ARE SCALING LAWS UNIVERSAL IN SPIKING LARGE LANGUAGE MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

The scaling law is a foundational property of large language models (LLMs), dictating predictable performance scaling in foundation models. Spiking LLMs are often presented as neuromorphic analogues of quantized models which convert continuous-valued LLM activations into discrete temporal spike trains. However, whether Spiking LLMs genuinely exhibit consistent scaling laws remains an open and unresolved question. In this paper, we demonstrate that the validity of scaling laws in Spiking LLMs depends fundamentally on whether the system is analyzed under a static conversion regime or a temporal dynamical regime. Under the static *accumulate-then-fire (ATF)* regime, where synaptic input is fully accumulated prior to firing, conversion acts effectively as a static quantizer. The excess loss exhibits a predictable representation-limited scaling law governed by network scale N and spike-code capacity $L(T, \phi)$. In contrast, when operating in the temporal *dynamical regime* via *online charge-and-fire (OCF)*, which represents the asynchronous parallel mode native to neuromorphic hardware, scaling laws become strictly *conditional*. In this dynamical regime, small time windows T and coarse threshold levels let temporal unevenness dominate, causing excess loss to deviate from power-law scaling against N or L. Power-law scaling is only restored when T and code capacity exceed critical dynamic thresholds. Our findings reveal that scaling claims for SLLMs cannot treat them simply as bio-inspired quantized LLMs, but must explicitly specify the dynamical regime and the (T, m) condition under OCF.

1 INTRODUCTION

Large language models follow relatively stable empirical scaling laws: training loss decreases as a power of parameter count and tokens (Kaplan et al., 2020; Hoffmann et al., 2022). Once a model is trained, a second family of laws describes what happens when it is served in reduced numerical precision. One line of work treats bit-width as contributing to an effective parameter count, so that a quantized network is a smaller network in disguise (Kumar et al., 2025). Another isolates quantization-induced degradation (QiD)—the excess loss of a q-bit model relative to a high-precision reference—and shows that this excess itself scales with model size, tokens, and bit-width (Ouyang et al., 2025). In both views, a quantized LLM still *has* a scaling law. The functional form may change, and the constants are recipe-dependent, but a power-law (or a close cousin) relating excess loss to scale and precision is assumed to exist. The practical consequence is that one can ask, and often answer, how much extra loss a given bit-width will buy at a given scale.

Spiking LLMs sit in an awkward middle ground. ANN-to-SNN conversion reuses a pretrained analog network and replaces analog activations by a finite-horizon spike code (Diehl et al., 2015; Rueckauer et al., 2017; Ding et al., 2021; Hao et al., 2023). Recent methods—SpikeLLM with Generalized Integrate-and-Fire (GIF) neurons (Xing et al., 2025), distribution-aware multi-granularity phase coding (Zheng et al., 2026), and multi-threshold (MTH) neurons (Huang et al., 2024)—scale that recipe to language models. These papers demonstrate that a converted LLM can match, and sometimes exceed, a comparably cheap quantized baseline. They do not, however, ask the scaling question that the precision literature has made standard: *how does the conversion gap of a spiking LLM change when the analog model grows, and when the spike code is enriched?* That question is not a matter of taste. If conversion error follows a law, one can allocate time steps, thresholds, and

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.

model size jointly, in the same way one allocates bits. If it does not, then reporting a single (T, ϕ) operating point, however strong, does not license an extrapolation to larger models or richer codes.

The answer is not a straightforward transplant of QiD scaling (Ouyang et al., 2025), because a spiking neuron is not only a static codebook. It is a dynamical system. Two implementations that share weights, calibration, and even the same finite set of representable values can disagree on every token if they disagree on *when* a spike is allowed to leave the membrane. We therefore separate two operating regimes that existing method papers often leave implicit.

Accumulate-then-fire (ATF). Synaptic inputs over T steps are aggregated before a spike is produced. The mapping is permutation-invariant: only the temporal mean matters. ATF is the spiking analogue of static post-training quantization. The residual is clipping plus quantization, and a representation-limited scaling law of the form

$$E_{\mathrm{ATF}} \;\simeq\; C\, N^{\beta}\, L(T,\phi)^{-\alpha} \tag{1}$$

is expected, with $L(T, \phi)$ the number of distinct states a neuron of type ϕ can realize in T steps. In this regime, GIF levels, phase bases, and MTH thresholds play the role of bit-width. A coarse codebook should produce a plateau rather than a law—exactly as a one- or two-bit quantizer does—and once L leaves that plateau, excess loss should fall with L and vary smoothly with N.

Online charge-and-fire (OCF). Membrane integration and firing occur concurrently. This is the regime that neuromorphic hardware actually runs: asynchronous, temporally causal, and parallel across neurons. Two input sequences with the same sum can then induce different spike trains. The extra discrepancy is temporal unevenness. We argue that this term is not a small perturbation of Eq. equation 1. It can dominate, and when it does, the power-law in N and L *fails to appear*. The failure is qualitative, not a mere shift of the prefactor C. Log–log plots against N or L need not be linear when T or m is small. When both are large, a scaling law in T reappears.

$$E_{\mathrm{OCF}} \;\simeq\; E_{\mathrm{rep}} + E_{\mathrm{uneven}} \tag{2}$$

This leads to the central claim of the paper. For a true (OCF) spiking LLM, a scaling law is *conditional*. For MTH neurons, the condition is operationally simple: when the time window T and the number of thresholds m are small, the neuron cannot discharge a bursty residual in one step, unevenness swamps representation error, and excess loss does not collapse onto a stable power law of model scale or capacity. When T and m are large enough that the largest threshold can absorb typical residuals, representation error again leads, and a scaling law reappears—with a form inherited from the ATF analysis. Conditionality is stronger than the statement "OCF is worse than ATF". Worse-but-lawful error is still a scaling object; unlawful error is not.

The distinction matters for how the field should read current spiking-LLM results. SpikeLLM can be implemented as a generalized quantizer (a GEMM with a multi-level code) (Xing et al., 2025); phase coding (Zheng et al., 2026) replaces a uniform codebook by a distribution-aware one. A GEMM-style or temporally aggregated implementation is close to ATF, hence close to static quantization, hence expected to exhibit scaling. Deploying the same weights as an online SNN is a different object. Reporting a scaling law without specifying the dynamical regime, or without checking the (T, m) condition, conflates a quantized LLM with a spike-driven one. The same (T, m) pair can be law-admissible under ATF and law-less under OCF, which is impossible in the QiD setting: bit-width does not create a dynamical backlog.

Our contributions are:

- A two-regime account of spiking-LLM conversion that places ATF next to precision scaling and OCF next to online integrate-and-fire dynamics, with a shared excess-loss observable $E = F_{\mathrm{SNN}} - F_{\mathrm{ANN}}$.
- A unified capacity variable $L(T, \phi)$ that lets GIF, phase-coding, and MTH neurons be compared on one axis, analogous to bit-width in QiD laws, together with the observation that matching L is not enough under OCF: the factorization of L into T and m decides whether a law exists.
- ATF measurements on phase coding, SpikeLLM, and MTH support the representation-limited half of the claim. For multi-threshold neurons, the scaling law is conditional on T and m, and it is restored when both are large enough.

