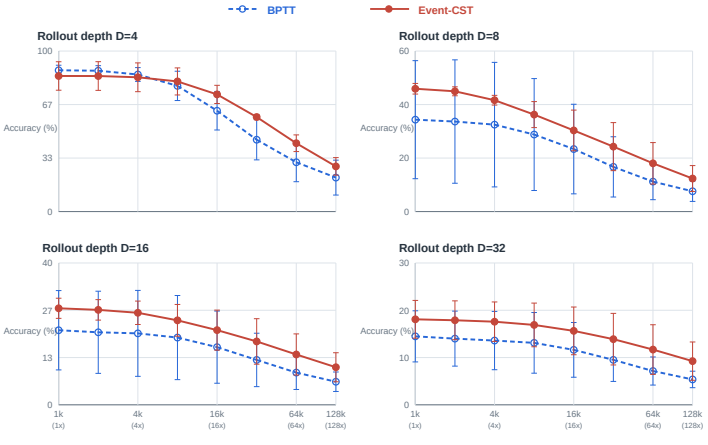
Evaluation

- Lengths 1K–128K.
- Depths $D = 4, 8, 16, 32$.
- 1024 examples per depth–length cell.
- Accuracy: mean and sample SD across training seeds.

Event-CST: $H = 32$, $\rho = 1.5$, $\gamma_{\max} = 2$.
EMA replay: same controller, $P = 100$, $\beta = 0.9$.

Paper, Sec. 5.2, PDF p. 8. No curriculum; 500-step warmup, then constant LR 3×10^{-4}.

Meta-FSA gains appear beyond the training horizon



16K–128K
All 16 mean cells improve.
+6.58 pp on average.

$D = 4$, **128K**
21.16% → 28.14%
+6.98 pp

At 1K, $D = 4$
−3.55 pp.

Original Fig. 1, PDF p. 8; Sec. 5.2.1, p. 9. Error bars: three-seed sample SD. Of 48 paired seed–cell comparisons at 16K–128K, 29 improve.

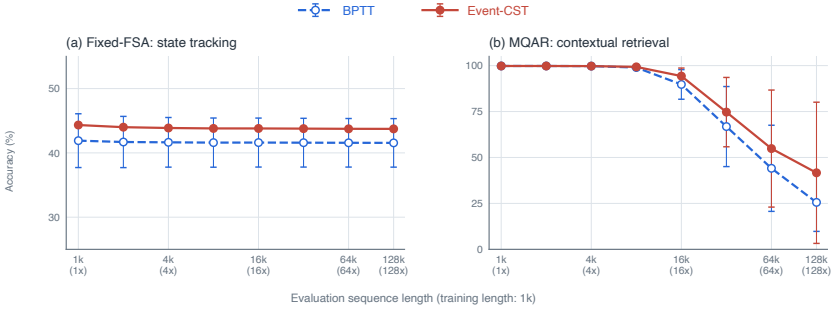
Accuracy improves at a substantial training cost

Method / implementation	Long-context accuracy	M tokens/s
BPTT, direct	18.5%	0.955
Adjacent-CST, CUDA Graph	22.1%	55.1% of BPTT
Event-CST, eager	25.1%	0.119
Event-CST, CUDA Graph	25.1%	0.526
EMA replay, sparse Graph	23.2%	0.808 projected

- Dense Graph Event-CST reaches 55.1% of BPTT throughput.
- EMA replay exposes 6.91 rather than 31 boundaries per step.
- EMA throughput range 0.75–0.82M is inferred from equivalent sparse graph benchmarks, not an end-to-end measurement.

Paper, Fig. 2 and Sec. 5.2.2, PDF p. 9. Single A100 80 GB, batch 64; implementations differ. Accuracy averages four depths and lengths 16K–128K.

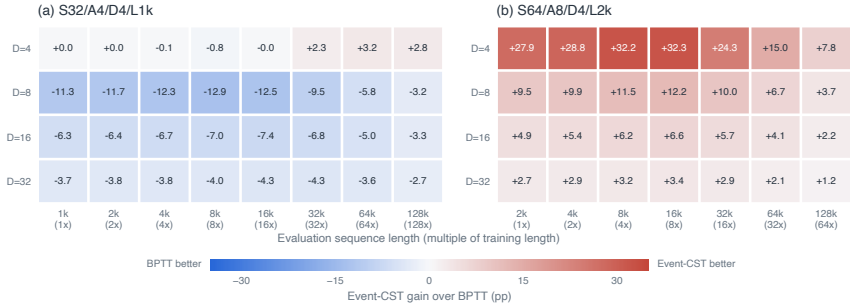
Retention tasks benefit more as the horizon grows



- Fixed-FSA: 41.58% → 43.76% over 16K–128K; length dependence stays nearly flat.
- MQAR at 128K: 25.56% → 41.66%, +16.10 pp. The 128 stored bindings remain fixed.
- Controllers were selected on the same extrapolation grids reported here.

