WHEN WEAK ATTENTION GOES SILENT: REPAIRING CONTEXT-AMPLIFIED ATTENTION DISTORTION IN SPIKING LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spiking neural networks (SNNs) offer sparse, event-driven inference for large language models (LLMs). Converting pretrained artificial neural networks (ANNs) to SNNs with few timesteps can preserve performance on mainstream language-modeling and question-answering benchmarks, yet we find that converted models increasingly fail to use relevant information as input length grows. Controlled interventions identify the few-spike encoding of post-Softmax attention as the dominant recoverable bottleneck: its zero-response region suppresses weak but collectively substantial attention probabilities. Uniformly rescaling the response to recover these probabilities compromises coverage of larger ones. We propose the Downstream-Calibrated Phase Neuron (DCPN) to repair both ends of the response at a fixed deployment budget. DCPN adjusts the final phase's threshold and readout to make weak probabilities accessible with appropriately small outputs, while coordinating earlier-phase readouts to preserve coverage of salient probabilities. It screens candidate responses by reconstructing attention-weighted value outputs, then confirms them using validation negative log-likelihood in the progressively converted model. DCPN changes only post-Softmax response parameters, preserving pretrained weights, timestep count, response-parameter storage, and few-spike inference structure. It improves conventional language-modeling and short-input task performance across evaluated settings. At eight timesteps, RULER and NeedleBench scores increase by 38.69 and 52.97 percentage points on LLaMA-2-7B at 4K, and by 24.44 and 38.07 points on LLaMA-3-8B at 8K, respectively, relative to the conversion baseline.

1 INTRODUCTION

Large language models (LLMs) perform well across language tasks (Touvron et al., 2023; Dubey et al., 2024), but their inference requires extensive dense computation. Inspired by biological neural signaling, spiking neural networks (SNNs) communicate through discrete spike events (Maass, 1997). Sparse spikes allow computation to be driven by activity, creating opportunities for lower energy use on suitable neuromorphic hardware (Yin et al., 2021; Schuman et al., 2022). These properties make SNNs a promising route to more efficient LLM inference, provided the capabilities of the original language model can be retained.

One practical route is conversion from artificial neural networks (ANNs) to SNNs: a pretrained model supplies the weights and target activations, while spiking neurons approximate its responses over a finite number of timesteps. This approach has been developed for Transformer architectures (Hwang et al., 2024; Huang et al., 2025; Wang et al., 2026) and extended to large language models (Zheng et al., 2026; Chen et al., 2026). Few-spike (FS) conversion uses spike timing as well as spike count to represent activations within short temporal windows (Stöckl & Maass, 2021). For LLMs, distribution-aware calibration further fits phase-neuron parameters to pretrained activation distributions without updating the model weights (Zheng et al., 2026). More timesteps can improve response fidelity but increase temporal computation, so the low-timestep regime requires careful control of conversion error.

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

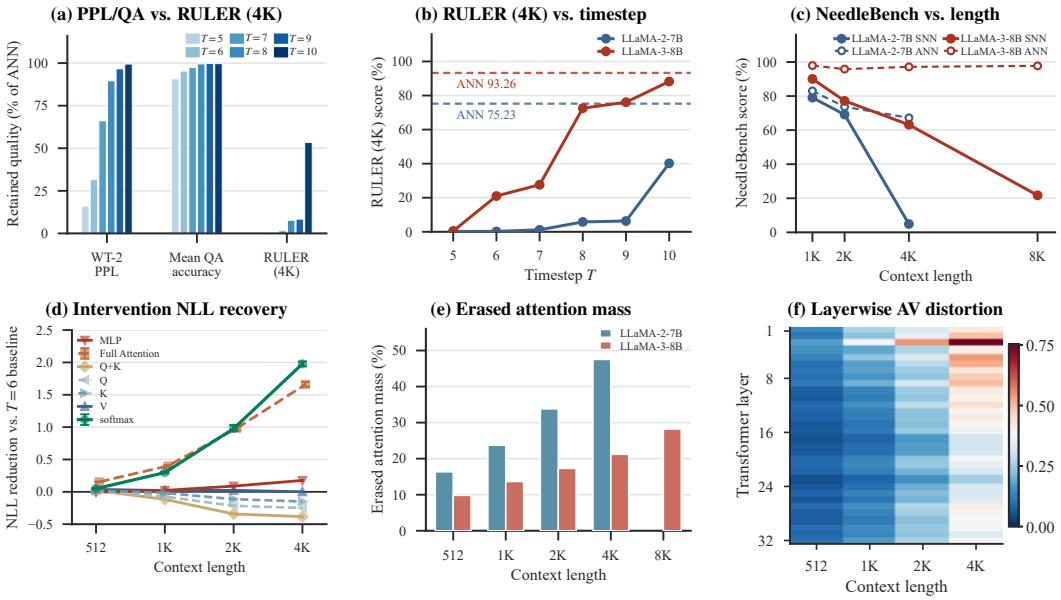

Figure 1: Diagnosing context-dependent failure in low-timestep conversion. (a) ANN-relative WikiText-2 (Merity et al., 2017) perplexity (PPL), four-task question-answering (QA) accuracy, and RULER at 4K across timesteps. (b) RULER at 4K versus timestep, with model-specific ANN ceilings. (c) NeedleBench (Li et al., 2024) scores versus context length at $T = 8$. Panels (b,c) probe temporal budget and input length separately. (d) Negative log-likelihood (NLL) recovery from component replacement relative to the LLaMA-2-7B $T = 6$ baseline; positive values indicate improvement. (e) Erased attention mass for both models. (f) Layerwise attention–value (AV) normalized mean squared error (NMSE) for LLaMA-2-7B at $T = 6$, measuring downstream output distortion rather than calibration effectiveness.

The challenge is that conventional benchmarks, dominated by short sequences (Appendix A.1), can overstate how well conversion preserves information use as input length grows. We study this low-timestep regime with the response-parameter count and FS inference structure fixed. In a representative setting, a converted model nearly matches its ANN reference on conventional tasks but scores only 5.80 versus 75.23 for the ANN on RULER (Hsieh et al., 2024) at 4K (Figure 1(a)). The deficit persists across timestep settings and worsens as input length grows (Figure 1(b,c)). Preserving this ability under a fixed deployment budget is a central conversion challenge.

We propose the Downstream-Calibrated Phase Neuron (DCPN) to repair this context-amplified loss of information. DCPN targets post-Softmax FS responses, which component-level interventions identify as the dominant recoverable bottleneck (Figure 1(d)). Their zero-response region erases weak attention probabilities that collectively carry substantial mass as context grows (Figure 1(e,f)). Recovering weak probabilities through a common scale adjustment also changes the response to salient probabilities, creating a *dual-endpoint conflict*. DCPN resolves this conflict by adjusting the final phase's threshold and readout to represent weak probabilities while coordinating earlier-phase readouts to preserve coverage of larger ones.

Beyond response construction, DCPN calibrates which candidate to deploy. Attention–value (AV) reconstruction screens phase responses by how well they preserve attention-weighted value outputs. Validation negative log-likelihood then confirms shortlisted candidates in the progressively converted model, with accepted updates committed sequentially. This model-level check matters because local reconstruction quality alone can mis-rank deployable responses.

Across evaluated models and timestep settings, DCPN improves information use as input length grows, along with conventional language-modeling and short-input task performance, while preserving the timestep count, response-parameter storage, and FS inference structure.

Our contributions are:

- **We identify post-Softmax weak-attention loss as the mechanism behind context-dependent conversion failure.** As input length grows, the FS zero-response region erases collectively substantial attention mass and amplifies attention–value distortion, even when conventional benchmarks remain close to the ANN reference.
- **We introduce DCPN to repair both ends of the post-Softmax response.** Its dual-endpoint phase response adjusts the final-phase threshold and readout to recover weak probabilities while coordinating earlier-phase readouts to preserve coverage of salient ones, without changing the timestep count, response-parameter storage, or FS inference structure.
- **We develop a two-stage, downstream-calibrated response-selection procedure.** Attention–value reconstruction screens feasible response candidates, and validation negative log-likelihood confirms shortlisted choices in the progressively converted model; accepted updates are committed sequentially.
- **We achieve consistent conventional-task gains and recover information use across native context windows.** DCPN improves perplexity and short-input accuracy across all evaluated model–timestep settings. At matched deployment budgets, its RULER and NeedleBench gains are 38.69 and 52.97 percentage points at 4K for LLaMA-2-7B, and 24.44 and 38.07 percentage points at 8K for LLaMA-3-8B, respectively.

2 Related Work

2.1 Few-Spike and Distribution-Aware Conversion

Few-spike conversion exploits spike timing as well as spike count to approximate activations with short temporal windows (Stöckl & Maass, 2021). Phase coding assigns timestep-dependent values to spikes; distribution-aware multi-granularity phase coding (DMPC) extends this representation with multiple learnable bases and decoupled threshold, reset, and readout parameters (Zheng et al., 2026). DMPC fits sampled hidden-layer activations without updating pretrained model weights. DCPN targets the phase-neuron response specifically at post-Softmax attention sites. Matched response-family and candidate-search comparisons isolate the effect of this design.

Transformer conversion also requires treating nonlinear computations explicitly. SpikedAttention develops a winner-oriented spike shift for Softmax (Hwang et al., 2024); expectation compensation handles Transformer nonlinearities (Huang et al., 2025); and multi-basis exponential decay supports nonlinear approximation in training-free conversion (Wang et al., 2026). loss-less ANN-SNN conversion (LAS) targets fully spike-driven LLMs (Chen et al., 2026). DCPN addresses the response of the post-Softmax FS neuron over both weak and salient attention inputs.

2.2 Operator-Aware Reconstruction

Quantization methods frequently replace elementwise error with a downstream or curvature-weighted objective. PTQ4ViT uses twin uniform quantization and Hessian-guided calibration (Yuan et al., 2022); PTQ4SAM evaluates attention quantization through value-weighted output (Lv et al., 2024); and aespa targets attention-wise reconstruction while retaining efficient layer-wise quantization (Kim et al., 2024). These works motivate measuring the computation induced by an approximation rather than the approximation alone. DCPN combines attention–value screening with model-level validation NLL to select deployable neurons.

2.3 Nonuniform Softmax Representation

Post-Softmax activations have motivated specialized quantization designs. FQ-ViT introduces Log-Int-Softmax for highly nonuniform attention maps (Lin et al., 2022). TSPTQ-ViT develops two-scaled quantization (Tai et al., 2023), AdaLog adapts the logarithmic base to activation statistics (Wu et al., 2024), and Softmax Bias Correction addresses systematic attention biases in quantized generative models (Pandey et al., 2023). These designs motivate our focus on weak-probability representation, output coverage, and aggregate distortion. DCPN addresses these considerations by calibrating the sequential FS neuron's triggering and readout while keeping its timestep count, parameter count, and inference structure unchanged.

3 PRELIMINARIES

Following DMPC (Zheng et al., 2026), we represent each converted activation with a few-spike (FS) phase neuron. For an input x and response configuration ϕ, spike generation and readout define

$$s_{1:T} = S_\phi(x), \qquad \widehat{x} = D_\phi(s_{1:T}) \equiv F_\phi(x). \tag{1}$$

For the nonnegative post-Softmax activations considered in this work, the neuron evolves as

$$u_1 = x + v_0, \quad s_t = \mathbf{1}[u_t > \theta_t], \quad u_{t+1} = u_t - h_t s_t, \quad F_\phi(x) = \sum_{t=1}^{T} d_t s_t. \tag{2}$$

where θ_t, h_t, and d_t denote the threshold, reset, and readout coefficients at phase t, respectively. Since $s_t \in \{0, 1\}$, a response naturally emits at most T spikes over its T phases.

With phase assignment $g_t \in \{1, \dots, G\}$, we follow the phase parameterization

$$d_t = \tau(b^d_{g_t})^{-t}, \qquad h_t = \tau(b^h_{g_t})^{-t}, \qquad \theta_t = \tau(b^\theta_{g_t})^{-t}. \tag{3}$$

The configuration ϕ therefore determines a finite response function F_ϕ through the sequential FS dynamics. DCPN modifies these response parameters only at post-Softmax FS sites; the pretrained weights and the underlying FS inference procedure remain unchanged.

4 DCPN: REPAIRING CONTEXT-AMPLIFIED POST-SOFTMAX FAILURE

DCPN translates the weak-attention diagnosis into a targeted phase-response design (Figure 2). We use the distinct roles of the early and final FS phases to repair weak-input responses and coordinate the representation of larger probabilities, then select responses through attention–value screening and model-level confirmation. The design preserves the deployment timestep count, response-parameter storage, and FS inference structure.

4.1 FROM WEAK-ATTENTION LOSS TO RESPONSE-SHAPE REPAIR

Preserving attention mass carried by weak probabilities requires two response changes: inputs inside the zero-response region must become able to trigger, and their first decoded output must match their small magnitude. We express these requirements through the end of the initial zero-response interval, β_0, and the first nonzero output, q_1. Lowering β_0 makes more weak inputs accessible, while calibrating q_1 controls how much mass each newly represented input contributes.