2 RELATED WORK

Scaling laws for language models. Empirical power laws relating training loss to parameter count and tokens are now a standard planning tool (Kaplan et al., 2020; Hoffmann et al., 2022). Subsequent work folds in inference compute (Sardana et al., 2024) and numerical precision. One useful idea is that training precision and post-training bit-width act through an effective parameter count N_{eff}, so that a quantized model is described by the same law as a smaller full-precision model (Kumar et al., 2025). Another isolates QiD, $\Delta_{\text{qLoss}} = \text{Loss}_q - \text{Loss}_{\text{16-bit}}$, and finds that this excess *increases* with training tokens and can *decrease* with model size, so that low-bit quantization favors undertrained LLMs (Ouyang et al., 2025). We borrow two ingredients and decline a third. The ingredients are: (i) excess loss relative to a high-precision reference is the right observable, because it subtracts the analog model's own scale dependence; (ii) precision enters as a capacity variable, not as an implementation detail. We decline the assumption that a capacity variable is enough. In static quantization, enlarging the codebook cannot create a new error source that grows with the number of sequential steps. In an online spiking neuron it can. Replacing bit-width by a spike-code capacity $L(T, \phi)$ is therefore only the first half of the story; the second half is a dynamical condition that has no counterpart in PTQ.

A further caveat, already familiar in the scaling-law literature, is that fitted constants depend on how and where the law is estimated (Sardana et al., 2024). Changing the tokenizer, the data mix, or the training horizon can move prefactors, and sometimes exponents, without destroying the existence of a law. We adopt the same discipline. When we say that a configuration *has* a scaling law, we mean that log excess loss is an approximately smooth, monotone function of $\log N$ or $\log L$, not that a single (α, β, C) will travel from WikiText-2 to every future corpus.

ANN-to-SNN conversion. Rate conversion maps activations to spike counts over a window T (Diehl et al., 2015; Rueckauer et al., 2017). Residual-membrane methods were introduced precisely because count-matching is not enough: two sequences with the same sum can leave different membranes if spikes are emitted online (Hao et al., 2023). A complementary line reduces conversion error at short latency by clip-floor-shift quantization of the analog activation (Bu et al., 2022). For transformers and LLMs, three neuron families have become representative. SpikeLLM compresses IF spike length with GIF neurons and, optionally, saliency-based step allocation; it can be realized either as event-driven spikes or as a GEMM with a multi-level code (Xing et al., 2025). Multi-granularity phase coding replaces a uniform rate or phase codebook by a distribution-aware codebook (Zheng et al., 2026). Multi-threshold (MTH) neurons convert transformers by emitting several spike amplitudes in one step (Huang et al., 2024). Training-free conversion of transformers, including attention, has also been studied (Jiang et al., 2024). These works optimize a method-specific (T, ϕ) pair and report that a converted LLM can be competitive with a quantized baseline at matched bit-width or matched spike length. We ask a different question: whether the resulting error, as a function of scale, is well described by a law. Competitiveness at one operating point does not imply that moving N or L will trace a power law.

Conversion of convolutional networks already showed that more time steps are not uniformly better once online dynamics are taken seriously: unevenness error, rather than count mismatch alone, degrades converted SNNs at short horizons (Hao et al., 2023), and conversion theory is used to shrink the inference window rather than to treat larger T as automatically safer (Ding et al., 2021). LLM conversion inherits the same tension at much larger N and with a different residual (next-token cross-entropy rather than classification error). The scaling-law lens makes the tension sharper. If E is dominated by a T-growing term, then plots against $L \propto T$ will not be decreasing, and the QiD analogy fails.

Spiking versus quantization. A recurring comparison in the SpikeLLM line of work is the right one to keep in mind: an IF neuron needs T binary steps to realize T levels, while a quantizer needs $\log_2 T$ bits in one step; GIF interpolates by emitting $\log_2 L$ bits per macro-step (Xing et al., 2025). If the converted model is evaluated after temporal aggregation, this comparison is fair and ATF scaling should resemble QiD scaling. If it is evaluated online, the extra degree of freedom—when the spike occurs—can destroy that resemblance. Matching L to 2^b then matches codebook size, not error. That is the gap we study. We do not claim that ATF is uninteresting: it is the correct control for

asking whether L is a good precision proxy at all. We do claim that ATF results should not be read as neuromorphic scaling laws.

3 TWO REGIMES FOR SPIKING LLMS

3.1 CONVERSION ERROR AND CAPACITY

Let F_{ANN} be the token-averaged cross-entropy of a frozen analog LLM, and F_{SNN} the same loss after replacing analog neurons by a spike code of horizon T. Following the QiD construction (Ouyang et al., 2025), the conversion excess loss is

$$E \;=\; F_{\mathrm{SNN}} - F_{\mathrm{ANN}}. \tag{3}$$

Weights, calibration data, and evaluation sequences are held fixed, so E isolates the spike approximation. A negative E is possible in principle (a spike code can regularize), but the conversion methods we study are designed to match analog activations, and we treat $E > 0$ as the typical case. Because F_{ANN} already carries the analog model's scale dependence, E is the quantity that can be compared across N without attributing the analog law to the spikes.

Let N denote the number of converted neuronal units; in experiments we take the non-embedding parameter count of the analog backbone. For a neuron family ϕ, define the finite-horizon representational capacity

$$L(T, \phi) \;=\; \big|S_{\phi}^{*T}\big|, \tag{4}$$

the number of distinct decoded states available in T steps. Concrete instances, used throughout, are

$$L_{\mathrm{G}} \;=\; qT_G + 1, \qquad L_{\mathrm{F}} \;=\; 2^T, \qquad L_{\mathrm{M}} \;\propto\; T \cdot m, \tag{5}$$

for GIF neurons with firing level q and T_G macro-steps, phase-coding neurons with a binary spike per step, and MTH neurons with m (signed) thresholds. L is the spiking counterpart of bit-width: it indexes codebook size, not the dynamical regime. Two configurations with the same L need not have the same error under OCF.

If conversion were only a static codebook, one would expect a precision-style law

$$E \;\simeq\; C\, N^{\beta}\, L(T, \phi)^{-\alpha}, \tag{6}$$

with $\alpha > 0$ and β capturing how error accumulates across units, analogous to N_{eff} (Kumar et al., 2025) or to the N-dependence of QiD (Ouyang et al., 2025). Equation equation 6 is our *null* scaling law. The rest of this section asks when it holds. We do not presuppose the sign of β: QiD of overtrained checkpoints can fall with N, while conversion of a fixed pretrained family can rise with N if mismatch accumulates over more layers. Existence of a law is about smoothness and monotonicity in $\log N$ and $\log L$, not about a preferred sign.