Original Fig. 3 and Sec. 5.3, PDF p. 10. Three-seed mean/SD; train at 1K. Fixed-FSA's roughly 42% score reflects incomplete rule learning.

A larger model reveals a clear failure regime



- 8 layers, width 1024, 32 heads; train at $D = 4$.
- S32/A4/1K: unseen-depth gains average -9.30, -6.39, -4.06 pp.
- S64/A8/2K: gains at all four depths; average $+7.14$ pp.

Original Fig. 4 and Sec. 5.4, PDF p. 11. Averages above use 16–64× training length. State count, action count, training length, and controller settings all change between regimes.

Real-text credit varies in both directions and across heads

Read-only BPTT probe, every 64 tokens

Median earliest/latest normalized norm ratio: 0.23.
At the final snapshot, 83.6% of sample–layer–head trajectories are weaker at the earliest boundary.

Adjacent absolute $\log_2$ change	Value	Multiplicative factor
Median	0.580	1.50
90th percentile	3.164	8.96

- This motivates gentle, symmetric, head-wise corrections.
- Probe normalization divides by $\sqrt{\text{remaining targets}}$; the controller uses the unnormalized signal.

Paper, Sec. 6.1, PDF pp. 11–12. Normalization is a heuristic for changing loss counts, not fixed-signal survival.

Symmetric CST shrinks deviations in log norm

For each layer and head, reduce over batch and within-head state dimensions:

$$n_k = \|\lambda_k\|_F, \qquad z_k = \log \max(n_k, \epsilon), \qquad d_k = z_k - m.$$

Controller

$$\gamma_k = \mathrm{clip}\Big(e^{-\alpha d_k}, c^{-1}, c\Big), \qquad \widetilde{\lambda}_k = \gamma_k \lambda_k$$

$$m \leftarrow \beta m + (1-\beta) z_k$$

- Without clipping: corrected log-norm deviation is $(1-\alpha)d_k$.
- $H = 64,\ \alpha = 0.10,\ \beta = 0.90,\ c = 1.10$; gains lie in $[0.9091, 1.10]$.
- Reference uses the current pre-intervention norm and resets each backward pass.

Paper, Sec. 6.1, Eqs. (16)–(17), PDF p. 12. Initialize $m = z_{K-1}$ with identity gain. Later-boundary corrections still create indirect feedback.

Language modeling tests 4K training through 128K inference

- **123.30M parameters**: one 32-token local-attention block, then 12 Gated DeltaNet blocks.
- **Training**: LongData, 4096-token sequences, 20K updates, 7.86B tokens, eight H100 GPUs.
- **Matched comparison**: same initialization and data seed; one final training run per method.

Evaluation	Domain	Cohort	Maximum horizon
LongData	In-domain	128 held-out windows	32K / 8×
Books3	Out-of-domain	128 book windows	128K / 32×
GovReport	Out-of-domain	50 documents	32K / 8×

Paper, Sec. 6.2, PDF p. 12; App. B.1, pp. 19–20. Controller selected among 18 settings using reported LongData-32K and Books3-128K endpoints.

Key tokens are labeled by an independent context test

Fixed external annotations

A token after 4K is selected when

$$\mathrm{NLL}_{\mathrm{short}} - \mathrm{NLL}_{\mathrm{full}} > 2.0 \quad \text{and} \quad \mathrm{NLL}_{\mathrm{full}} < 2.0.$$

- LongData / Books3: pinned Qwen2-72B-Instruct annotator.
- GovReport: released LongPPL labels.
- Across tokenizers, require full containment of the target's UTF-8 byte span.
- Masks are fixed before comparing BPTT and CST.

All-token and key-token NLL are token-weighted; lower is better.

At 4K there are no eligible key targets, so key NLL is undefined.

Paper, App. B.2, PDF p. 20; Table 4, p. 13.

Longest-context gains are larger on key tokens

$$\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{BPTT}} - \mathrm{NLL}_{\mathrm{CST}} \qquad \text{(positive favors CST)}$$

Dataset / length	All-token gain	Key-token gain	Key / all gain
LongData / 32K	0.006843	0.023932	3.50×
Books3 / 128K	0.005949	0.014770	2.48×
GovReport / 32K	0.009726	0.044182	4.54×

- All-token NLL improves in all 14 dataset–length settings.
- Mean all-token gain: 0.00724 NLL; overall improvement is modest.
- Example: Books3-128K all-token NLL is 3.524265 → 3.518316.

Paper, Tables 3–4 and Sec. 6.4, PDF p. 13. Gains use unrounded losses, then independent rounding. Ratios compare NLL reductions, not accuracies.