A common scale changes both properties together. In exact arithmetic, uniformly scaling the thresholds, resets, readouts, and initial potential by $r > 0$ gives

$$F_{\phi_r}(x) = rF_\phi(x/r). \tag{4}$$

Shrinking the scale moves triggering toward smaller inputs but also reduces the readout amplitudes used to represent larger probabilities. This motivates a response-shape adjustment that targets the weak end while coordinating high-end coverage. DCPN realizes this adjustment through the different roles of the existing FS phases.

4.2 TAIL-TARGETED REPAIR WITH COMPLEMENTARY COVERAGE

Phase roles provide a direction for repair. In the two-grain response ($G = 2$), the first $T - 1$ phases share a head coefficient group and the final phase has a tail group:

$$F_\phi(x) = \underbrace{\sum_{t=1}^{T-1} d_t s_t}_{\text{head}} + \underbrace{d_T \mathbf{1}[u_T > \theta_T]}_{\text{tail}}. \tag{5}$$

For a weak input that triggers none of the earlier phases, $u_T = x + v_0$. The tail threshold therefore directly controls whether this input is represented, and the tail readout controls its first decoded value. As the last decision in the sequence, this adjustment cannot change earlier spike decisions; moving the repair to an earlier phase can also alter the membrane state seen by later phases. Larger inputs can activate multiple head phases and accumulate their readout contributions. This difference gives the design a direct connection to the observed distortion: target weak-input accessibility and precision through the tail, and use the head contributions to support larger outputs.

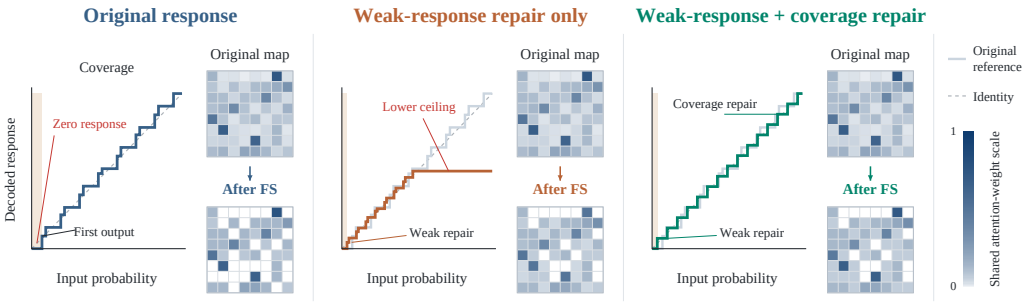

Figure 2: FS response design at a fixed timestep count and coefficient parameterization ($G = 2$). In this toy simulation, weak-response repair reduces zero responses but lowers the output ceiling; complementary coverage repair restores the high end. Maps share the same synthetic input and color scale, without post-encoding renormalization; white denotes zero.

Repairing the weak endpoint. DCPN uses the weak endpoint $(\beta^{\mathrm{ref}}, q^{\mathrm{ref}})$ of a higher-timestep response as a target for finer weak-input representation. We transfer these endpoint targets to the existing tail coefficients at deployment timestep T:

$$\theta_T^* = \beta^{\mathrm{ref}} + v_0, \qquad d_T^* = aq^{\mathrm{ref}}, \quad a > 0. \tag{6}$$

The threshold places the trigger near the reference weak boundary, while the factor a varies the first output around the reference value. These adjustments recover weak probabilities and calibrate their assigned mass. Appendix B.1 details parameter conversion and endpoint matching.

Coordinating coverage after weak repair. The tail also contributes to larger inputs whenever their residual potential $u_T = x + v_0 - \sum_{t=1}^{T-1} h_t s_t$ crosses its threshold. Changing the tail readout therefore changes these inputs' decoded values as well. DCPN complements weak repair by varying the overall response scale and, in joint candidates, adjusting the head readout to balance the change in the tail contribution. This coordinates the output available from earlier phases with the repaired weak endpoint, allowing the response to represent both weak and salient probabilities. The resulting design prioritizes weak-mass recovery and adds coverage correction to improve the fidelity of retained attention, matching the roles observed in the response diagnostics (Appendix C.1).

4.3 Downstream-Calibrated Response Selection

The construction yields candidates with different weak-end precision and high-end coverage. DCPN selects among them by first measuring their effect on attention aggregation, then evaluating how the shortlisted responses fit the current model configuration.

Screening by attention–value reconstruction. For attention A and its corresponding value tensor V, a candidate response changes the attention output by

$$\Delta O = (F_\phi(A) - A)V. \tag{7}$$

Using calibration pairs (A_i, V_i) from the current converted model with response configuration Φ, we score candidates at site l by the normalized aggregation error:

$$\mathcal{L}_{\mathrm{AV},l}(\phi; \Phi) = \frac{\sum_i \|(F_\phi(A_i) - A_i)V_i\|_F^2}{\sum_i \|A_i V_i\|_F^2}. \tag{8}$$

This score measures how well a candidate preserves the immediate computation performed with attention. Screening forms a small shortlist $\mathcal{P}_l$ from the weak-response, coverage, and joint construction families, together with the current response; the exact rule is given in Appendix B.2.

Confirming responses in the current model. Model-level confirmation selects a response in the presence of previously accepted updates. With only site l replaced by a shortlisted candidate, DCPN proposes

$$\phi_l^{\mathrm{prop}} = \arg\min_{\phi \in \mathcal{P}_l} \mathcal{L}_{\mathrm{val}}\left(M_{\Phi_{l \leftarrow \phi}}\right). \tag{9}$$

Table 1: RULER macro score (%, ↑) by context length. Mean averages lengths within each model. †LLaMA-2 1K uses the eligible subset; details in Appendix E.1.

		LLaMA-2-7B				LLaMA-3-8B				
T	Method	1K†	2K	4K	Mean	1K	2K	4K	8K	Mean
5	DMPC	14.34	3.14	0.21	5.90	6.07	1.17	0.47	0.12	1.96
	DCPN (ours)	**29.14**	**7.62**	**1.87**	**12.88**	**14.90**	**5.27**	**6.90**	**1.10**	**7.04**
6	DMPC	39.15	5.42	0.21	14.93	74.77	36.34	20.99	8.82	35.23
	DCPN (ours)	**72.68**	**45.91**	**4.85**	**41.15**	**91.12**	**79.78**	**58.49**	**29.34**	**64.68**
7	DMPC	69.86	26.43	1.17	32.49	89.91	71.06	27.63	9.25	49.46
	DCPN (ours)	**85.25**	**61.23**	**7.17**	**51.22**	**92.13**	**87.91**	**77.26**	**51.01**	**77.08**
8	DMPC	74.33	43.98	5.80	41.37	94.35	91.46	72.57	36.91	73.82
	DCPN (ours)	**90.21**	**80.21**	**44.49**	**71.64**	**94.46**	**93.22**	**89.40**	**61.35**	**84.61**
8	LAS	75.00	35.98	3.55	38.18	94.16	88.47	66.26	23.34	68.06
8	SpikedAttention	30.22	2.76	0.48	11.15	83.79	42.89	6.13	0.04	33.21
–	ANN	70.58	66.18	75.23	70.66	94.15	94.22	93.26	91.55	93.30

Candidate rankings motivate confirmation: local AV and model-level preferences can differ (Appendix B.3). Updates meeting the validation-improvement criterion are committed sequentially (Appendix D.1), and local pairs for subsequent sites are recollected from the updated model. Each decision combines attention-output fidelity with compatibility with the adapted model.

5 Experiments

5.1 Experimental Setup

We evaluate LLaMA-2-7B and LLaMA-3-8B (Touvron et al., 2023; Dubey et al., 2024) at $T \in \{5, 6, 7, 8\}$ against DMPC (Zheng et al., 2026), framework-matched LAS and SpikedAttention encoding controls (Chen et al., 2026; Hwang et al., 2024), and ANN references. RULER (Hsieh et al., 2024) and NeedleBench (Li et al., 2024) measure information use across the models' native context windows (4K and 8K, respectively), complemented by fit-window LongBench (Bai et al., 2024). Conventional evaluation reports mean perplexity over four corpora and mean accuracy over four downstream tasks. Paired DCPN–DMPC comparisons share pretrained weights, tokenization, evaluation examples, deployment timesteps, coefficient count and precision, and FS inference structure, with disjoint calibration, selection, and test data. Protocols and datasets are detailed in Appendix A.

Table 2: Native-window NeedleBench macro score (%, ↑) by context budget. Mean averages the reported lengths within each model.

		LLaMA-2-7B				LLaMA-3-8B				
T	Method	1K	2K	4K	Mean	1K	2K	4K	8K	Mean
5	DMPC	5.24	0.13	0.25	1.87	0.13	0.00	0.00	0.00	0.03
	DCPN (ours)	**24.39**	**0.40**	**0.76**	**8.52**	**1.96**	**1.55**	0.00	0.00	**0.88**
6	DMPC	48.66	1.06	0.13	16.62	62.35	25.04	1.29	0.76	22.36
	DCPN (ours)	**71.25**	**55.53**	**8.62**	**45.13**	**88.83**	**79.62**	**45.25**	**17.27**	**57.74**
7	DMPC	79.62	32.55	0.00	37.39	76.57	39.01	8.56	0.38	31.13
	DCPN (ours)	**85.27**	**83.39**	**23.55**	**64.07**	**93.45**	**85.04**	**71.29**	**36.24**	**71.51**
8	DMPC	79.10	69.14	4.89	51.04	90.08	77.19	63.18	21.74	63.05
	DCPN (ours)	**87.52**	**82.06**	**57.86**	**75.81**	**95.17**	**90.23**	**86.82**	**59.81**	**83.01**
8	LAS	88.35	50.76	1.10	46.74	87.50	74.09	42.20	13.02	54.20
8	SpikedAttention	30.27	0.60	0.00	10.29	73.26	9.17	0.57	0.00	20.75
–	ANN	83.01	73.75	67.22	74.66	97.92	95.83	97.16	97.73	97.16

5.2 Length-Resolved Capability Recovery

DCPN improves RULER performance over DMPC at all four deployment timesteps (Table 1). At $T = 8$, the 4K score rises from 5.80 to 44.49 on LLaMA-2-7B and from 72.57 to 89.40 on LLaMA-3-8B. Gains also extend across the shorter lengths and LLaMA-3's 8K window, showing that the response design improves information use across the evaluated contexts.

NeedleBench shows increasing recovery with context length at $T = 8$ (Table 2): gains grow from 8.42 points at 1K to 52.97 at 4K on LLaMA-2-7B, and from 5.09 at 1K to 38.07 at 8K on LLaMA-3-8B. Full results, including $T = 9, 10$ references, appear in Appendix E. On untruncated, model-specific fit-window LongBench subsets, DCPN improves the sample-weighted score in all eight model–timestep pairs (Appendix E.3).

5.3 Conventional Benchmark Performance

DCPN improves mean perplexity and accuracy in all eight model–timestep settings, with lower perplexity in all 32 individual corpus comparisons (Table 3). For LLaMA-2-7B at $T = 6$, mean perplexity falls from 18.84 to 6.74 and mean accuracy rises from 64.16% to 65.56%. These gains show that the long-context recovery is accompanied by improvements on conventional benchmarks; full per-corpus and per-task results appear in Appendix E.4.

Table 3: Mean perplexity and accuracy. Paired entries show DMPC / **DCPN** at the same T; means weight the four corpora or tasks equally.

		LLaMA-2-7B		**LLaMA-3-8B**	
T	Method	Mean PPL ↓	Mean Acc. (%) ↑	Mean PPL ↓	Mean Acc. (%) ↑
5	DMPC / DCPN	26.75 / **13.26**	61.07 / **63.18**	40.48 / **11.56**	48.14 / **59.47**
6	DMPC / DCPN	18.84 / **6.74**	64.16 / **65.56**	8.43 / **7.72**	69.29 / **69.70**
7	DMPC / DCPN	9.21 / **5.91**	65.62 / **66.80**	7.53 / **7.20**	70.36 / **70.55**
8	DMPC / DCPN	6.54 / **5.74**	67.00 / **67.28**	6.97 / **6.91**	71.42 / **71.61**
8	LAS	7.22	66.22	7.00	71.09
8	SpikedAttention	21.74	63.26	11.23	67.26
–	ANN	5.66	67.30	6.79	71.17

5.4 Ablation Studies

5.4.1 Response Construction and Phase Placement

The endpoint ablation links response design to recovery as context grows (Table 4). At $T = 8$, weak-response repair improves NeedleBench over DMPC by 4.92, 7.00, and 45.39 points at 1K, 2K, and 4K; adding coverage contributes another 3.50, 5.92, and 7.58 points. General search achieves the highest 1K score but falls to 2.10 at 4K, compared with 57.86 for dual endpoint, despite matching its 260 non-current candidates per layer. Thus, targeting weak attention is valuable as context grows. Gains hold under mean squared error (MSE) and AV screening (Appendix B.4).

Table 4: Endpoint ablation on LLaMA-2-7B at $T = 8$: NeedleBench performance, held-out AV NMSE, and WikiText-2 NLL/PPL. Protocol: Appendix B.5.