3.2 ACCUMULATE-THEN-FIRE: A STATIC QUANTIZER

ATF first forms the temporal mean $\bar{x} = T^{-1} \sum_{t=1}^{T} x_t$ and then produces spikes from $W\bar{x}$. For any permutation π,

$$F_{\mathrm{ATF}}(x_{1:T}) \;=\; F_{\mathrm{ATF}}\big(x_{\pi(1:T)}\big). \tag{7}$$

The mapping is therefore a (possibly non-uniform) quantizer of a static activation. The residual decomposes as $E_{\mathrm{ATF}} = E_{\mathrm{clip}} + E_{\mathrm{quant}}$: clipping from a finite range, quantization from a finite codebook of size $L(T, \phi)$. This is exactly the setting in which QiD laws are derived (Ouyang et al., 2025), with L in place of 2^b. We therefore hypothesize that ATF obeys Eq. equation 6 *unconditionally* (given a calibrated range):

$$E_{\mathrm{ATF}} \;\simeq\; C_{\mathrm{ATF}}\, N^{\beta}\, L(T, \phi)^{-\alpha}. \tag{8}$$

Two qualitative predictions follow from the precision literature and should be checked empirically. (i) For fixed L, E_{ATF} should vary smoothly with N. We do not presuppose the sign of β. (ii) For fixed N, E_{ATF} should decrease in L once L leaves a coarse plateau on which the codebook is too small to track the activation distribution. Below that plateau, no power law is expected even under ATF—not because of spikes, but because the quantizer is simply too coarse. We treat $L < 32$ as

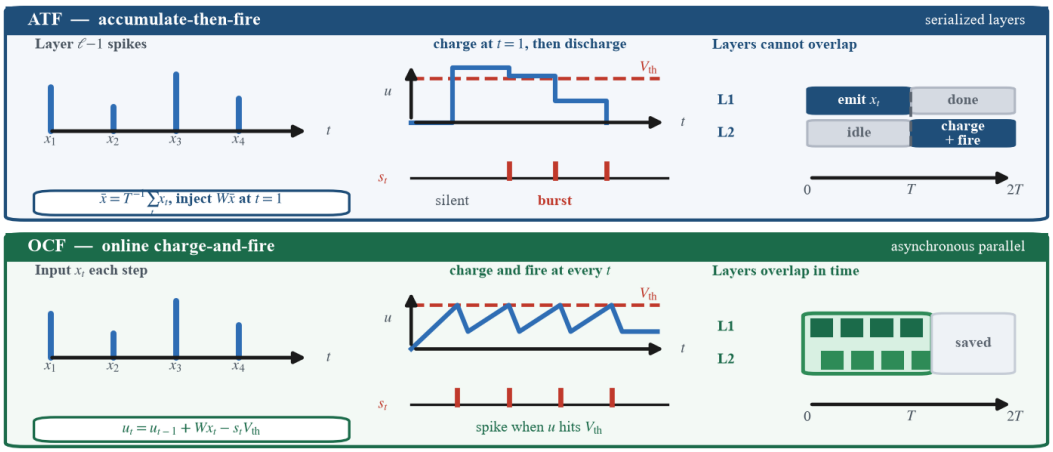


Figure 1: Two dynamical regimes. **ATF** accumulates the whole window and only then emits a spike burst, so a downstream neuron stays idle until T (wall-clock $\sim 2T$). **OCF** charges and fires online: a spike can leave as soon as the membrane hits threshold, so successive layers overlap in time (asynchronous parallelism, wall-clock $\sim T$).

this coarse bin, matching the scale at which a five-bit codebook begins to have room to move. A saturating form

$$E_{\mathrm{ATF}}(L) \;=\; \frac{\Delta_0 - \varepsilon}{1 + (L/L_0)^{\alpha}} + \varepsilon \tag{9}$$

is allowed and still counts as a law: the floor ε is the analogue of an irreducible calibration mismatch, and a large R^2 on Eq. equation 9 is a stronger description of a plateau-then-decay shape than a pure power law fitted through the knee.

ATF is a diagnostic, not a deployment mode. It answers: *if* we strip out time, does a spiking LLM scale like a quantized LLM? A yes supports using L as a precision proxy. It does not license claims about neuromorphic inference. Several published conversions are closer to ATF than their "spiking" label suggests: SpikeLLM notes that GIF can be realized as a GEMM with a multi-level code (Xing et al., 2025); phase coding supplies a distribution-aware codebook (Zheng et al., 2026), which becomes a static quantizer if it is evaluated after temporal aggregation. MTH conversion of transformers is likewise a finite codebook of thresholds (Huang et al., 2024). Those systems should, and in Section 4 do, look like quantizers.

3.3 Online charge-and-fire: a conditional law

Unlike ATF, OCF couples membrane integration with causal spike emission,

$$u_t = u_{t-1} + Wx_t - s_t V_{\mathrm{th}}, \qquad s_t = \Theta(u_{t-1} + Wx_t - V_{\mathrm{th}}). \tag{10}$$

Hence, two input sequences with the same temporal sum can induce different membrane trajectories and outputs (Hao et al., 2023). OCF therefore introduces an additional error:

$$E_{\mathrm{OCF}} = E_{\mathrm{rep}} + E_{\mathrm{uneven}}, \tag{11}$$

where E_{uneven} captures the discrepancy induced by causal temporal dynamics. ATF with the same weights and calibration serves as a natural counterfactual for isolating this temporal component.

Representation capacity does not determine temporal dynamics. This distinction is particularly clear for MTH neurons. With m uniformly spaced thresholds calibrated by $m\theta_{\mathrm{M}} = R_{\mathrm{M}}$, the number of representable states is $L_{\mathrm{M}}(T, m) = 2mT + 1$, and the ATF representation error follows

$$E_{\mathrm{rep}} = C_{\mathrm{M}} N^{\beta_{\mathrm{M}}} (mT)^{-\alpha_{\mathrm{M}}}. \tag{12}$$

Thus, under ATF, m and T enter only through the effective code resolution mT. OCF breaks this equivalence. Although L_{M} grows with mT, the maximum membrane charge that can be discharged within one timestep remains

$$C_{\mathrm{M}}^{\mathrm{temp}} = m\theta_{\mathrm{M}} = R_{\mathrm{M}}, \tag{13}$$

independent of both m and T, revealing a *fundamental decoupling between representation capacity and temporal processing capability that does not arise in static quantization*.