Key-token gains include two reported reversals

Dataset	Length	Key targets	Key NLL gain
LongData	16K	7,130	+0.025832
	32K	39,054	+0.023932
Books3	16K	2,836	−0.019084
	32K	11,074	+0.007481
	64K	34,101	+0.010558
	128K	95,993	+0.014770
GovReport	8K	277	−0.049017
	32K	3,147	+0.044182

Sparse short-context annotations give uncertain estimates.
Books3 results have no paired document confidence intervals.

Selected rows from paper Table 4, PDF p. 13; App. B.4, p. 21. Negative values are retained rather than assumed to be noise.

Context removal supports better distant use on LongData

Re-score the same key targets with full context and a clean-state 4K window.

$$\Delta_{\text{context}} = \underbrace{(\text{NLL}_{\text{short}} - \text{NLL}_{\text{full}})_{\text{CST}}}_{\text{CST context benefit}} - \underbrace{(\text{NLL}_{\text{short}} - \text{NLL}_{\text{full}})_{\text{BPTT}}}_{\text{BPTT context benefit}}$$

Setting	Context-use gain	95% paired document CI
LongData / 16K	+0.019479	[0.006917, 0.032245]
LongData / 32K	+0.014703	[0.009604, 0.020999]
GovReport / 32K	+0.002684	[−0.002305, 0.007471]

LongData excludes zero; GovReport does not.

These intervals describe document sampling, conditional on the selected runs.

Paper, Sec. 6.4, PDF p. 14; Table 5 and App. B.4, p. 21. 20,000 paired document bootstrap resamples; no removed-prefix state is reused.

The evidence supports a useful but selective intervention

- **Scale control has limits.** It cannot restore missing directions or resolve temporal interference.
- **Controller strength matters.** Only 5 of 18 language configurations improve both selection endpoints; $\alpha = 0.30$ performs worse on average.
- **Language evidence is exploratory.** One 123M architecture; one final run per method; reported endpoints also used for selection.
- **Transfer is conditional.** Larger-model unseen-depth failures and the unresolved GovReport control matter for deployment decisions.

Paper, Secs. 5.4, 6.2–6.4 and 8, PDF pp. 11–15; App. B, pp. 19–21.

Take home: diagnose credit demand before stabilizing scale

1. **Separate task need from signal decay.**
   Dense local supervision can train a shared recurrent rule.
2. **State credit is a direct intervention point.**
   CST modifies the reverse pass while retaining the forward model.
3. **Match the controller to observed dynamics.**
   Event-based contraction repair and symmetric head-wise smoothing serve different regimes.

A concrete next experiment

Compare one-sided vs. two-sided controllers with independent selection, multiple seeds, context-removal tests, and measured training throughput.

Next experiment is this reading's proposal. Primary source: Hanwen Jiang, Learning Length-Extrapolatable Recurrent Models, arXiv:2609.09157v1 (2026).

Backup: a capped event resets to the attainable norm

$$\text{Adjacent-CST:} \qquad \gamma_k = \operatorname{clip}\left(\frac{\widetilde{n}_{k+1}}{\max(n_k, \epsilon)}, 1, \gamma_{\max}\right)$$

Event-CST teaching example: $A = 10,\ \rho = 1.5,\ \gamma_{\max} = 2.$

Incoming n_k	Event?	Gain	Corrected norm	New anchor
8	No	1	8	10
6	Yes	10/6	10	10
3	Yes, capped	2	6	6

The anchor follows achieved restoration when the gain cap binds.
Invalid or near-zero norms keep gain 1 and leave the anchor unchanged.

Teaching example, not paper data. Controller equations: Sec. 5.1, Eqs. (12)–(14), PDF p. 7. Incoming norms are live observations at each boundary.

Backup: local scaling can change the summed direction

Teaching example

$$q = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \bar{J}^{\top} \lambda = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\text{BPTT: } \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \longrightarrow \quad \gamma = 2 : \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

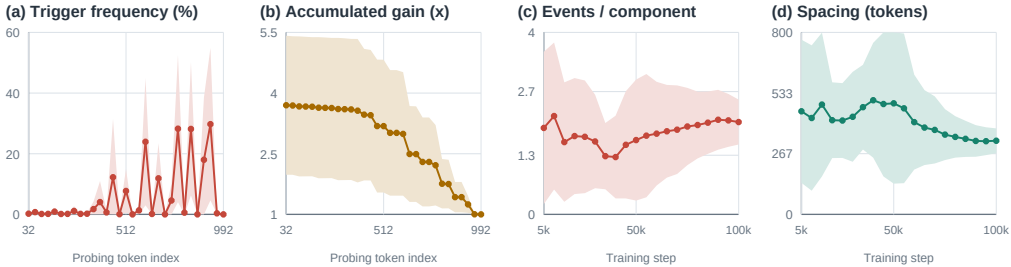
The scaled component keeps its direction; the sum changes direction.