	NeedleBench (%, ↑)			Held-out / test (↓)		
Construction	1K	2K	4K	AV NMSE	NLL	PPL
DMPC (current)	79.10	69.14	4.89	.02059	1.80964	6.1082
Coverage only	73.84	73.19	10.45	.01921	1.74695	5.7371
Weak-response only	84.02	76.14	50.28	.00454	1.71416	5.5520
Dual endpoint	87.52	**82.06**	**57.86**	**.00443**	**1.70753**	**5.5153**
General search	**88.98**	68.43	2.10	.02058	1.78047	5.9326

Repair-phase placement. We compare front, middle, and tail repair at $T = 8$ with matched parameter counts, candidate counts, and selection procedures (Table 5). Tail repair achieves the lowest selected-response attention–value (AV) normalized mean squared error (NMSE), the highest NeedleBench score at every length, and a slightly lower WikiText-2 perplexity. This supports the final phase as the repair target under matched budgets.

Table 5: Repair-phase ablation on LLaMA-2-7B at $T = 8$ with matched budgets and selection. Local AV NMSE is measured during selection, unlike held-out AV NMSE in Table 4. Tail is DCPN.

Repair phase	WikiText-2 PPL ↓	Local AV NMSE ↓	NeedleBench (%, ↑)		
			1K	2K	4K
Front ($t = 1$)	5.5227	0.01307	84.30	79.87	54.68
Middle ($t = 4$)	5.5206	0.01448	84.15	79.26	53.26
Tail ($t = 8$, DCPN)	**5.5153**	**0.01243**	**87.52**	**82.06**	**57.86**

The length sweep clarifies this effect (Appendix C.1). From 512 tokens to 4K at $T = 6$, DMPC's erased mass rises from 16.347% to 47.521% and AV NMSE from 0.0739 to 0.3599. Weak repair restores most lost mass; the coverage adjustment further reduces AV distortion at every length.

5.4.2 Model-Level Confirmation

Model-level confirmation improves retrieval beyond local attention–value screening at both tested timestep budgets (Figure 3). Relative to local-only selection, forward confirmation raises NeedleBench scores by 3.50, 9.47, and 5.38 percentage points at 1K, 2K, and 4K for $T = 6$, and by 5.11, 7.76, and 12.15 points for $T = 8$. These paired comparisons support confirming shortlisted responses in the progressively converted model. The full $T = 6$ comparison with independent and reverse sequential selection, including perplexity, appears in Appendix D.2; ranking details appear in Appendix B.3.

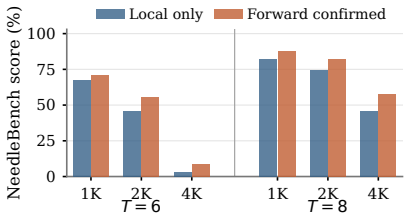

Figure 3: Model-level confirmation on LLaMA-2-7B.

5.4.3 Calibration-Length Transfer

We select responses using 511–4,088-token windows at $T = 8$ and evaluate at all listed lengths (Table 6). Across calibration windows, PPL varies by at most 0.0160 and stays below DMPC; every selection improves 2K and 4K NeedleBench. At 1K, only shorter calibration windows improve retrieval, with 511-token calibration achieving the highest score of 81.62. Exact target-length calibration is therefore unnecessary for gains at the longer tested lengths.

Table 6: Calibration-length transfer on LLaMA-2-7B at $T = 8$, evaluated across lengths. WikiText-2 PPL uses 16 test windows per length rather than the full main-result protocol, yielding different baseline PPL values. DMPC is the baseline; bold marks the best DCPN values.

	WikiText-2 PPL (↓)				NeedleBench (%, ↑)		
Calibration window	512	1K	2K	4K	1K	2K	4K
DMPC (baseline)	7.9151	6.8930	6.8023	9.4749	79.10	69.14	4.89
511	7.8392	6.7168	6.1513	6.0295	**81.62**	80.51	54.29
1,022	7.8450	6.7226	6.1442	6.0341	81.46	**80.66**	54.05
2,044	7.8498	6.7226	6.1511	6.0298	77.54	80.55	**58.43**
4,088	**7.8338**	**6.7099**	**6.1385**	**6.0200**	78.43	80.32	53.32

5.4.4 WEAK-ATTENTION RECOVERY AND SPIKE ACTIVITY

At $T = 8$, dual-endpoint repair nearly doubles post-Softmax events in paired LLaMA-2-7B checkpoints (Table 7), supporting weak-probability recovery alongside reduced erased mass (Appendix C.1). Whole-model few-spike (FS) events increase by only 1.19%, 1.33%, and 2.01% at 1K, 2K, and 4K, remaining below weak-only repair; estimated energy stays within 0.2% of DMPC. Counting details appear in Appendix F.1.

Table 7: FS events (billions, B) and theoretical energy on LLaMA-2-7B at $T = 8$. Energy changes are relative to DMPC.

		FS spike activity					Theoretical energy	
Length	Method	Post-Softmax	Other FS	All FS	Softmax share	vs. DMPC	J	Δ (%)
1K	DMPC	0.0505	3.8139	3.8644	1.31%	1.0000×	18.1531	–
	Weak only	0.0953	3.8196	3.9149	2.43%	1.0131×	18.1609	+0.043
	Dual endpoint	0.0955	3.8148	3.9103	2.44%	1.0119×	18.1467	−0.035
2K	DMPC	0.1048	7.6532	7.7580	1.35%	1.0000×	36.7229	–
	Weak only	0.2071	7.6589	7.8660	2.63%	1.0139×	36.7144	−0.023
	Dual endpoint	0.2075	7.6539	7.8615	2.64%	1.0133×	36.6976	−0.069
4K	DMPC	0.2383	15.2894	15.5276	1.53%	1.0000×	74.7585	–
	Weak only	0.4539	15.3898	15.8437	2.86%	1.0204×	74.8960	+0.184
	Dual endpoint	0.4548	15.3849	15.8398	2.87%	1.0201×	74.8797	+0.162

5.5 THEORETICAL ENERGY ANALYSIS

Following DMPC (Zheng et al., 2026), we estimate energy from multiply–accumulate (MAC) and spike-driven accumulate (AC) counts (Appendix F.2). DCPN differs from DMPC by -0.25% to $+1.03\%$ across evaluated lengths and reduces ANN energy by 45.97% at LLaMA-3 8K (Table 8).

Table 8: Theoretical inference energy (J, ↓); parentheses give percentage reductions relative to the matching ANN baseline.

		LLaMA-2-7B			**LLaMA-3-8B**			
T	Method	1K	2K	4K	1K	2K	4K	8K
5	DMPC	9.12 (71.2)	18.42 (71.5)	37.46 (72.2)	10.42 (71.0)	20.96 (71.4)	42.50 (71.9)	87.67 (72.9)
	DCPN	9.14 (71.1)	18.48 (71.4)	37.66 (72.0)	10.42 (71.0)	21.01 (71.3)	42.71 (71.8)	88.31 (72.7)
6	DMPC	11.43 (63.9)	23.05 (64.4)	46.76 (65.2)	14.19 (60.5)	28.55 (61.0)	57.69 (61.9)	118.04 (63.5)
	DCPN	11.40 (64.0)	23.12 (64.3)	47.24 (64.9)	14.19 (60.5)	28.59 (60.9)	57.94 (61.8)	118.72 (63.3)
7	DMPC	13.90 (56.1)	28.11 (56.5)	57.20 (57.5)	15.48 (56.9)	31.26 (57.3)	63.58 (58.0)	131.52 (59.3)
	DCPN	13.88 (56.2)	28.14 (56.5)	57.58 (57.2)	15.49 (56.9)	31.29 (57.3)	63.77 (57.9)	132.11 (59.1)
8	DMPC	18.15 (42.7)	36.72 (43.2)	74.76 (44.4)	20.74 (42.3)	41.80 (42.9)	84.83 (44.0)	174.25 (46.1)
	DCPN	18.15 (42.7)	36.70 (43.3)	74.88 (44.3)	20.75 (42.3)	41.82 (42.9)	84.92 (43.9)	174.65 (46.0)
8	LAS	18.15 (42.7)	36.71 (43.2)	74.66 (44.5)	20.74 (42.3)	41.80 (42.9)	84.79 (44.0)	174.09 (46.1)
8	SpikedAttention	11.40 (64.0)	23.04 (64.4)	46.90 (65.1)	13.02 (63.8)	26.25 (64.1)	53.36 (64.8)	110.11 (65.9)
–	ANN	31.66 (0.0)	64.68 (0.0)	134.52 (0.0)	35.95 (0.0)	73.20 (0.0)	151.48 (0.0)	323.23 (0.0)

6 CONCLUSION

Low-timestep ANN-to-SNN conversion can obscure context-length-dependent failure. DCPN combines weak-attention repair, salient-response coverage, local AV screening, and model-level confirmation to recover information use across two LLaMA models, with unchanged timesteps, response storage, and FS inference structure.

Limitations. We evaluate two LLaMA models within their native 4K and 8K windows. Future work will test larger models and longer contexts to assess DCPN's scalability and generality.

AI USE STATEMENT

AI tools assisted with identifying related literature and refining the discussion of experimental results. The authors take responsibility for the final content.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, et al. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.172. URL https://doi.org/10.18653/v1/2024.acl-long.172.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, et al. PIQA: reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7432–7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL https://doi.org/10.1609/aaai.v34i05.6239.

Long Chen, Xiaotian Song, and Yanan Sun. LAS: loss-less ANN-SNN conversion for fully spike-driven large language models. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1730–1738. AAAI Press, 2026. doi: 10.1609/AAAI.V40I3.37151. URL https://doi.org/10.1609/aaai.v40i3.37151.

Peter Clark, Isaac Cowhey, Oren Etzioni, et al. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL http://arxiv.org/abs/1803.05457.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Leo Gao, Stella Biderman, Sid Black, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021. URL https://arxiv.org/abs/2101.00027.

Mark Horowitz. Computing's energy problem (and what we can do about it). In *IEEE International Solid-State Circuits Conference*, pp. 10–14. IEEE, 2014. doi: 10.1109/ISSCC.2014.6757323.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, et al. RULER: what's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. doi: 10.48550/ARXIV.2404.06654. URL https://doi.org/10.48550/arXiv.2404.06654.

Zihan Huang, Xinyu Shi, Zecheng Hao, et al. Towards high-performance spiking transformers from ANN to SNN conversion. *arXiv preprint arXiv:2502.21193*, 2025. URL https://arxiv.org/abs/2502.21193.

Sangwoo Hwang, Seunghyun Lee, Dahoon Park, et al. SpikedAttention: Training-free and fully spike-driven transformer-to-SNN conversion with winner-oriented spike shift for Softmax operation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/file/7c9341ad0263428b5057d92f4d88dfa0-Paper-Conference.pdf.

Junhan Kim, Chungman Lee, Eulrang Cho, et al. Towards next-level post-training quantization of hyper-scale transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ab6eba9a853087993addff937c8cec87-Abstract-Conference.html.

Mo Li, Songyang Zhang, Yunxin Liu, et al. NeedleBench: Can LLMs do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*, 2024. doi: 10.48550/ARXIV.2407.11963. URL https://arxiv.org/abs/2407.11963v1.

Yang Lin, Tianyu Zhang, Peiqin Sun, et al. FQ-ViT: Post-training quantization for fully quantized vision transformer. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022. URL https://www.ijcai.org/proceedings/2022/0164.pdf.

Chengtao Lv, Hong Chen, Jinyang Guo, et al. PTQ4SAM: Post-training quantization for segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2405.03144.

Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. doi: 10.1016/S0893-6080(97)00011-7. URL https://doi.org/10.1016/S0893-6080(97)00011-7.

Stephen Merity, Caiming Xiong, James Bradbury, et al. Pointer sentinel mixture models. In *International Conference on Learning Representations*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Nilesh Prasad Pandey, Marios Fournarakis, Chirag Patel, et al. Softmax bias correction for quantized generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023. URL https://arxiv.org/abs/2309.01729.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, et al. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8732–8740. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6399. URL https://doi.org/10.1609/aaai.v34i05.6399.

Catherine D. Schuman, Shruti R. Kulkarni, Maryam Parsa, et al. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2(1):10–19, 2022. doi: 10.1038/S43588-021-00184-Y. URL https://doi.org/10.1038/s43588-021-00184-y.

Christoph Stöckl and Wolfgang Maass. Optimized spiking neurons can classify images with high accuracy through temporal coding with two spikes. *Nature Machine Intelligence*, 3(3): 230–238, 2021. doi: 10.1038/S42256-021-00311-4. URL https://doi.org/10.1038/s42256-021-00311-4.

Yu-Shan Tai, Ming-Guang Lin, and An-Yeu Wu. TSPTQ-ViT: Two-scaled post-training quantization for vision transformer. *arXiv preprint arXiv:2305.12901*, 2023. URL https://arxiv.org/abs/2305.12901.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Jingya Wang, Xin Deng, Wenjie Wei, et al. Training-free ANN-to-SNN conversion for high-performance spiking transformers. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2128–2136. AAAI Press, 2026. doi: 10.1609/AAAI.V40I3.37195. URL https://doi.org/10.1609/aaai.v40i3.37195.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, et al. RedPajama: an open dataset for training large language models. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d34497330b1fd6530f7afd86d0df9f76-Abstract-Datasets_and_Benchmarks_Track.html.