Temporal unevenness induces a dynamical scaling condition. To characterize the temporal error introduced by OCF, we model the input as $I_t = \mu + \xi_t$, where ξ_t is a zero-mean conditionally sub-Gaussian fluctuation. Under the stable condition $|\mu| < R_{\mathrm{M}}$, Appendix D.3 shows that the membrane backlog behaves as a negative-drift reflected random walk, yielding

$$E_{\mathrm{uneven}} \le K_{\mathrm{M}} \left(A_{\mathrm{M}} + \frac{B_{\mathrm{M}}}{m^2} \right) T^{-2}, \tag{14}$$

where $A_{\mathrm{M}} = \sigma^4/\varepsilon_{\mathrm{M}}^2$, $B_{\mathrm{M}} = R_{\mathrm{M}}^2$, and $\varepsilon_{\mathrm{M}} = R_{\mathrm{M}} - |\mu|$. Importantly, Equation 14 shows that temporal unevenness decays differently from the representation error. Their competition therefore determines whether the ATF scaling behavior survives under causal OCF dynamics. The following theorem characterizes the corresponding scaling-admissible regime in terms of the temporal horizon T and threshold number m.

Theorem 3.1 (Conditional scaling under OCF) *Suppose* $E_{\mathrm{rep}} = C_{\mathrm{M}} N^{\beta_{\mathrm{M}}} (mT)^{-\alpha_{\mathrm{M}}}$ *with* $0 < \alpha_{\mathrm{M}} < 2$ *and* $\beta_{\mathrm{M}} > 0$. *Let* $F_{\mathrm{M}}(m) = A_{\mathrm{M}} m^{\alpha_{\mathrm{M}}} + B_{\mathrm{M}} m^{\alpha_{\mathrm{M}}-2}$ *and* $\Gamma_{\mathrm{M}} = K_{\mathrm{M}}/C_{\mathrm{M}}$. *For* $N \in [N_{\min}, N_{\max}]$ *and* $\eta \in (0, 1]$, *the representation-based scaling law is preserved under OCF if*

$$T \ge \left[\frac{\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m)}{\eta N_{\min}^{\beta_{\mathrm{M}}}} \right]^{\frac{1}{2-\alpha_{\mathrm{M}}}}, \qquad m \gtrsim \frac{1}{T} \left[\frac{\Gamma_{\mathrm{M}} B_{\mathrm{M}}}{\eta N_{\min}^{\beta_{\mathrm{M}}}} \right]^{\frac{1}{2-\alpha_{\mathrm{M}}}} \tag{15}$$

where the second condition is the explicit form in the finite-threshold-dominated regime $A_{\mathrm{M}} m^2 \ll B_{\mathrm{M}}$. *Under these conditions,* $E_{\mathrm{rep}} \le E_{\mathrm{OCF}} \le (1+\eta) E_{\mathrm{rep}}$, $E_{\mathrm{OCF}} = \Theta\big(N^{\beta_{\mathrm{M}}}\big)$ *for fixed* (T, m).

Remark 3.2 *Theorem 3.1 shows that OCF scaling is jointly controlled by the temporal horizon T and threshold number m, rather than by nominal code capacity alone. In particular, two configurations with the same mT can have comparable ATF representation error but different OCF behavior because temporal unevenness still depends explicitly on T. Therefore, equal representation capacity does not imply equivalent temporal behavior. Scaling emerges only when the temporal dynamics satisfy the admissible condition. This distinction between representation capacity and temporal dynamics fundamentally separates spiking scaling from static quantization.*

4 Experiments

This section reports conversion sweeps on the OPT family. Fitted constants are in Appendix C.

4.1 Common setup

Backbone. OPT (Zhang et al., 2022) provides a single recipe over several orders of magnitude in N. We convert the non-embedding stack and report N as that count. The scale points that appear on the N-plots are

$$\{125\mathrm{M},\ 1.3\mathrm{B},\ 2.7\mathrm{B},\ 6.7\mathrm{B},\ 13\mathrm{B},\ 30\mathrm{B}\},$$

with 2.7B on the OCF N-sweep. Fine (T, ϕ) sweeps are run on 125M; 6.7B repeats the T-sweep (Figure 4).

Data and metric. F_{ANN} and F_{SNN} are token-averaged cross-entropies on WikiText-2 (Merity et al., 2016), RefinedWeb (Penedo et al., 2023), and Dolma (Soldaini et al., 2024). E is computed on identical sequences. In the plots, $\Delta\mathcal{L}$ is this excess.

Neurons. We convert with GIF (SpikeLLM without salient-channel search or weight quantization) (Xing et al., 2025), multi-granularity phase coding (Zheng et al., 2026), and MTH (Huang et al., 2024). A globally unified (T, q) (respectively l or m) keeps L from being confounded with mixed-precision allocation. For MTH, ATF and OCF share weights and calibration and differ only in the forward dynamics. Phase coding and SpikeLLM are run as ATF.

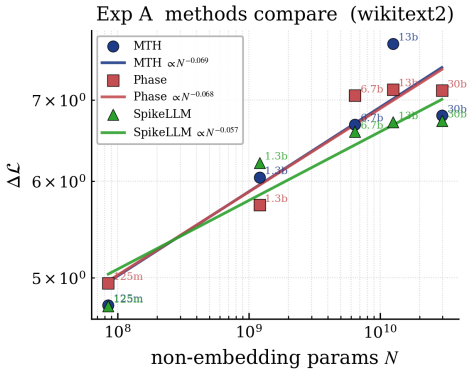
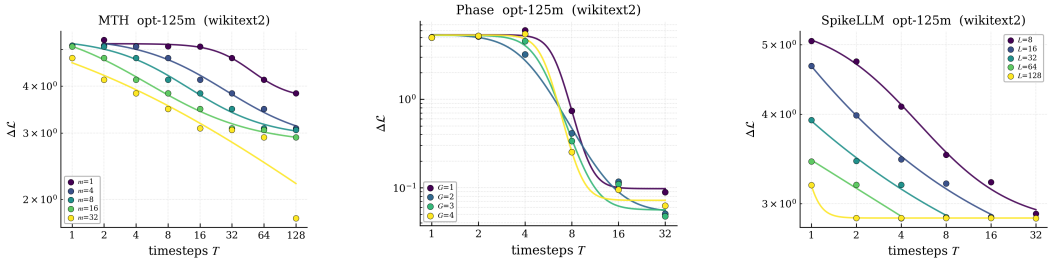


Figure 2: ATF excess loss versus non-embedding parameter count on WikiText-2, one frozen setting per family. All three families rise slowly with N. The MTH point at 13B lies above the line.



(a) MTH (b) Phase (c) SpikeLLM / GIF

Figure 3: ATF excess loss versus T on OPT-125M / WikiText-2. Colour is m, granularity, or the GIF level. Curves fall or flatten; they do not rise with T.