Implementation

- Clone boundary state without detaching.
- A backward hook rescales its incoming credit.
- Event-CST: one gain per state tensor, batch included.
- Real text: one gain per layer and recurrent head.

Example is explanatory, not an experiment. Paper, Eq. (11), PDF p. 6; App. C.2, p. 22. References reset every backward pass.

Backup: Event-CST makes few, position-dependent corrections



- 5.8% of boundary–component decisions trigger: 1.79 events per state-component trajectory.
- Mean gain over all candidates: $1.053\times$.
- Panel (b) is the product of controller gains, about $3.7\times$ at early boundaries; it is not a CST/BPTT gradient-norm ratio.

Original Fig. 6, PDF p. 23; App. C.4, pp. 22–23. Panels (a,b): token position; panels (c,d): training step. Shading: sample SD across seeds.

Backup: sparse replay plans reduce boundary exposure



- Late-training plans retain 4, 8, and 8 boundaries across three seeds.
- Including periodic dense probes: 31 $\rightarrow$ 6.91 hooks per step.
- Positions are reused; current minibatch signals determine live gains.

Original Fig. 7 and App. C.5, PDF p. 23. Probe period $P = 100$; 77.7% fewer hooks. Throughput recovery remains a projection in this paper.

Backup: model and search budgets define the LM comparison

- Width 768; 16 recurrent heads of dimension 48; SwiGLU width 1792; tied GPT-2 embeddings, vocabulary 50,304.
- LongData: 458 training shards, 16 held-out shards; packed 4096-token sequences, EOT document separators.
- AdamW: peak LR 6×10^{-4}, 67-step warmup, cosine decay to 6×10^{-5}, gradient norm clip 1.

18 full-training configurations

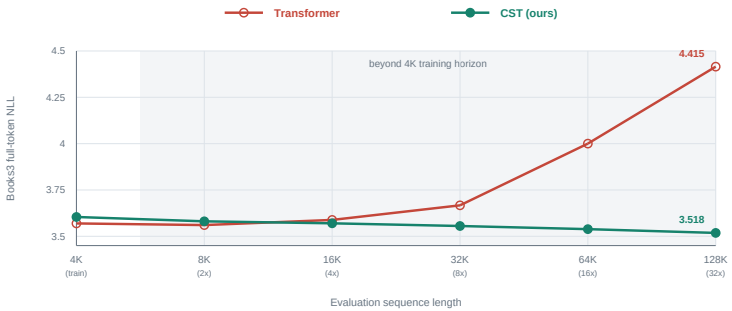
$$\alpha \in \{0.03, 0.10, 0.30\}, \quad \beta \in \{0.90, 0.97\}, \quad c \in \{1.10, 1.25, 1.50\}.$$

All trained for 20K updates; selection uses LongData-32K and Books3-128K.

Reported intervals condition on the selected configuration.

Paper, App. B.1, PDF pp. 19–20. AdamW betas $(0.9, 0.95)$, weight decay 0.1; global batch 96.

Backup: the Transformer comparison changes the architecture



- Matched data, roughly 123M parameters, 4K training, 7.86B tokens.
- Transformer uses dynamic YaRN beyond 4K; CST uses recurrent inference.
- Isolated CST gain at 128K is $3.524265 \rightarrow 3.518316$ against recurrent BPTT.

Original Fig. 5 and App. B.3, PDF p. 20; recurrent baseline from Table 3, p. 13. The 4.415 vs. 3.518 architecture gap is not the isolated effect of CST.

Backup: confidence intervals measure document uncertainty

- 20,000 paired document bootstrap resamples.
- Windows from the same document remain grouped.
- Each replicate recomputes total NLL / total eligible target count.
- No training-seed or hyperparameter-selection uncertainty is included.

LongData, only positions 16K–32K:

Quantity	Gain	95% CI
Key-token NLL	0.023537	[0.011159, 0.037012]
Context use	0.013666	[0.008616, 0.020309]

The later position bucket supports the cumulative-prefix result.

Paper, App. B.4, PDF p. 21. Books3 has no paired document intervals.

Backup: unsuccessful explorations suggest concrete bottlenecks

- **Length curriculum:** varying growth schedules did not yield consistent extrapolation gains.
- **State self-distillation:** useful privileged target states were difficult to define.
- **Quantized state:** simultaneously learning the quantizer, codebook, and recurrence created a coupled optimization problem.
- **Fully adaptive credit control:** removing a fixed trigger or gain cap increased sensitivity to noisy boundary statistics and added complexity.

These are outcomes of the authors' tested formulations, not impossibility results.

Paper, App. D, PDF pp. 23–24.