Zhuguanyu Wu, Jiaxin Chen, Hanwen Zhong, et al. AdaLog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2407.12951.

Bojian Yin, Federico Corradi, and Sander M. Bohté. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence*, 3(10): 905–913, 2021. doi: 10.1038/S42256-021-00397-W. URL https://doi.org/10.1038/s42256-021-00397-w.

Zhihang Yuan, Chenhao Xue, Yiqi Chen, et al. PTQ4ViT: Post-training quantization for vision transformers with twin uniform quantization. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2111.12293.

Hanyuan Zheng, Haozhen Zhang, Tianshuo Chen, et al. Distribution-aware multi-granularity phase coding: Towards lower conversion error for spike-driven large language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/81858886d6a6f22c2e555e9d8f5172d9-Paper-Conference.pdf.

A Full Experimental Protocols

The complete experimental setup below records example counts, window pairing, statistical uncertainty, and matched deployment and calibration settings. The principal text uses the abbreviated description in Section 5.1.

A.1 Models and Conventional Evaluation

The hardware and principal settings are summarized in Table 9; experiment-specific settings are stated in the corresponding protocols.

Table 9: Experimental configuration. The lower group applies specifically to the controlled response-selection experiment.

Configuration	Setting
Hardware	8 × NVIDIA L20 GPUs
Models	LLaMA-2-7B; LLaMA-3-8B
Deployment timesteps	$T \in \{5, 6, 7, 8\}$
Native context windows	4K tokens (LLaMA-2-7B); 8K tokens (LLaMA-3-8B)
Response parameterization	Two coefficient groups ($G = 2$): head and tail
Adaptation scope	Post-Softmax FS response coefficients
Pretrained weights	Fixed during response adaptation
Data separation	Disjoint calibration, model-level selection, and test sets
Paired comparisons	Same weights, tokenizer, inputs, timesteps, coefficient count and precision, and FS inference structure
Controlled response-selection experiment	
Model and timesteps	LLaMA-2-7B at $T = 8$
Execution precision	16-bit floating point (FP16)
Calibration / local holdout	16 / 8 windows
Sampled attention rows	64 per layer per window
Stored response coefficients	Eight scalars per post-Softmax site
WikiText-2 test	166 windows of 2,048 tokens; 339,802 predicted tokens

Models and primary tasks. The evaluation covers LLaMA-2-7B and LLaMA-3-8B (Touvron et al., 2023; Dubey et al., 2024). Language modeling is measured on WikiText-2 (Merity et al., 2017), C4 (Raffel et al., 2020), RedPajama (Weber et al., 2024), and Pile (Gao et al., 2021) using perplexity. ARC-Challenge and ARC-Easy (Clark et al., 2018), PIQA (Bisk et al., 2020), and WinoGrande (Sakaguchi et al., 2020) measure downstream accuracy. Information use as input length grows within the models' native windows is evaluated with RULER and native-window NeedleBench at 1K–4K for LLaMA-2-7B and 1K–8K for LLaMA-3-8B, plus complementary fit-window LongBench subsets. All length-dependent conclusions refer to these evaluated windows. Low-timestep configurations use $T \in \{5, 6, 7, 8\}$; higher-timestep models and the ANN provide performance references.

Table 10: Scored-sequence lengths for the four conventional accuracy tasks, in model-specific tokens. For each example, the reported length is the maximum across answer candidates using the same zero-shot prompt templates as evaluation. Special tokens are excluded. The four-task mean accuracy weights tasks equally, not by sample count.

		Mean / median / P95 / max (tokens)	
Task	n	LLaMA-2-7B	LLaMA-3-8B
PIQA	1,838	43.22 / 35 / 100 / 297	36.77 / 30.5 / 83 / 244
ARC-Easy	2,376	37.89 / 33 / 74 / 207	32.91 / 28 / 64 / 163
ARC-Challenge	1,172	43.88 / 39 / 82 / 221	38.36 / 34 / 72 / 175
WinoGrande	1,267	25.64 / 25 / 34 / 45	22.04 / 21 / 30 / 37

A.2 Native-Window Retrieval Protocols

RULER protocol. We evaluate the RULER task suite at total context lengths of 1K, 2K, and 4K for LLaMA-2-7B, and 1K, 2K, 4K, and 8K for LLaMA-3-8B (1K = 1,024 tokens). The suite contains 13 tasks: eight needle-in-a-haystack variants, variable tracking, common-word extraction, frequent-word extraction, and two question-answering tasks. Each model–length configuration contains 500 examples per task (6,500 in total), except the LLaMA-2 1K eligible subset: 11 tasks contain 500 examples each and the two QA tasks contain 128 each, yielding 5,756 untruncated examples. Scores are unweighted macro averages over the 13 task scores. The 4K evaluation is the formal paired protocol; the other lengths provide complementary length-sweep comparisons. Within each model and length, evaluation examples, generation settings, and scoring settings are fixed across configurations, and every configuration completes all eligible examples.

NeedleBench protocol. We evaluate fixed native-window NeedleBench cases at total input-plus-generation budgets of 1K, 2K, and 4K for LLaMA-2-7B and 1K, 2K, 4K, and 8K for LLaMA-3-8B. Each length contains 110 single-retrieval, 10 multi-retrieval, and 88 multi-reasoning cases. Task content, evidence placement, and scoring follow the benchmark, with identical untruncated inputs across compared configurations. We report an unweighted macro average over the three task groups. Comparisons are paired within a model and length; cross-model scores use only the shared 1K–4K lengths.

LongBench protocol. We complement RULER with TREC, TriviaQA, SAMSum, Qasper, HotpotQA, GovReport, passage retrieval, and LCC. We request up to 200 examples per task and retain an example only when the complete official prompt and generation budget fit the native context window; documents are never truncated. This yields 343 evaluated examples for LLaMA-2-7B at 4K and 765 for LLaMA-3-8B at 8K. Passage retrieval has no eligible example under either native window and is reported as $n = 0$ rather than assigned a score. To prevent very small fit-window subsets from receiving the same influence as tasks with hundreds of eligible examples, the aggregate is weighted by the effective task sample count: $S_{\mathrm{LB}} = \sum_j n_j s_j / \sum_j n_j$ over the seven non-empty tasks, where s_j is the official task score. The exact model-specific sample counts and all per-task scores are reported in Tables 30 and 31.

A.3 Diagnostic Evaluation Protocols

Length-wise intervention protocol. For LLaMA-2-7B at $T = 6$, we compare the native baseline, a post-Softmax-only $T' = 8$ response transplant, and a full-attention $T' = 8$ response transplant on identical non-overlapping WikiText-2 test windows within each length. The paired window counts at lengths 512, 1,024, 2,048, and 4,096 are 670, 334, 166, and 83, respectively. We define recovery as baseline minus intervention NLL. Uncertainty uses 20,000 paired circular moving-block bootstrap resamples with block size four. The 2K intervention reported separately in Table 18 uses its own window protocol and simultaneous intervals; likewise, this intervention study has different window counts from the weak-mass response sweep below. Their numerical estimates should be interpreted within their respective matched protocols.

Weak-mass length sweep. We evaluate the DMPC, weak-response, coverage, and dual-endpoint responses from the LLaMA-2-7B $T = 6$ construction setting on 1,055 non-overlapping held-out windows: 564, 281, 140, and 70 windows at lengths 512, 1,024, 2,048, and 4,096, respectively. Every window covers all 32 layers with 64 sampled attention rows per layer. Lost mass is the row-wise probability mass assigned to positive attention entries that the FS response maps to zero. Row-mass error measures the deviation of the reconstructed row sum from the Softmax unit-mass constraint, and AV NMSE measures the normalized perturbation after multiplication by the corresponding value tensor. Weak-response and dual-endpoint adaptation are compared using 10,000 window-level paired bootstrap resamples.

A.4 Data Separation, Matched Budgets, and Uncertainty

Data separation and matched configurations. Calibration samples support local response scoring, a disjoint selection set supports model-level confirmation and hyperparameter choice, and locked test sets measure final performance after the configuration is frozen. Within each comparison, tokenization, sequence construction, pretrained weights, weight preprocessing, activation precision,

and evaluation settings are fixed. We distinguish the released DMPC checkpoint, a DMPC response refitted under our data and precision settings, a general matched-candidate search, and DCPN. The general-search and DCPN comparisons use matched candidate counts and the same shortlist and model-level evaluation budgets, so differences in quality cannot be attributed to a larger search budget alone.

Deployment scope. Same-T DCPN and DMPC execute the same FS dynamics with the same coefficient count, storage precision, and operator structure. DCPN changes the response coefficients and their selection while preserving the deployed inference graph. Section 5.4.4 reports realized spike counts and theoretical energy.

Controlled response-selection protocol. The response-by-objective factorial experiments use LLaMA-2-7B at $T = 8$, with 16 calibration windows and eight local holdout windows. Each layer samples 64 attention rows per window. Candidate scores share the same initial deployment trajectory, and the calibration objective alone determines the per-layer selection in this controlled experiment. The model is evaluated on the complete WikiText-2 test protocol: 166 windows of length 2,048 and 339,802 predicted tokens. Parameter changes are confined to post-Softmax neurons. Both candidate families use the same FS dynamics, a two-grain ($G = 2$) coefficient parameterization, the same phase assignment, and eight stored scalars. Computation uses FP16 execution.

Statistical uncertainty. Paired NLL differences for the completed factorial study use a circular moving-block bootstrap with blocks of eight evaluation windows and 5,000 resamples. These intervals characterize window-sampling variability for the recorded evaluation; downstream accuracies are accompanied by paired example-level bootstrap intervals.

B Response Construction and Selection Details

B.1 Response Characterization and Construction Details

For response characterization, let $\mathcal{X}$ be a finite ordered nonnegative deployment input grid containing zero and supported values up to one. For responses with $F_\phi(0) = 0$ and at least one positive output on this grid, define

$$\begin{aligned} x_+(\phi) &= \min\{x \in \mathcal{X} : F_\phi(x) > 0\}, \\ \beta_0(\phi) &= \max\{x \in \mathcal{X} : x < x_+(\phi)\}, \\ q_1(\phi) &= F_\phi(x_+(\phi)), \qquad R_{\max}(\phi; \mathcal{X}) = \max_{x \in \mathcal{X}} F_\phi(x). \end{aligned} \tag{10}$$

For one attention row a, erased mass is

$$\mathcal{M}_{\text{erase}}(a; \phi) = \sum_j a_j \mathbf{1}[a_j > 0]\mathbf{1}[F_\phi(a_j) = 0]. \tag{11}$$

These quantities characterize realized responses; candidate screening and confirmation use the AV and validation-NLL objectives in Section 4.3.

For weak-response construction, the tail threshold is placed around $\beta^{\text{ref}} + v_0$, and the target readout is $q^* = aq^{\text{ref}}$. In exact arithmetic, the corresponding readout base is

$$b_{\text{tail}}^{d,*} = \left(\frac{\tau}{q^*}\right)^{1/T}. \tag{12}$$

The implementation searches neighboring representable bases and retains responses whose realized (β_0, q_1) best match the intended weak endpoint. For coverage construction, threshold, reset, readout, and explicit initial-potential scales follow $\tau' = r\tau$. Joint candidates may additionally adjust the head readout to preserve the original coefficient-sum envelope. Because this envelope need not be attainable under the sequential threshold and reset dynamics, coverage is checked from realized outputs. Parameter settings that produce identical mappings are removed.

The scale identity in Equation 4 follows directly from the FS recurrence in exact arithmetic. With the same T, the uniformly scaled neuron at input x has potential $u_t^{(r)}(x) = ru_t(x/r)$ and the same spike decisions as the original neuron at input x/r, by induction over phases. Scaling the readout coefficients then gives $F_{\phi_r}(x) = rF_\phi(x/r)$. Finite-precision effects are handled by inspecting the realized mappings.

B.2 Candidate Pool and Shortlist

At each post-Softmax site l, the pool contains weak-response, coverage, and joint candidates, together with the current response. Each construction reuses the existing phase coefficients at fixed T, G, and stored parameter count. Let $\mathcal{C}_l^f$ denote the candidate pool from construction family f, and let Best_m retain the m distinct realized responses with the lowest AV loss. The shortlist is

$$\mathcal{P}_l = \{\phi_l^{\mathrm{current}}\} \cup \bigcup_{f\in\{\mathrm{weak,cov,dual}\}} \mathrm{Best}_m\left(\mathcal{C}_l^f; \mathcal{L}_{\mathrm{AV},l}\right). \tag{13}$$

Retaining candidates from each family preserves alternatives from the different construction directions while forming a compact shortlist. Including the current response permits model-level confirmation to retain it when no proposal meets the acceptance criterion in Appendix D.1.