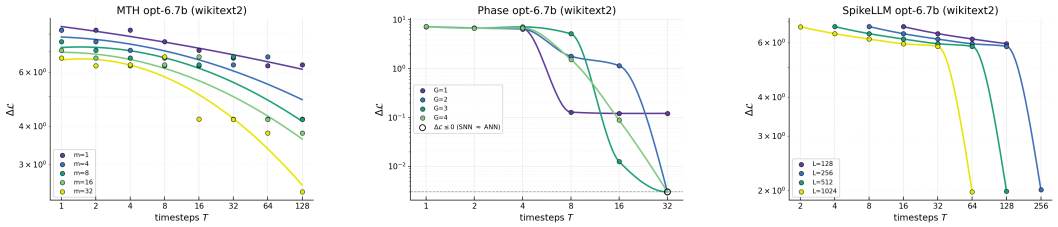
Grids. ATF covers phase coding, SpikeLLM, and MTH, with $T \in \{1, 2, 4, \ldots, 256\}$ and method-specific ϕ. The online sweeps use multi-threshold neurons: $\Delta\mathcal{L}$ versus N at $m \in \{1, 2, 4\}$, and versus T on OPT-125M.

4.2 ATF: A SPIKING LLM SCALES LIKE A QUANTIZER

Network scale. Fixing (T, ϕ) holds L constant, so Eq. equation 8 predicts a smooth trend of E_{ATF} in N. Figure 2 shows log–log plots of excess loss versus non-embedding parameter count on WikiText-2 for the three families, at representative frozen settings (MTH $(T, m) = (4, 8)$, phase $(T, l) = (8, 2)$, GIF $(T, q) = (2, 8)$). Excess loss *increases* slowly with N: larger analog networks are slightly harder to convert at fixed neuronal capacity. Occasional checkpoints (e.g. 13B) sit off the line; we do not over-interpret a single point. The qualitative message is that, with time stripped out, conversion error is a regular function of N, as a QiD-style law requires. The sign of β is positive here, opposite to the "larger models quantize more easily" slogan for overtrained PTQ, and consistent with mismatch accumulating over a deeper stack at matched L.

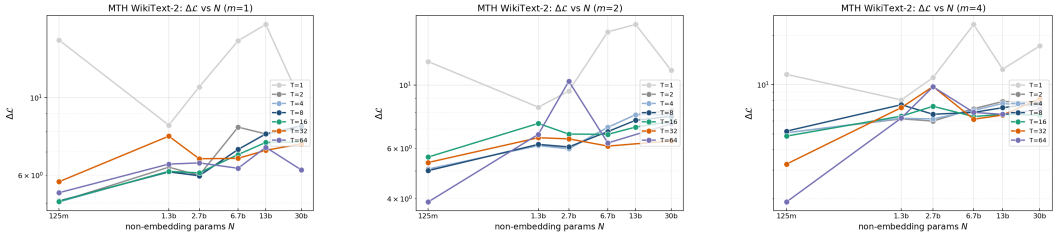
Time slices at fixed model size. Figure 3 slices ATF excess loss against T for the three neuron families on OPT-125M / WikiText-2. With time stripped out of the dynamics, increasing T does not raise error. MTH falls with T, and a larger m lies lower. Phase stays flat at small T and then drops. SpikeLLM / GIF falls with T and flattens earlier when the level in the legend is larger. None of the ATF curves turns up as T grows. Figure 4 is the same three families on OPT-6.7B / WikiText-2: again the curves fall or flatten with T. The matched RefinedWeb and Dolma panels, at both scales, are Figures 7 and 8.

Stripped of time, a converted spiking LLM on these plots behaves like a static quantizer: error rises slowly with N at a frozen code, and it does not rise with T.

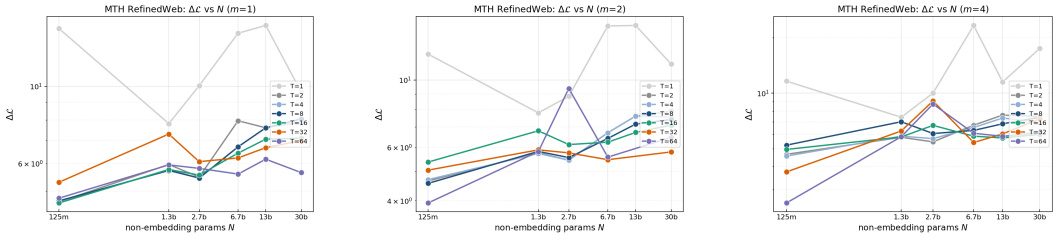


(a) MTH (b) Phase (c) SpikeLLM / GIF

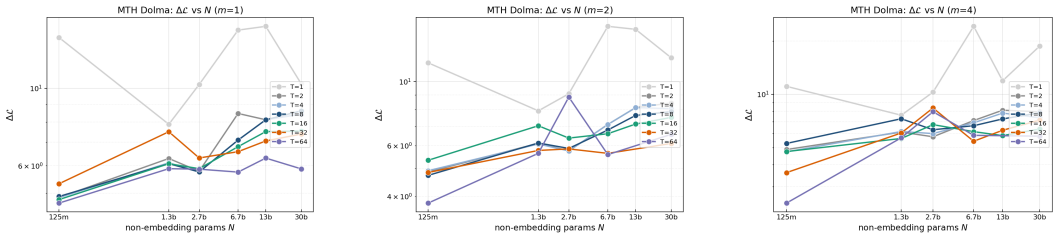
Figure 4: ATF excess loss versus T on OPT-6.7B / WikiText-2. Colour is the second knob, as in Figure 3. Curves fall or flatten with T.



(a) WikiText-2, $m = 1$ (b) WikiText-2, $m = 2$ (c) WikiText-2, $m = 4$



(d) RefinedWeb, $m = 1$ (e) RefinedWeb, $m = 2$ (f) RefinedWeb, $m = 4$



(g) Dolma, $m = 1$ (h) Dolma, $m = 2$ (i) Dolma, $m = 4$

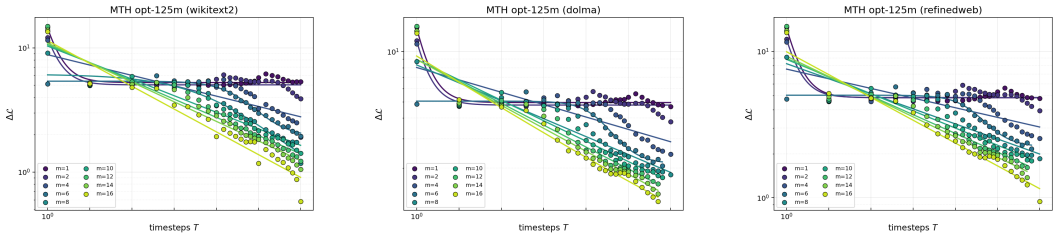
Figure 5: OCF MTH excess loss versus N. Rows are WikiText-2, RefinedWeb, and Dolma. Colour is T. $T = 1$ is irregular. For $T \geq 2$ the curves separate at 125M and bunch up at larger N.

4.3 OCF FOR MULTI-THRESHOLD NEURONS

Eq. equation 11 writes the online excess as representation error plus temporal unevenness, so a law in N and L depends on T and m.

Figure 5 is the online N-sweep for MTH at $m \in \{1, 2, 4\}$ on WikiText-2, RefinedWeb, and Dolma. At each m, $T = 1$ is high and non-monotone, while $T \geq 2$ sits lower. Those larger-T curves still separate at 125M and bunch up from 1.3B onward. Extra T on this grid does not line the points up in $\log N$.