B.3 Candidate Ranking and Shortlist Quality

To assess how reliably local objectives order candidates by model quality, we compare elementwise and attention–value reconstruction:

$$\begin{aligned} \mathcal{L}_{\mathrm{MSE}}(\phi) &= \mathbb{E}\|F_\phi(A) - A\|_F^2, \\ \mathcal{L}_{\mathrm{AV}}(\phi) &= \mathbb{E}\|(F_\phi(A) - A)V\|_F^2. \end{aligned} \tag{14}$$

Every candidate receives both local scores and a validation NLL under the same surrounding configuration. We report rank association, top-m recall of the model-best response, and selection regret. These measurements quantify whether a local proxy can reliably preserve the model-level ordering needed for deployment.

Table 11 shows only moderate local-to-model rank association at $T = 8$. For Dual/AV, Spearman correlation is $\rho = .356$ and R@2 is .125. Confirming the two locally shortlisted candidates lowers mean validation-NLL regret from .00639 to .00375, although the shortlist often excludes the model-best response. These measurements diagnose candidate selection within a fixed surrounding model; matched test-set deployment results are reported separately in Table 13.

Table 11: Candidate-level local-to-global ranking on held-out layers of LLaMA-2-7B at $T = 8$. ρ is Spearman rank correlation between the indicated local proxy and model validation NLL; R@2 is the fraction of layers for which the model-best candidate survives a two-candidate local shortlist. Regret is validation-NLL regret relative to the model-best candidate; "Conf." evaluates the two shortlisted candidates in the deployed model. Lower regret is better.

Family/proxy	ρ	R@2	Local regret	Conf. regret
Dual/AV	.356	.125	.00639	.00375
Dual/MSE	.312	.031	.00346	.00284
Fixed/AV	.483	.094	.00523	.00422
Fixed/MSE	.416	.094	.00669	.00648

The factorial experiment provides a complementary example for activation MSE: dual-endpoint selection has slightly higher local activation NMSE than fixed-response selection, 0.001042 versus 0.001022, while achieving lower test perplexity, 5.606667 versus 5.828505. This MSE example and the AV candidate-ranking measurements support checking selected responses in the model; neither establishes that AV is universally superior to MSE.

B.4 Response Construction across Local Objectives

For post-Softmax responses, attention–value reconstruction is closer to the immediate computation than elementwise activation error, but it remains a local proxy. In general,

$$\mathcal{L}_{\mathrm{AV}}(\phi_1) < \mathcal{L}_{\mathrm{AV}}(\phi_2) \not\Rightarrow \mathcal{L}_{\mathrm{val}}(M_{\Phi_{l\leftarrow\phi_1}}) < \mathcal{L}_{\mathrm{val}}(M_{\Phi_{l\leftarrow\phi_2}}). \tag{15}$$

We test this mismatch directly among realizable FS candidates at the same timestep count and parameterization, measuring candidate-level rank association, top-candidate quality, and shortlist recall.

The response-by-objective factorial experiment crosses fixed and dual-endpoint response families with MSE and AV calibration on LLaMA-2-7B at $T = 8$. Each cell evaluates 45 candidates: the fixed-response family searches 45 scales, while the dual-endpoint family searches 15 scales and three tail factors. The families share the original 15 fixed-response candidates. All layers use the same initial model trajectory, so the experiment isolates local selection and response construction before introducing sequential confirmation.

Table 12: Response construction $\times$ calibration objective. WikiText-2 perplexity (↓) on LLaMA-2-7B at $T = 8$; bold marks the lower value in each column.

Response	MSE	AV
Fixed	5.829	5.902
Dual-endpoint	**5.607**	**5.688**

Dual-endpoint construction reduces perplexity by 3.81% under MSE and 3.62% under AV. The corresponding paired NLL differences are -0.038804 with interval $[-0.041267, -0.036513]$ and -0.036868 with interval $[-0.040716, -0.033379]$. Thus the response construction improves model quality under both local objectives in this setting. AV screening does not outperform MSE in this experiment.

The shared-grid comparison gives a second view of the construction effect. Relative to the original 15-scale MSE reference, dual-endpoint responses improve perplexity from 5.788838 to 5.606667 at the same $T = 8$.

B.5 Extended Endpoint Ablation

Table 4 reports the $T = 8$ construction comparison on LLaMA-2-7B, with all metrics in each row evaluated from the same selected checkpoint. NeedleBench uses the unweighted three-task macro score over the same 208 fixed cases per length. AV NMSE uses eight held-out 2K windows and 64 sampled attention rows per layer and window; WikiText-2 test NLL and PPL use all 166 windows of length 2,048 (339,802 predicted tokens). The dual-endpoint row uses the same checkpoint as the $T = 8$ forward-confirmed result in Figure 3. Joint construction improves PPL over weak-response only by 0.0367 and raises the 4K NeedleBench score by 7.58 points. These results support the second endpoint as a smaller but consistent addition to weak-response repair in this protocol.

The candidate-ranking and matched-confirmation analyses examine whether local reconstruction identifies responses that improve model quality in the deployed model.

B.6 Matched Model-Level Confirmation

Table 13: Matched local-only versus model-confirmed selection on LLaMA-2-7B at $T = 8$ using the complete WikiText-2 test protocol (perplexity, ↓). Confirmation uses the same full-model evaluation budget within each comparison. Bold marks the better test result, but confirmation decisions use validation NLL rather than test perplexity.

Family/proxy	Local only	Model confirmed
Dual/AV	5.6523	**5.6325**
Dual/MSE	**5.6151**	5.6171
Fixed/AV	5.7841	**5.7544**
Fixed/MSE	5.8186	**5.7866**

The matched deployment experiment in Table 13 is separate from the one-shot factorial comparison in Table 12: candidates are screened and confirmed under a matched layerwise selection trajectory, so the absolute perplexities need not equal the factorial cells. For Dual/AV, confirmation lowers perplexity from 5.6523 to 5.6325, a 0.35% reduction. Fixed/AV and Fixed/MSE also improve, while Dual/MSE slightly worsens from 5.6151 to 5.6171 despite lower mean validation regret at shortlist size two. Model-level confirmation therefore provides a useful selection check in these experiments, without guaranteeing an improvement on the locked test set.

Together, the factorial and candidate-ranking results distinguish response construction from response selection. Dual-endpoint construction improves model quality under both local objectives, while local AV scores only partially preserve model-level candidate rankings. These observations support the model-level confirmation stage in Section 4 as a complement to local screening, with the response design providing the central correction to weak-attention distortion.

C Bottleneck Localization and Response Diagnostics

The following analyses cover baseline bottleneck localization, response diagnostics, and residual-error interventions after adaptation. Each comparison uses its stated checkpoint and evaluation protocol, with separate window sets where specified.

C.1 Weak-Mass Erasure and AV Distortion

We next test whether the diagnosed post-Softmax failure becomes more severe as context grows. Table 15 reports the complete length sweep. DMPC degrades monotonically on all three metrics: from 512 to 4,096 tokens, erased probability mass increases from 16.35% to 47.52%, row-mass error from 0.1433 to 0.4499, and AV NMSE from 0.0739 to 0.3599. Thus, the fixed response loses not only a growing number of small entries but an increasing fraction of the total attention mass, and the error is amplified after value aggregation.

Table 14: Length-wise response ablation on LLaMA-2-7B at $T = 6$: erased attention mass and AV distortion (↓).

	Lost mass (%)				AV NMSE			
Construction	512	1K	2K	4K	512	1K	2K	4K
DMPC	16.347	23.746	33.800	47.521	0.0739	0.1230	0.2218	0.3599
Coverage only	11.425	17.961	26.886	40.194	0.0520	0.0799	0.1405	0.2406
Weak-response only	6.553	11.445	18.376	29.028	0.0196	0.0252	0.0369	0.0594
Dual endpoint	**6.531**	**11.414**	**18.341**	**28.989**	**0.0163**	**0.0211**	**0.0324**	**0.0561**

Table 15: Full held-out weak-mass length sweep for the LLaMA-2-7B $T = 6$ construction setting (↓). Results aggregate 1,055 non-overlapping windows; every window covers 32 layers and 64 attention rows per layer. Bold marks the best response at each length.

Length	Response	Lost mass (%)	Row-mass error	AV NMSE
512	DMPC	16.347	0.1433	0.0739
	Weak-response	6.553	0.0480	0.0196
	Coverage	11.425	0.0995	0.0520
	Dual-endpoint	**6.531**	**0.0369**	**0.0163**
1,024	DMPC	23.746	0.2150	0.1230
	Weak-response	11.445	0.0649	0.0252
	Coverage	17.961	0.1597	0.0799
	Dual-endpoint	**11.414**	**0.0548**	**0.0211**
2,048	DMPC	33.800	0.3122	0.2218
	Weak-response	18.376	0.0978	0.0369
	Coverage	26.886	0.2433	0.1405
	Dual-endpoint	**18.341**	**0.0896**	**0.0324**
4,096	DMPC	47.521	0.4499	0.3599
	Weak-response	29.028	0.1537	0.0594
	Coverage	40.194	0.3728	0.2406
	Dual-endpoint	**28.989**	**0.1497**	**0.0561**

The ablations separate the roles of the two construction directions. Weak-response adaptation provides most of the recovered probability mass: at 4K it reduces lost mass by 18.49 percentage points and AV NMSE by 83.5% relative to DMPC. Coverage adaptation recovers less weak mass, but improves reconstruction of salient attention; across the four lengths its salient-band NMSE remains approximately 0.0055–0.0056, compared with 0.0146–0.0149 for DMPC. Dual-endpoint adaptation combines these roles and attains the lowest row-mass error and AV NMSE at every length.

The paired analysis in Table 16 clarifies that the additional dual-endpoint benefit is not caused by producing substantially more nonzeros. Relative to weak-response adaptation, dual-endpoint adaptation recovers only 0.022–0.038 additional percentage points of mass. Nevertheless, it consistently reduces row-mass error and AV NMSE, with every paired interval excluding zero. The extra gain therefore comes primarily from reconstructing retained nonzero attention more faithfully and protecting the salient region, rather than from changing which weak entries cross the zero-response boundary.

These results support a deliberately asymmetric interpretation of the dual-endpoint construction. Context growth primarily amplifies weak-mass loss, so weak-response adaptation accounts for most of the recovery. Coverage adaptation supplies a smaller but statistically consistent complementary correction to the fidelity of retained attention, which dual-endpoint adaptation inherits without sacrificing weak-response accessibility.

Table 16: Window-level paired bootstrap differences, Δ = Dual-endpoint − Weak-response, using 10,000 resamples. Negative values favor dual-endpoint adaptation. Lost-mass differences are percentage points; brackets are paired 95% confidence intervals.

Length	Δ Lost mass	Δ Row-mass error	Δ AV NMSE
512	-0.02164 $[-0.02184, -0.02144]$	-0.01108 $[-0.01115, -0.01101]$	-0.00327 $[-0.00343, -0.00311]$
1,024	-0.03111 $[-0.03145, -0.03077]$	-0.01008 $[-0.01019, -0.00998]$	-0.00415 $[-0.00434, -0.00396]$
2,048	-0.03446 $[-0.03507, -0.03387]$	-0.00824 $[-0.00841, -0.00807]$	-0.00451 $[-0.00469, -0.00433]$
4,096	-0.03848 $[-0.03937, -0.03762]$	-0.00390 $[-0.00416, -0.00366]$	-0.00324 $[-0.00343, -0.00305]$

C.2 Component Interventions Across Input Lengths

We next localize the conversion error exposed by the length-resolved evaluations. In this controlled localization experiment, the baseline is LLaMA-2-7B at $T = 6$, and selected module families are replaced by their $T' = 8$ responses while the rest of the deployed model is held fixed. Evaluation uses the complete WikiText-2 protocol: 166 windows and 339,802 predicted tokens. We define

$$\Delta\mathcal{L}(\mathcal{S}) = \mathcal{L}(M_{T;\mathcal{S}\leftarrow T'}) - \mathcal{L}(M_T), \tag{16}$$

so negative values indicate recovery. Equivalently, the recovery magnitude used in the diagnostic discussion is $\mathcal{R}(\mathcal{S}) = -\Delta\mathcal{L}(\mathcal{S})$.

The intervention establishes a sharply concentrated bottleneck in this setting. Softmax replacement alone accounts for 99.6% of the NLL recovery obtained by replacing the full attention stack ($0.9802/0.9846$). By contrast, replacing V alone yields only a small improvement, replacing the MLP gives a modest improvement, and both isolated Q and K replacements significantly worsen the model. The Q-side transplant is more harmful ($+0.2122$ NLL) than the K-side transplant ($+0.1072$ NLL); their joint harm ($+0.3359$ NLL) is close to the sum of the two isolated effects. Thus, the adverse Q+K result is not attributable to only one side of the bilinear interaction. The incremental interventions lead to the same conclusion. Starting from the V intervention, additionally replacing Softmax changes NLL by -1.0105 with paired 95% CI $[-1.0404, -0.9806]$. Starting from Softmax, additionally replacing V yields a much smaller but still significant -0.0590 change with CI $[-0.0622, -0.0559]$. Finally, full-attention replacement is not detectably better than Softmax alone: their paired NLL difference has 95% CI $[-0.0089, +0.0003]$. Softmax + V achieves the best perplexity, 6.1318, even outperforming the full-attention transplant.

These results support the statement that post-Softmax encoding is the dominant low-timestep bottleneck at 2K for *LLaMA-2-7B at* $T = 6 \to 8$ *on WikiText-2*. We next evaluate how this localization changes with attention length; the claim is confined to the studied intervention setting.

Length-wise localization. Table 17 repeats the matched Softmax-only and full-attention comparison at four attention lengths. At 512 tokens, Softmax-only replacement accounts for 31.7% of full-attention recovery; the fraction rises to 74.3% at 1K and 99.5% at 2K. At 4K, Softmax-only recovery is 1.9781 NLL, greater than the 1.6604 from full-attention replacement. The ratio of 119.1% is a relative comparison of two positive recovery values, not a recovery beyond a theoretical limit.

Table 17: Length-wise paired intervention on LLaMA-2-7B at $T = 6 \to 8$ on WikiText-2. All three configurations use identical windows within each length. PPL is lower-is-better; recovery is baseline NLL minus intervention NLL. Ratio is Softmax-only recovery divided by full-attention recovery.

Length	Windows	Base PPL	Softmax PPL	Full PPL	Softmax rec.	Full rec.
512	670	8.758	8.341	**7.507**	0.0488	0.1541
1,024	334	9.739	7.234	**6.528**	0.2974	0.4001
2,048	166	17.246	6.491	**6.462**	0.9772	0.9817
4,096	83	53.765	**7.437**	10.219	**1.9781**	1.6604

The paired direct contrasts in Appendix Table 19 show a change in the relative effectiveness of the two interventions: full attention is better at 512 and 1K, the difference is small at 2K, and Softmax-only is substantially better at 4K (contrast $+0.3177$, 95% CI $[+0.2979, +0.3383]$). Thus, additional attention-component replacements have a negative *net* effect relative to Softmax-only replacement at 4K. The separate 2K module interventions show that both Q and K teacher transplants are harmful, with a larger Q-side effect; they establish a sensitive bilinear subpath but do not by themselves attribute the full 4K net difference among all additional attention components. The length-wise evidence motivates concentrating construction on post-Softmax responses while keeping claims about other model families and tasks open.

Table 18: Controlled module intervention for LLaMA-2-7B on WikiText-2, replacing selected $T = 6$ responses with $T' = 8$ references. ΔNLL is relative to the unchanged $T = 6$ baseline; negative values indicate improvement. Confidence intervals are paired 95% simultaneous intervals over the full 166-window evaluation.

Intervention	PPL ↓	ΔNLL ↓	95% simultaneous CI
Baseline	17.3354	0	–
Softmax	6.5045	-0.9802	$[-1.0229, -0.9376]$
V	16.8431	-0.0288	$[-0.0407, -0.0169]$
Q-only	21.4327	+0.2122	$[+0.2022, +0.2222]$
K-only	19.2961	+0.1072	$[+0.0989, +0.1154]$
Q+K	24.2557	+0.3359	$[+0.3209, +0.3509]$
MLP	15.8810	-0.0876	$[-0.1001, -0.0752]$
Softmax + V	**6.1318**	**-1.0393**	$[-1.0843, -0.9942]$
Full attention	6.4765	-0.9846	$[-1.0283, -0.9409]$

Table 19: Paired 95% intervals for the length-wise intervention using 20,000 circular moving-block bootstrap resamples (block size four). Recovery is baseline minus intervention NLL. The direct contrast is $\mathrm{NLL}_{\mathrm{full}} - \mathrm{NLL}_{\mathrm{Softmax}}$; positive values favor Softmax-only. Ratio intervals are computed from paired resamples.

Length	Softmax recovery [95% CI]	Full-attention recovery [95% CI]	Softmax/Full [95% CI]	Direct contrast [95% CI]
512	$0.0488\,[0.0457, 0.0520]$	$0.1541\,[0.1494, 0.1589]$	$31.7\%\,[30.0\%, 33.3\%]$	$-0.1053\,[-0.1091, -0.1016]$
1,024	$0.2974\,[0.2840, 0.3108]$	$0.4001\,[0.3846, 0.4158]$	$74.3\%\,[73.4\%, 75.2\%]$	$-0.1028\,[-0.1072, -0.0984]$
2,048	$0.9772\,[0.9428, 1.0115]$	$0.9817\,[0.9471, 1.0162]$	$99.5\%\,[99.1\%, 100.0\%]$	$-0.0045\,[-0.0090, -0.0002]$
4,096	$1.9781\,[1.9415, 2.0152]$	$1.6604\,[1.6251, 1.6981]$	$119.1\%\,[117.8\%, 120.6\%]$	$+0.3177\,[+0.2979, +0.3383]$

C.3 Cross-Model and Streamed Diagnostics

Residual-error interventions after adaptation. For LLaMA-3-8B, Table 20 starts from the final adapted DCPN checkpoint at $T = 6$ and replaces either its post-Softmax responses or its full-attention responses with those from the native $T = 8$ donor checkpoint. All other responses remain at the DCPN baseline configuration. Evaluation uses full-window forward passes on identical non-overlapping WikiText-2 windows within each length. Softmax-only replacement improves NLL throughout 512–4K, with recovery rising from .0097 to .0207, while full-attention replacement produces larger recovery at every evaluated length. These interventions probe the encoding error remaining after DCPN adaptation: further post-Softmax correction is beneficial, but does not account for all recoverable attention error in the adapted model. They do not localize the error sources of the original DMPC checkpoint.

Separate 8K streamed comparison. Table 21 reports the 8K comparison under its streamed evaluation protocol. Softmax-only replacement improves NLL, while full-attention replacement recovers an additional 0.0380 NLL. These within-protocol contrasts are reported separately from the full-window 512–4K intervention results.

Table 22 independently confirms context-amplified degradation under the same LLaMA-3 $T = 8$ setting used by the primary DCPN results: lost mass rises from 11.42% at 1K to 25.09% at 8K, while AV NMSE increases from .0184 to .0760. These diagnostics use a separate streamed evaluation protocol and are presented as corroborating evidence rather than as a replacement for the paired LLaMA-2 intervention analysis.

Original DMPC versus final DCPN. We additionally compare the original DMPC checkpoint with the final model-confirmed DCPN checkpoint on LLaMA-3-8B at $T = 6$, using the same WikiText-2 windows and sampled attention rows. Table 23 shows that DCPN lowers perplexity, erased mass, row-mass error, and attention–value distortion at every length. The absolute improvements grow with context: at 8K, DCPN reduces erased mass by 17.37 percentage points and AV NMSE by .2112 relative to raw DMPC. This paired comparison measures the overall effect of adaptation and does not isolate individual error sources or endpoint contributions. Its DCPN checkpoint is identical to the intervention baseline in Table 20; the small PPL differences between the two tables reflect full-window versus streamed evaluation.

For the response-level diagnosis, we report the fraction of positive attention entries encoded as zero, the probability mass carried by those entries, reconstruction error over fixed weak-input and high-amplitude regions, row-mass error, and the full value-weighted perturbation. These measurements quantify the dominant weak-input erasure and test whether the coverage adjustment in Section 4.2 preserves salient-response fidelity alongside weak-response repair, rather than merely reparameterizing a common scale.

Table 20: Post-adaptation residual-error interventions on LLaMA-3-8B. The baseline is the final DCPN checkpoint at $T = 6$. Selected post-Softmax or full-attention responses are replaced with responses from the native $T = 8$ donor; all other responses remain fixed. All configurations use full-window forward passes on identical non-overlapping WikiText-2 windows within each length. Recovery is DCPN baseline NLL minus intervention NLL; positive values indicate improvement.

Length	Windows	DCPN PPL	Softmax PPL	Full PPL	Softmax rec.	Full rec.
512	565	9.025	8.938	**8.489**	.0097	.0613
1,024	282	7.693	7.608	**7.223**	.0112	.0630
2,048	141	6.984	6.877	**6.538**	.0156	.0661
4,096	70	6.656	6.520	**6.213**	.0207	.0688

Table 21: 8K streamed intervention on LLaMA-3-8B at $T = 6 \rightarrow 8$ on WikiText-2. Each arm scores 35 aligned non-overlapping windows. Recovery is baseline minus intervention NLL. This memory-bounded protocol is reported separately and is not numerically mixed with the full-window 512–4K protocol above.

Configuration	Mean NLL ↓	PPL ↓	Recovery ↑
Baseline	1.8915	6.629	0
Softmax-only	1.8548	6.390	+0.0367
Full attention	**1.8167**	**6.152**	**+0.0747**

Table 22: LLaMA-3-8B $T = 8$ DMPC post-Softmax error trajectory on WikiText-2 (↓). This uses the same $T = 8$ timestep count as the primary LLaMA-3 DCPN evaluations, with four deterministic attention rows per layer and window. Lost mass is the mass carried by positive ANN attention entries encoded as zero.

Length	Windows	Lost mass (%)	Row-mass error	AV NMSE
1,024	282	11.42	.0868	.0184
2,048	141	15.27	.1239	.0362
4,096	70	19.06	.1597	.0532
8,192	35	25.09	.2175	.0760

Table 23: Paired streamed response diagnostic for LLaMA-3-8B at $T = 6$ (↓). Each cell is *original DMPC / final DCPN*; both checkpoints use identical WikiText-2 test windows, four sampled attention rows per layer and window. Lost mass is in percent.

Length	Windows	PPL	Lost mass (%)	Row-mass error	AV NMSE
1,024	282	8.2013 / **7.6929**	24.627 / **13.658**	.2149 / **.1618**	.1198 / **.0338**
2,048	141	7.5500 / **6.9831**	30.440 / **17.289**	.2739 / **.1898**	.1798 / **.0508**
4,096	70	7.4362 / **6.6551**	35.940 / **21.204**	.3288 / **.2167**	.2366 / **.0724**
8,192	35	8.0394 / **6.6292**	45.588 / **28.219**	.4251 / **.2711**	.3115 / **.1003**

D CALIBRATION SENSITIVITY AND CONFIRMATION SCHEDULES

D.1 CONFIRMATION ACCEPTANCE CRITERIA

The conservative variant accepts a proposal only when the upper endpoint of a paired block-bootstrap interval for its validation-NLL difference from the current response is below $-\delta$; the point-improvement variant compares the corresponding mean difference directly. Otherwise, the current response is retained. These decisions use the selection split; final test sets remain locked during calibration.

D.2 CONFIRMATION ORDER AND RETRIEVAL EVALUATION

At $T = 6$, the four selection schedules share the candidate-construction and screening rules in Section 4.3. Local calibration pairs (A, V) are collected from the current converted model. For sequential confirmation, accepted updates are committed to the model, and local pairs for subsequent sites are recollected from the updated model before candidate screening and confirmation. Forward and reverse confirmation traverse layers in orders $1 \to L$ and $L \to 1$, respectively. Table 24 reports WikiText-2 validation/test perplexity and NeedleBench scores for the resulting checkpoints. Forward sequential confirmation yields the lowest perplexity and the highest NeedleBench scores at all three lengths. Independent confirmation also improves perplexity and NeedleBench at all three lengths relative to local-only selection, while forward confirmation achieves larger gains. The forward-confirmed NeedleBench scores match the main DCPN results at $T = 6$ (Table 2).

The additional $T = 8$ pair compares local-only and forward-confirmed selection (Figure 3). The confirmed checkpoint is the same as the dual-endpoint row in Table 4, whose WikiText-2 test PPL is 5.5153. Forward confirmation improves NeedleBench by 5.11, 7.76, and 12.15 percentage points at 1K, 2K, and 4K, respectively. These NeedleBench results are reported separately from the response-family deployment comparison in Table 13. Across the four $T = 6$ schedules and the two $T = 8$ configurations, every configuration completes all 624 NeedleBench cases, with 208 cases at each length. Scores are the unweighted macro average of the three NeedleBench task groups, following the evaluation protocol in Appendix A.

Table 24: Confirmation schedules on LLaMA-2-7B at $T = 6$: WikiText-2 perplexity (↓) and NeedleBench scores (%, ↑).

	WikiText-2 PPL		NeedleBench		
Selection schedule	Validation	Test	1K	2K	4K
Local only	7.5078	6.5748	67.75	46.06	3.24
Independent	7.4389	6.5308	70.22	48.39	3.79
Reverse sequential	7.4588	6.5598	68.91	44.79	7.21
Forward sequential	**7.2881**	**6.4264**	**71.25**	**55.53**	**8.62**

D.3 FIXED-TIMESTEP DEPLOYMENT AND CALIBRATION SCOPE

The main experiments span four deployment timestep counts. Within each paired DCPN–DMPC comparison, the pretrained weights, timestep count, coefficient storage, and FS inference structure are fixed. The response is constructed and calibrated offline through candidate screening and model-level confirmation.

D.4 CALIBRATION-DATA SENSITIVITY

We further vary the number of fixed WikiText-2 calibration windows used for two-grain dual-endpoint AV selection on LLaMA-2-7B at $T = 8$. The four calibration sets are nested subsets of the same 32-window sample and share eight disjoint selection windows, the same reference model and candidate pool, and the complete WikiText-2 test set. This is a fixed-split sensitivity analysis, so it measures variation with calibration-set size but does not estimate across-seed variance.