Figure 6 is the online T-sweep on OPT-125M, the same weights as Figure 3, for MTH. On WikiText-2, Dolma, and RefinedWeb, $m = 1$ and $m = 2$ stay high across T. Larger m decreases, and $m = 16$



(a) WikiText-2 (b) Dolma (c) RefinedWeb

Figure 6: OCF MTH excess loss versus T on OPT-125M. Colour is m. Small m stays high; large m decreases with T.

falls the most. Extra time steps lower online error when m is large enough, and they do not when m is small. At $m \in \{1, 2, 4\}$ the N-sweep does not form a power law.

Figure 5 repeats the three corpora of the ATF sweeps at $m \in \{1, 2, 4\}$, with colour T. The $T = 1$ trace lies above the others and is not monotone in N. From $T = 2$ the level drops, mostly at 125M: by 1.3B the remaining colours lie close together and stay close through 30B. WikiText-2, RefinedWeb, and Dolma bunch in the same way.

Figure 6 fixes the checkpoint at OPT-125M, the same weights as Figure 3, and varies m. At $m = 1$ and $m = 2$ the excess stays high across the whole range of T on every corpus. Larger m lowers the curve, and $m = 16$ falls the most. Under ATF, GIF, phase coding, and MTH all fall or flatten with T on this checkpoint and on OPT-6.7B (Figure 4). Online, that decay shows up only once m is large. A long window at $m = 1$ does not produce it. What the two T-sweeps share is the ordering in the second knob: a finer code lies lower, and extra steps reduce error on that finer code.

Appendix C fits the online decay. For $T \geq 8$ and $m \geq 6$, $\Delta\mathcal{L} \propto T^{-\gamma}$ on all three corpora, with γ between 0.50 and 0.71 and $R^2 \geq 0.91$. At $T \leq 8$ the same form has $R^2 < 0.82$ for every m. At $T \geq 8$ and $m \leq 2$ the fit stays weak, and γ is negative at $m = 1$. Figure 5 is that lower corner. The falling curves in Figure 6 are the corner in which the power of T holds.

The ordering is the same on WikiText-2, Dolma, and RefinedWeb, in both sweeps. In Figure 5 the colours separate most at 125M and collapse from 1.3B onward. In Figure 6 the gap among values of m grows with T: the coarse codes stay flat while $m = 16$ keeps falling. The loss at a short window and a small m and the loss at a long window and $m = 16$ are different regimes of the same weights, and only the second sits on the power of T above. Under ATF the second knob changes the level, but GIF, phase coding, and MTH still move down as T increases. Online, m decides whether extra steps move the loss at all.

The three corpora agree on the shape. At $m = 16$, OPT-125M falls on WikiText-2, Dolma, and RefinedWeb alike, and the traces at $m = 1$ and $m = 2$ stay flat on all three. Past 125M, raising T from 2 to 16 barely moves the excess, at every m in the N-grid. Lengthening a coarse code does not recover that decay.

5 Conclusion

A spiking LLM scales like a quantized LLM only under a stated dynamical regime. Accumulate-then-fire is a static codebook: on OPT, excess loss rises slowly with N at a frozen code and falls or flattens with T, for GIF, phase coding, and MTH, on WikiText-2, RefinedWeb, and Dolma, at 125M and at 6.7B. For MTH, a scaling law exists only when T and m are large enough that temporal unevenness no longer swamps representation error. When both are small, excess loss does not collapse onto a stable power law of N or L; the N-sweep at $m \in \{1, 2, 4\}$ is that failure, not the restored law. When both are large enough, representation error leads again and the law inherited from ATF reappears. ATF, including GEMM-style GIF, is the comparison to a quantizer. Claims about spike-driven LLMs should be made in OCF and should name the (T, m) regime: more time steps are not automatically more law-like, and a model that looks quantized in a simulator may not look so on a chip. This is a conversion study of one analog family, post-training rather than trained from scratch.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing prose and for organizing the manuscript into ICLR LaTeX. We have not used generative AI tools to generate experimental numbers or figures. Quantitative results are taken from the conversion sweeps in Section 4. We reviewed the text and take responsibility for the final content.

ETHICS STATEMENT

This work studies post-training conversion of publicly released language models on standard text corpora. It does not involve human subjects or private data.

REPRODUCIBILITY STATEMENT

Regimes, capacity definitions, and experimental grids are specified in Sections 3 and 4. Multi-threshold conditions are in Appendix D.3. GIF and phase-coding bounds are in Appendix A. Code will be submitted as supplementary material.

REFERENCES

Tong Bu, Wei Fang, Jianhao Ding, PengLin Dai, Zhaofei Yu, and Tiejun Huang. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. In *International Conference on Learning Representations*, 2022.

Peter U Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2015.

Jianhao Ding, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Optimal ANN-SNN conversion for fast and accurate inference in deep spiking neural networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2328–2336, 2021.

Zecheng Hao, Tong Bu, Jianhao Ding, Tiejun Huang, and Zhaofei Yu. Reducing ANN-SNN conversion error through residual membrane potential. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11–21, 2023.

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

Zihan Huang, Xinyu Shi, Zecheng Hao, Tong Bu, Jianhao Ding, Zhaofei Yu, and Tiejun Huang. Towards high-performance spiking transformers from ANN to SNN conversion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10688–10697, 2024.

Yizhou Jiang, Kunlin Hu, Tianren Zhang, Haichuan Gao, Yuqian Liu, Ying Fang, and Feng Chen. Spatio-temporal approximation: A training-free SNN conversion for transformers. In *International Conference on Learning Representations*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Xu Ouyang, Tao Ge, Thomas Hartvigsen, Zhisong Zhang, Haitao Mi, and Dong Yu. Low-bit quantization favors undertrained LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32338–32348, 2025.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, 11:682, 2017.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 43445–43460, 2024. URL `https://proceedings.mlr.press/v235/sardana24a.html`.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024.

Xingrun Xing, Boyan Gao, Zheng Liu, David Clifton, Shitao Xiao, Wanpeng Zhang, Li Du, Zheng Zhang, Guoqi Li, and Jiajun Zhang. SpikeLLM: Scaling up spiking neural network to large language models via saliency-based spiking. In *International Conference on Learning Representations*, 2025.

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

Hanyuan Zheng, Haozhen Zhang, Tianshuo Chen, Zhaogeng Liu, Yi Chang, and Bin Gu. Distribution-aware multi-granularity phase coding: Towards lower conversion error for spike-driven large language models. In *International Conference on Learning Representations*, 2026.

A GIF AND PHASE-CODING BOUNDS

We record the GIF and phase-coding counterparts of the MTH condition $2^{m-1}\theta > |\mu|$. Phase coding and SpikeLLM are measured as accumulate-then-fire, so these bounds are not the online measurements.