Table 25: Calibration-data sensitivity for two-grain dual-endpoint AV selection on LLaMA-2-7B at $T = 8$. Each row uses a nested subset of the same 32-window calibration sample and eight disjoint selection windows. Test PPL uses complete WikiText-2 (166 windows; 339,802 predicted tokens); AV NMSE is measured on the held-out selection windows.

Calibration windows	Changed layers	Held-out AV NMSE ↓	WT-2 PPL ↓
4	32	.00550	5.5294
8	32	.00499	5.5316
16	32	.00480	**5.5291**
32	32	**.00471**	5.5302

Table 25 shows stable locked-test PPL across 4–32 calibration windows (range 0.0025 PPL), even as held-out AV NMSE declines with additional calibration data. This supports that the observed endpoint-selection result is not driven by a large calibration corpus.

D.4.1 INDEPENDENT DRAWS AND CROSS-MODEL CHECK

To assess whether this conclusion depends on the particular nested draw, we repeat the LLaMA-2-7B $T = 8$ two-grain dual-endpoint protocol for two independent calibration draws at $n \in \{4, 16, 32\}$, while retaining the same eight selection windows, teacher, candidate pool, and locked WikiText-2 test set. Table 26 reports the observed mean and sample standard deviation. The two draws have low variability and preserve the same weak dependence on calibration-set size. We also perform a minimal LLaMA-3-8B $T = 8$ cross-model check with $n = 4$ and $n = 16$, using its fixed eight-window selection split and the same 141-window WikiText-2 test protocol. The two settings differ by only 0.0009 PPL. These completed checks strengthen the claim that a small calibration set can suffice in the studied configurations, but do not constitute a full cross-model seed sweep.

Table 26: Independent-draw calibration sensitivity for two-grain dual-endpoint AV selection. LLaMA-2-7B $T = 8$ entries are mean $\pm$ sample standard deviation over two independent calibration draws; all use the same eight selection windows and complete WikiText-2 test set. LLaMA-3-8B $T = 8$ entries are a completed single-split cross-model check on 141 WikiText-2 windows, not a seed estimate.

Model	Calibration windows	Draws	WT-2 PPL ↓
LLaMA-2-7B $T = 8$	4	2	5.5344 ± 0.0033
	16	2	5.5310 ± 0.0008
	32	2	$\mathbf{5.5298 \pm 0.0001}$
LLaMA-3-8B $T = 8$	4	1	$\mathbf{6.2162}$
	16	1	6.2171

E EXTENDED BENCHMARK RESULTS

This appendix provides the complete RULER and NeedleBench comparisons, including higher-timestep references, followed by LongBench and conventional language-modeling and task results.

E.1 COMPLETE RULER RESULTS

Tables 1 and 27 share the same evaluation protocol; the latter additionally includes native $T = 9$ and $T = 10$ DMPC references.

At 4K, LLaMA-2-7B DCPN at $T = 8$ exceeds the native $T = 10$ reference (44.49 versus 40.20), while LLaMA-3-8B DCPN at $T = 7$ exceeds the native $T = 9$ reference (77.26 versus 76.08).

Table 27: Complete RULER macro score (%, ↑) by total context length. Numeric cells contain 13 tasks with 500 examples per task, except the LLaMA-2 **1K eligible subset (5,756 / 6,500 cases)**†: all untruncated examples available there are used (11 tasks $\times 500$, two QA tasks $\times 128$; $n = 5{,}756$), and the score is an unweighted macro over the 13 task scores. *Mean* is the unweighted average over the reported length columns within each model (three for LLaMA-2 and four for LLaMA-3). The 4K column is the formal paired protocol; completed 1K/2K cells and the LLaMA-3 8K cells are complementary length-sweep references under the same task suite. Native $T = 9$ and $T = 10$ DMPC configurations and ANN are reference models. LAS and SpikedAttention are framework-matched encoding controls.

		LLaMA-2-7B				**LLaMA-3-8B**				
T	Method	1K†	2K	4K	Mean	1K	2K	4K	8K	Mean
5	DMPC	14.34	3.14	0.21	5.90	6.07	1.17	0.47	0.12	1.96
	DCPN (ours)	**29.14**	**7.62**	**1.87**	**12.88**	**14.90**	**5.27**	**6.90**	**1.10**	**7.04**
6	DMPC	39.15	5.42	0.21	14.93	74.77	36.34	20.99	8.82	35.23
	DCPN (ours)	**72.68**	**45.91**	**4.85**	**41.15**	**91.12**	**79.78**	**58.49**	**29.34**	**64.68**
7	DMPC	69.86	26.43	1.17	32.49	89.91	71.06	27.63	9.25	49.46
	DCPN (ours)	**85.25**	**61.23**	**7.17**	**51.22**	**92.13**	**87.91**	**77.26**	**51.01**	**77.08**
8	DMPC	74.33	43.98	5.80	41.37	94.35	91.46	72.57	36.91	73.82
	DCPN (ours)	**90.21**	**80.21**	**44.49**	**71.64**	**94.46**	**93.22**	**89.40**	**61.35**	**84.61**
8	LAS	75.00	35.98	3.55	38.18	94.16	88.47	66.26	23.34	68.06
8	SpikedAttention	30.22	2.76	0.48	11.15	83.79	42.89	6.13	0.04	33.21
9	DMPC	83.51	61.94	6.34	50.60	94.40	92.29	76.08	47.49	77.57
10	DMPC	90.46	79.07	40.20	69.91	94.46	93.45	88.23	65.61	85.44
–	ANN	70.58	66.18	75.23	70.66	94.15	94.22	93.26	91.55	93.30

E.2 Complete NeedleBench Results

Table 28 extends the main-text comparison with native $T = 9$ and $T = 10$ DMPC references under the same evaluation cases.

Table 28: Complete native-window NeedleBench macro score (%, ↑) by total context budget. Scores are unweighted means over single retrieval, multi-retrieval, and multi-reasoning on the same 208 fixed cases per model–length cell. DMPC and DCPN are paired at each $T \in \{5, 6, 7, 8\}$; native $T = 9$, $T = 10$, and ANN are references. Mean averages the reported lengths within each model. LAS and SpikedAttention are framework-matched encoding controls.

		LLaMA-2-7B				**LLaMA-3-8B**				
T	Method	1K	2K	4K	Mean	1K	2K	4K	8K	Mean
5	DMPC	5.24	0.13	0.25	1.87	0.13	0.00	0.00	0.00	0.03
	DCPN (ours)	**24.39**	**0.40**	**0.76**	**8.52**	**1.96**	**1.55**	0.00	0.00	**0.88**
6	DMPC	48.66	1.06	0.13	16.62	62.35	25.04	1.29	0.76	22.36
	DCPN (ours)	**71.25**	**55.53**	**8.62**	**45.13**	**88.83**	**79.62**	**45.25**	**17.27**	**57.74**
7	DMPC	79.62	32.55	0.00	37.39	76.57	39.01	8.56	0.38	31.13
	DCPN (ours)	**85.27**	**83.39**	**23.55**	**64.07**	**93.45**	**85.04**	**71.29**	**36.24**	**71.51**
8	DMPC	79.10	69.14	4.89	51.04	90.08	77.19	63.18	21.74	63.05
	DCPN (ours)	**87.52**	**82.06**	**57.86**	**75.81**	**95.17**	**90.23**	**86.82**	**59.81**	**83.01**
8	LAS	88.35	50.76	1.10	46.74	87.50	74.09	42.20	13.02	54.20
8	SpikedAttention	30.27	0.60	0.00	10.29	73.26	9.17	0.57	0.00	20.75
9	DMPC	87.71	77.31	14.19	59.74	87.37	77.27	65.64	31.46	65.44
10	DMPC	83.47	75.42	49.91	69.60	91.17	86.52	84.56	54.29	79.14
–	ANN	83.01	73.75	67.22	74.66	97.92	95.83	97.16	97.73	97.16

For LLaMA-3-8B at 8K, DCPN $T = 8$ exceeds native $T = 10$ (59.81 versus 54.29), although the ANN remains higher. On LLaMA-2-7B, DCPN $T = 8$ exceeds native $T = 10$ at all three lengths; it also exceeds the ANN at 1K and 2K but remains 9.36 points below it at 4K.

E.3 LongBench Results

Table 29 summarizes the sample-weighted scores. These model-specific fit-window evaluations support paired comparisons within each model rather than full LongBench leaderboard claims.

Table 29: Expanded LongBench performance (sample-count-weighted score, %, ↑). LLaMA-2-7B and LLaMA-3-8B evaluate 343 and 765 fit-window examples, respectively; task scores are weighted by effective sample counts. LAS and SpikedAttention are framework-matched encoding controls. Per-task results appear in Tables 30 and 31.

T	Method	**LLaMA-2-7B**	**LLaMA-3-8B**
5	DMPC	21.47	9.20
	DCPN (ours)	**21.98**	**13.83**
6	DMPC	20.92	22.75
	DCPN (ours)	**42.55**	**41.67**
7	DMPC	31.68	30.54
	DCPN (ours)	**47.71**	**45.70**
8	DMPC	42.62	44.78
	DCPN (ours)	**53.43**	**48.04**
8	LAS	41.35	43.08
8	SpikedAttention	19.31	17.93
9	DMPC	50.51	46.22
10	DMPC	52.03	47.60
–	ANN	56.60	49.35

Tables 30 and 31 report every requested LongBench task. The effective sample count n is fixed across ANN, DMPC, and DCPN configurations within each model. Scores are percentages under the official task metric: classification score for TREC, QA F1 for TriviaQA, Qasper, and HotpotQA, ROUGE for SAMSum and GovReport, retrieval score for passage retrieval, and code similarity for LCC. The aggregate weights each non-empty task score by its effective sample count, n; consequently, the three-example LLaMA-2 HotpotQA subset does not receive the same weight as the 200-example LCC subset.

Table 30: Detailed expanded LongBench results for LLaMA-2-7B (%, ↑). n is the number of complete prompts that fit the native 4K context window and the weight used in the aggregate, shared by every configuration in the row. DMPC and DCPN are paired on the identical examples. Passage retrieval has no fit-window example and receives zero weight.

			$T = 5$		$T = 6$		$T = 7$		$T = 8$		Higher-T DMPC	
Task	n	ANN	DMPC	DCPN (ours)	DMPC	DCPN (ours)	DMPC	DCPN (ours)	DMPC	DCPN (ours)	$T = 9$	$T = 10$
TREC	33	66.67	9.09	12.12	0.00	45.45	9.09	54.55	42.42	60.61	60.61	57.58
TriviaQA	21	88.89	0.00	2.15	1.19	12.61	17.79	55.56	38.62	85.19	66.27	75.96
SAMSum	25	34.22	0.31	0.67	0.69	25.67	2.72	31.42	20.37	32.20	21.80	33.43
Qasper	52	10.39	2.51	2.92	2.35	6.83	1.38	4.59	4.84	8.63	5.02	5.44
HotpotQA	3	0.00	2.67	7.41	0.00	31.67	10.53	5.56	10.26	9.88	19.59	2.47
GovReport	9	28.47	5.48	3.79	2.55	10.97	2.71	15.27	13.15	17.93	15.20	18.67
Passage retrieval-en	0	–	–	–	–	–	–	–	–	–	–	–
LCC	200	68.48	34.34	34.35	34.94	58.20	49.99	61.09	57.49	65.47	64.65	65.28
Sample-weighted	343	56.60	21.47	21.98	20.92	42.55	31.68	47.71	42.62	53.43	50.51	52.03

Table 31: Detailed expanded LongBench results for LLaMA-3-8B (%, ↑). n is the number of complete prompts that fit the native 8K context window and the weight used in the aggregate, shared by every configuration in the row. DMPC and DCPN are paired on the identical examples. Passage retrieval has no fit-window example and receives zero weight.

			$T=5$		$T=6$		$T=7$		$T=8$		Higher-T DMPC	
Task	n	ANN	DMPC	DCPN (ours)	DMPC	DCPN (ours)	DMPC	DCPN (ours)	DMPC	DCPN (ours)	$T=9$	$T=10$
TREC	127	74.02	2.36	8.66	22.05	56.69	33.86	66.14	65.35	72.44	69.29	71.65
TriviaQA	59	85.25	2.81	5.87	36.57	79.79	61.02	84.75	84.12	83.90	86.16	86.16
SAMSum	84	43.74	0.29	2.29	7.19	34.02	14.51	38.59	40.23	44.58	43.78	44.49
Qasper	186	14.35	0.58	3.68	7.17	9.80	7.47	13.96	11.38	12.16	11.35	11.72
HotpotQA	29	13.61	0.69	0.00	8.32	12.31	9.84	11.82	9.45	14.15	12.66	12.19
GovReport	80	31.07	1.21	7.22	19.92	29.97	19.48	28.23	28.26	31.76	29.87	31.89
Passage retrieval-en	0	–	–	–	–	–	–	–	–	–	–	–
LCC	200	70.48	31.60	38.41	43.35	62.66	55.07	65.61	64.82	68.23	64.65	67.04
Sample-weighted	765	49.35	9.20	13.83	22.75	41.67	30.54	45.70	44.78	48.04	46.22	47.60

E.4 Conventional Language Modeling and Task Results

Tables 32 and 33 provide the complete perplexity and downstream-task results discussed in the main text, including their unweighted means across timesteps and available methods.

Table 32: Language-modeling perplexity (↓). Mean is the unweighted arithmetic mean of the four reported corpus PPL values, not pooled-token perplexity. DCPN is compared with DMPC at the same deployment timestep T. Shading marks DCPN; bold marks the better value in each pair. LAS and SpikedAttention are controlled encoding adaptations evaluated in the same model-specific pipeline.

		LLaMA-2-7B					**LLaMA-3-8B**				
T	Method	WT-2	C4	RP	Pile	Mean	WT-2	C4	RP	Pile	Mean
5	DMPC	34.16	32.53	23.27	17.04	26.75	40.91	52.25	40.83	27.94	40.48
	DCPN (ours)	**12.76**	**17.28**	**13.21**	**9.790**	**13.26**	**10.57**	**14.98**	**11.97**	**8.726**	**11.56**
6	DMPC	17.33	26.10	18.85	13.07	18.84	7.554	10.67	8.759	6.726	8.43
	DCPN (ours)	**6.426**	**8.266**	**6.775**	**5.477**	**6.74**	**6.983**	**9.536**	**8.080**	**6.275**	**7.72**
7	DMPC	8.274	11.98	9.346	7.221	9.21	6.778	9.298	7.867	6.172	7.53
	DCPN (ours)	**5.683**	**7.043**	**5.916**	**5.002**	**5.91**	**6.498**	**8.829**	**7.554**	**5.923**	**7.20**
8	DMPC	6.108	8.002	6.601	5.429	6.54	6.298	8.484	7.294	5.805	6.97
	DCPN (ours)	**5.515**	**6.818**	**5.747**	**4.881**	**5.74**	**6.225**	**8.395**	**7.231**	**5.785**	**6.91**
8	LAS	6.665	8.984	7.326	5.900	7.22	6.326	8.518	7.315	5.824	7.00
8	SpikedAttention	21.38	27.23	21.88	16.48	21.74	10.13	14.69	11.57	8.524	11.23
Higher-timestep and ANN references											
9	DMPC	5.661	6.965	5.871	4.971	5.87	6.233	8.393	7.226	5.738	6.90
10	DMPC	5.502	6.779	5.721	4.859	5.72	6.182	8.327	7.176	5.705	6.85
–	ANN	5.472	6.704	5.653	4.819	5.66	6.135	8.257	7.112	5.664	6.79

WT-2: WikiText-2; RP: RedPajama. LAS uses fixed binary phase encoding; SpikedAttention uses fixed binary encoding with at most two output spikes. These are framework-matched encoding controls, not reproductions of the original complete architectures.

Table 33: Downstream task accuracy (%, ↑). Mean is the unweighted average across ARC-C, ARC-E, PIQA, and WinoGrande. DCPN is compared with DMPC at the same deployment timestep T. Shading marks DCPN; bold marks the better value in each pair. LAS and SpikedAttention are controlled encoding adaptations evaluated in the same model-specific pipeline.

		LLaMA-2-7B					**LLaMA-3-8B**				
T	Method	ARC-C	ARC-E	PIQA	WG	Mean	ARC-C	ARC-E	PIQA	WG	Mean
5	DMPC	40.36	67.68	74.05	62.19	61.07	27.73	44.40	63.06	57.38	48.14
	DCPN (ours)	**42.06**	**68.98**	**75.52**	**66.14**	**63.18**	**39.51**	**61.62**	**71.16**	**65.59**	**59.47**
6	DMPC	42.06	71.25	77.04	66.30	64.16	50.17	**76.68**	79.27	71.03	69.29
	DCPN (ours)	**44.37**	**73.11**	**77.97**	**66.77**	**65.56**	**51.11**	76.64	**79.92**	**71.11**	**69.70**
7	DMPC	43.43	72.56	**78.51**	67.96	65.62	**51.96**	77.10	80.09	72.30	70.36
	DCPN (ours)	**45.65**	**74.41**	78.24	**68.90**	**66.80**	51.71	**77.19**	**80.69**	**72.61**	**70.55**
8	DMPC	46.25	73.61	**78.35**	69.77	67.00	**54.10**	77.57	81.01	73.01	71.42
	DCPN (ours)	**46.59**	**73.99**	**78.35**	**70.17**	**67.28**	53.92	**77.95**	**81.07**	**73.48**	**71.61**
8	LAS	43.77	73.11	77.69	70.32	66.22	51.62	80.43	78.73	73.56	71.09
8	SpikedAttention	37.88	72.05	75.24	67.88	63.26	45.90	77.65	76.44	69.06	67.26
Higher-timestep and ANN references											
9	DMPC	46.59	73.95	78.24	69.93	67.18	53.92	77.31	80.74	72.38	71.09
10	DMPC	46.08	73.95	78.24	70.09	67.09	53.75	77.53	80.63	73.56	71.37
–	ANN	46.42	74.62	79.11	69.06	67.30	53.50	77.69	80.79	72.69	71.17

WG: WinoGrande. LAS uses fixed binary phase encoding; SpikedAttention uses fixed binary encoding with at most two output spikes. These are framework-matched encoding controls, not reproductions of the original complete architectures.

E.4.1 Per-Corpus and Per-Task Interpretation

We next check whether the gains observed in length-resolved evaluations are accompanied by conventional language-modeling and task quality at the same deployment timestep count. Complete per-corpus perplexity and per-task accuracy results for both models are provided in Appendix Tables 32 and 33, covering every native and adapted configuration at $T \in \{5, 6, 7, 8\}$. They also include LAS fixed phase encoding and a SpikedAttention two-spike encoding as controlled $T = 8$ baselines under the same model-specific pipelines, together with native $T = 9$, $T = 10$, and ANN references.

Consistent recovery across corpora. DCPN reduces perplexity in all 32 comparisons formed by two models, four timestep counts, and four evaluation corpora. The gain depends on the model and timestep count. For LLaMA-2-7B at $T = 6$, perplexity decreases from 17.33 to 6.426 on WikiText-2, from 26.10 to 8.266 on C4, from 18.85 to 6.775 on RedPajama, and from 13.07 to 5.477 on Pile. The corresponding reductions are 62.92%, 68.33%, 64.06%, and 58.09%. At $T = 5$, the reductions remain substantial, ranging from 42.55% to 62.65% across the four corpora.

Model-dependent timestep sensitivity. LLaMA-3-8B shows its largest recovery at $T = 5$: WikiText-2 perplexity falls from 40.91 to 10.57, C4 from 52.25 to 14.98, RedPajama from 40.83 to 11.97, and Pile from 27.94 to 8.726. These changes correspond to reductions of 74.16%, 71.33%, 70.68%, and 68.77%. As DMPC becomes more accurate at higher timestep counts, the gains become smaller: 6.71%–10.63% at $T = 6$, 3.98%–5.04% at $T = 7$, and 0.34%–1.16% at $T = 8$. LLaMA-2-7B retains larger gains at these timestep counts, including 9.71%–14.80% at $T = 8$. The contrasting profiles show that the value of response adaptation depends on the conversion regime of the particular model.

Approaching higher-timestep models. For LLaMA-2-7B, the adapted $T = 6$ model reaches perplexities within 5.21% of the native $T = 8$ model on all four corpora. Its adapted $T = 8$ configuration reaches 5.515, 6.818, 5.747, and 4.881, compared with 5.502, 6.779, 5.721, and 4.859 for the native $T = 10$ reference. The relative differences are 0.24%, 0.58%, 0.45%, and 0.45%, respectively. The same adapted $T = 8$ configuration is within 1.71% of ANN perplexity across the

four corpora. Thus, selected configurations approach higher-timestep performance while retaining the lower deployment timestep count.

Downstream task recovery. Mean accuracy across ARC-Challenge, ARC-Easy, PIQA, and WinoGrande improves in all eight model–timestep settings. For LLaMA-2-7B, the mean gains are 2.11, 1.39, 1.19, and 0.28 percentage points at $T = 5, 6, 7, 8$, respectively. LLaMA-3-8B gains 11.33 percentage points at $T = 5$, increasing from 48.14% to 59.47% mean accuracy; the gains at $T = 6, 7, 8$ are 0.41, 0.19, and 0.18 percentage points. Across the 32 individual task comparisons, 27 improve, one is unchanged, and four decrease. The remaining task variation is small relative to the strongest recovery: at $T = 5$, LLaMA-3-8B improves ARC-Challenge by 11.78 and ARC-Easy by 17.22 percentage points. Together, the language-modeling and task results establish recovery across both token prediction and downstream inference.

F Spike Activity and Theoretical Energy

F.1 Spike Activity Measurement

Table 7 reports events across all 289 FS sites for the same LLaMA-2-7B, $T = 8$ checkpoints, input protocol, and dual-endpoint construction settings used in the NeedleBench endpoint ablation (Table 4). The event counts in Table 7 and retrieval scores in Table 4 are paired by method and length. "Other FS" covers all FS sites outside post-Softmax. Component counts are rounded separately to four decimal places in billions and may differ from the rounded total by 0.0001 billion; the relative column reproduces the supplied ratios rather than recalculating them from rounded totals. The energy columns report the corresponding theoretical energy estimates. Dual-endpoint energy changes are -0.035%, -0.069%, and $+0.162\%$ at 1K, 2K, and 4K, respectively.

F.2 Theoretical Synaptic Operations and Energy Calculation

We follow the MAC/AC operation-counting approach used by DMPC (Zheng et al., 2026). A multiply–accumulate (MAC) operation multiplies an activation by a weight and adds the result to an accumulator; a spike-driven accumulate (AC) operation adds the corresponding weight when a spike arrives. The theoretical inference energy is

$$E_{\text{total}} = E_{\text{MAC}} N_{\text{MAC}} + E_{\text{AC}} N_{\text{AC}}, \tag{17}$$

where N_{MAC} and N_{AC} count the respective operations over the evaluated inference sequence. We use the 45 nm CMOS reference costs (Horowitz, 2014)

$$E_{\text{MAC}} = 4.6\,\text{pJ}, \qquad E_{\text{AC}} = 0.9\,\text{pJ}. \tag{18}$$

Spike-driven synaptic operations. For an input spike train at site l, let N_l be the number of input neuron–position pairs and let $s_{l,i,t} \in \{0, 1\}$ indicate a spike at phase t. Its average firing rate is

$$f_r(l) = \frac{1}{T N_l} \sum_{i=1}^{N_l} \sum_{t=1}^{T} s_{l,i,t}. \tag{19}$$

If each input has the same synaptic fan-out, the accumulate count is

$$\text{SOPs}(l) = f_r(l)\, T\, C_{\text{dense}}(l), \tag{20}$$

where synaptic operations (SOPs) count spike-driven accumulations and $C_{\text{dense}}(l)$ is the equivalent dense MAC count for that operator at the same sequence length. For input-dependent fan-out, including causal attention, we instead count operations per emitted spike:

$$\text{SOPs}(l) = \sum_{i=1}^{N_l} \sum_{t=1}^{T} s_{l,i,t}\, k_{l,i}, \tag{21}$$

where $k_{l,i}$ is the number of active synaptic destinations for input i. This weighting connects spike activity to executed synaptic operations.

Whole-model accounting. Let $\mathcal{S}$ contain the spike-driven operators and $\mathcal{D}$ the operators executed with dense arithmetic. We compute

$$N_{\mathrm{AC}} = \sum_{l \in \mathcal{S}} \mathrm{SOPs}(l), \qquad N_{\mathrm{MAC}} = \sum_{l \in \mathcal{D}} C_{\mathrm{dense}}(l). \tag{22}$$

Dense arithmetic in the embedding stage and vocabulary projection, together with retained dense operators, contributes to N_{MAC}. Spike-driven attention and feed-forward operators contribute to N_{AC} according to their executed operations. For the ANN reference, all counted synaptic operations use the dense MAC cost:

$$E_{\mathrm{ANN}} = E_{\mathrm{MAC}} \sum_{l} C_{\mathrm{dense}}(l). \tag{23}$$

DMPC and DCPN use the same operator partition, pretrained weights, timestep count, tokenizer, inputs, and output-scoring scope within each comparison. Counts are collected separately at 1K, 2K, and 4K for each model and timestep setting, with additional LLaMA-3 8K comparisons at all four timesteps; the reported energy is the mean per evaluated sequence, converted from picojoules to joules. Relative energy changes use $(E_{\mathrm{DCPN}}/E_{\mathrm{DMPC}} - 1) \times 100\%$. ANN-relative reductions use $(1 - E/E_{\mathrm{ANN}}) \times 100\%$ and are calculated before rounding. The LLaMA-2 $T = 8$ DMPC and DCPN energies use the same underlying values as Table 7. Table 8 reports the complete model–timestep results. At $T = 8$ and 4K, DCPN reduces ANN energy by 44.33% on LLaMA-2 and 43.94% on LLaMA-3. At LLaMA-3 8K, DMPC, DCPN, LAS, and SpikedAttention reduce energy relative to ANN by 46.09%, 45.97%, 46.14%, and 65.94%, respectively. At 8K, DCPN increases energy over the same-T DMPC by 0.73%, 0.57%, 0.46%, and 0.23% for $T = 5, 6, 7, 8$, respectively.