GIF. A GIF macro-step merges q binary IF steps (Xing et al., 2025). The codebook size is $L_G = qT_G + 1$, and the local mean-squared distortion of a floor quantizer is $O((qT_G)^{-2})$. Increasing q refines the grid but does not raise the maximum analog discharge per macro-step, which remains the calibrated range θ_G. Read online, the two decoded outputs differ only by a terminal residual. Modelling the input as $\mu + \xi_t$ with sub-Gaussian ξ_t, a reflected-walk bound in the stable regime $|\mu| < \theta_G$ gives

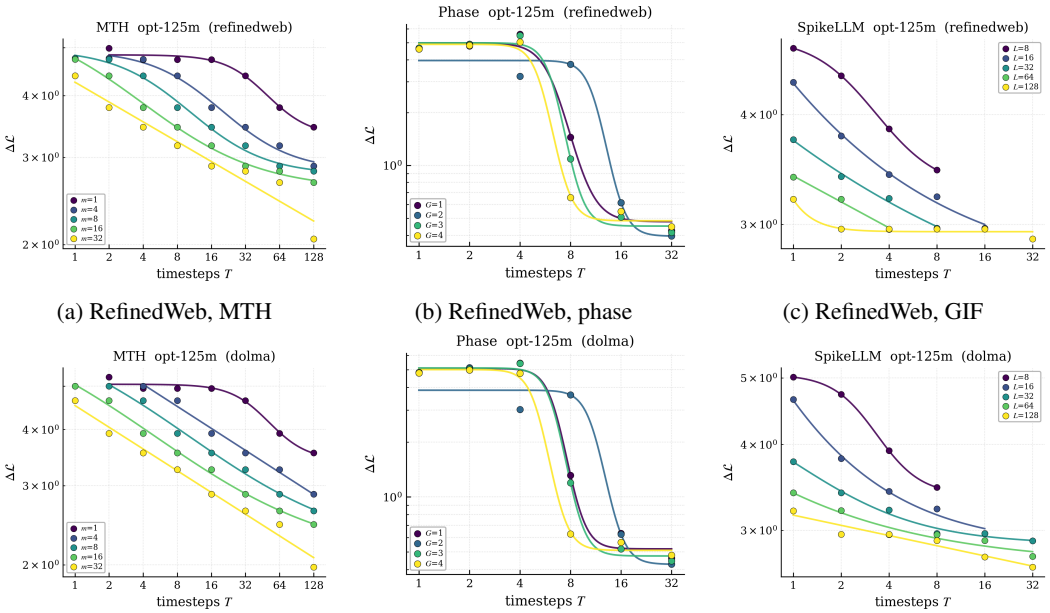
$$E_{\text{uneven}} \;=\; O\big(T_G^{-2}(\sigma^4/\varepsilon_G^2 + \theta_G^2/q^2)\big), \qquad \varepsilon_G = \theta_G - |\mu|.$$

Matching this to $E_{\text{ATF}} = \Theta((qT_G)^{-\alpha})$ shows that scaling is preserved only for $q = O(T_G^{2/\alpha-1})$. Large q at small T_G refines the grid and leaves that backlog in place. If $|\mu| > \theta_G$, the backlog grows linearly and no q restores a representation-led law.

Phase coding. A T-step binary phase code has $L_F = 2^T$ patterns; granularity l reshapes their locations, not their count (Zheng et al., 2026). Read step by step, a bit flips relative to accumulate-then-fire only if unseen future input exceeds the current margin $m_t(l)$. Sub-Gaussian tails then bound the flip probability by $\exp(-m_t(l)^2/(2\sigma^2(T-t)))$. Small T or a codebook that parks decision boundaries near typical activations makes unevenness dominate, again breaking Eq. equation 6 even when L_F is large.

B ADDITIONAL PLOTS

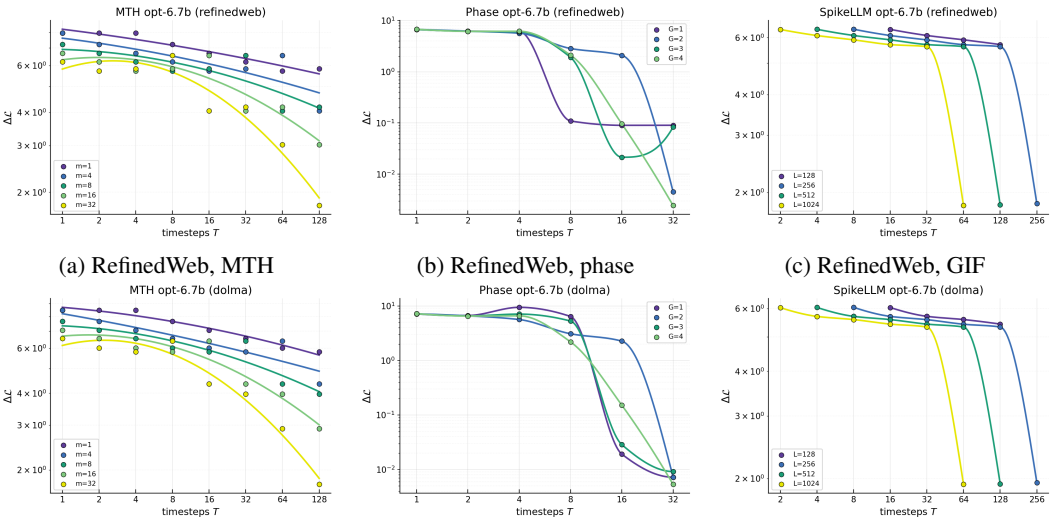
Figure 7 is the OPT-125M T-sweep of Figure 3 on RefinedWeb and Dolma, for MTH, phase coding, and GIF. Figure 8 is the same three families on OPT-6.7B.



(a) RefinedWeb, MTH (b) RefinedWeb, phase (c) RefinedWeb, GIF

(d) Dolma, MTH (e) Dolma, phase (f) Dolma, GIF

Figure 7: ATF excess loss versus T on OPT-125M, RefinedWeb and Dolma. Same sweep as Figure 3.



(a) RefinedWeb, MTH (b) RefinedWeb, phase (c) RefinedWeb, GIF

(d) Dolma, MTH (e) Dolma, phase (f) Dolma, GIF

Figure 8: ATF excess loss versus T on OPT-6.7B, RefinedWeb and Dolma. Same sweep as Figure 4.

C REPORTED FITS

The excess in the tables is the plotted quantity $\Delta\mathcal{L} = F_{\text{SNN}} - F_{\text{ANN}}$. For the MTH fits the spike budget is $L = T \cdot m$. Log-quadratic and PCHIP curves drawn only to connect markers are omitted.

ATF excess versus N. At a frozen code, MTH uses $(T, m) = (4, 8)$ and SpikeLLM uses $T = 2$ with level 8. The five scales are 125M, 1.3B, 6.7B, 13B, and 30B,

$$\Delta\mathcal{L} = C\, N^{\beta}. \tag{16}$$

MTH and SpikeLLM both rise slowly with N, with β between 0.055 and 0.076.

Table 1: ATF fit of Eq. equation 16. Phase coding is not given an exponent: RefinedWeb and Dolma have two models, and WikiText-2 has three points ($\beta = 0.059$, $R^2 = 0.32$).

Method	Corpus	C	β	R^2
MTH	WikiText-2	1.397	0.0694	0.787
MTH	RefinedWeb	1.181	0.0740	0.734
MTH	Dolma	1.186	0.0761	0.729
SpikeLLM	WikiText-2	1.793	0.0565	0.885
SpikeLLM	RefinedWeb	1.733	0.0547	0.877
SpikeLLM	Dolma	1.819	0.0548	0.918

ATF MTH excess versus spike budget. On OPT-125M and OPT-6.7B, for $L \geq 32$, the saturating law fits better than a pure power law on all three corpora,

$$\Delta\mathcal{L}(L) = \frac{\Delta_0 - \varepsilon}{1 + (L/L_0)^{\alpha}} + \varepsilon. \tag{17}$$

The joint form across scale and budget has a floor that the fit drives to 10^{-15}, written here as zero:

Table 2: Saturating ATF MTH fit of Eq. equation 17, $L \geq 32$.

Corpus	Model	Δ_0	ε	L_0	α	R^2
WikiText-2	125M	4.092	0.00595	2866	1.034	0.948
WikiText-2	6.7B	6.850	0.000139	1876	0.890	0.975
RefinedWeb	125M	3.959	0.0837	3130	0.719	0.946
RefinedWeb	6.7B	6.239	0.0779	1732	1.086	0.972
Dolma	125M	4.620	0.0862	1483	0.566	0.958
Dolma	6.7B	6.566	0.0469	1497	0.993	0.986

$$\Delta\mathcal{L} = C\, N^{\beta}\, L^{-\gamma}. \tag{18}$$

Here β agrees with the 0.07 scale in Table 1, and γ lies between 0.21 and 0.23. The exponents are

Table 3: Joint ATF MTH fit of Eq. equation 18.

Corpus	C	β	γ	R^2
WikiText-2	3.108	0.0715	0.225	0.872
RefinedWeb	2.615	0.0735	0.212	0.866
Dolma	2.765	0.0758	0.228	0.901

consistent across the three corpora.

OCF MTH excess versus T on OPT-125M. A power of T appears only at high m and high T. On low m or low T, the same power law does not hold. Every R^2 in the three tables below is computed for $\ln \Delta\mathcal{L}$ against $\ln T$,

$$\Delta\mathcal{L} = C\, T^{-\gamma}. \tag{19}$$

The model is OPT-125M, on all three corpora. The low-T window is $T \leq 8$ (five points). The high-T window is $T \geq 8$.

Low T: no scaling law. For $T \leq 8$, all nine values of m have R^2 below 0.82. At $m = 6$ the exponent γ is negative, so the excess does not fall on that slice.

Table 4: OCF MTH, $T \leq 8$ on OPT-125M. Not a scaling law: every $R^2 < 0.82$, and $\gamma < 0$ at $m = 6$.

	WikiText-2		RefinedWeb		Dolma	
m	γ	R^2	γ	R^2	γ	R^2
1	0.436	0.608	0.470	0.606	0.431	0.604
2	0.369	0.629	0.410	0.639	0.378	0.651
4	0.330	0.591	0.339	0.520	0.316	0.542
6	−0.056	0.402	−0.077	0.505	−0.060	0.410
8	0.252	0.638	0.237	0.508	0.234	0.592
10	0.380	0.579	0.410	0.565	0.416	0.622
12	0.441	0.622	0.467	0.618	0.469	0.665
14	0.448	0.685	0.477	0.692	0.475	0.716
16	0.548	0.807	0.524	0.794	0.524	0.814

Table 5: OCF MTH power law of Eq. equation 19 for $T \geq 8$ and $m \leq 4$ on OPT-125M. Not a scaling law.

m	Corpus	C	γ	Points	T	R^2
1	WikiText-2	4.28	−0.067	21	8–64	0.407
1	RefinedWeb	4.08	−0.055	20	8–64	0.331
1	Dolma	4.65	−0.021	20	8–64	0.069
2	WikiText-2	7.11	0.098	21	8–64	0.277
2	RefinedWeb	6.09	0.063	20	8–64	0.152
2	Dolma	6.93	0.109	20	8–64	0.406
4	WikiText-2	18.15	0.470	21	8–64	0.819
4	RefinedWeb	15.58	0.419	20	8–64	0.868
4	Dolma	15.99	0.439	20	8–64	0.885

High T, low m: still no scaling law. On $T \geq 8$, $m = 1$ and $m = 2$ still have R^2 between 0.07 and 0.41. At $m = 1$, γ is negative, so $\Delta\mathcal{L}$ does not fall with T. At $m = 4$, R^2 rises to 0.82–0.88: close to 0.9, and not reported as a law. If $m = 1$ is fitted anyway on the whole range $T = 1$–64, γ pins at the optimizer upper bound 5 and the curve falls to a floor. The floors are 5.26 on WikiText-2, 4.83 on RefinedWeb, and 4.94 on Dolma. That is saturation, not a power-law exponent.

High T, high m: eighteen laws. For $T \geq 8$ and $m \geq 6$, three corpora and six values of m give eighteen fits. All have $R^2 \geq 0.91$, and sixteen have $R^2 \geq 0.95$. The exponent γ lies between 0.50 and 0.71 and is not pinned at a bound. This is the OCF scaling law: $\Delta\mathcal{L} \propto T^{-\gamma}$ with $\gamma \approx 0.5$–0.7. At large T and large m these measurements follow a scaling law in T, reported in Table 6.

What is not a reported law. The log-quadratic drawn on the 6.7B $\Delta\mathcal{L}$–T panels, and the PCHIP curves for phase coding and SpikeLLM, only connect markers. Phase saturating and joint fits pin an exponent at a bound, or give a negative R^2 along a T or m slice; on those panels $\Delta\mathcal{L}$ is already near zero. The SpikeLLM fit against L pins α at 8 on 125M, and the pure power law on 6.7B has R^2 below 0.1. An OCF fit forced over the whole range $T = 1$–64 is not reported: low T pulls a high-m power law down, and low m drives γ to the upper bound 5.

D Neuron-Specific Scaling Laws under Temporal Unevenness

In the ATF regime, temporal inputs are first aggregated before spike generation. Therefore, the conversion error is dominated by the finite representational capability of the spiking neuron, including clipping and quantization errors. In contrast, in the OCF regime, spike generation is performed concurrently with membrane integration, introducing an additional temporal unevenness error.

For a neuron family j, we decompose the OCF conversion error as:

$$\mathcal{E}^{(j)}_{\mathrm{OCF}} = \mathcal{E}^{(j)}_{\mathrm{rep}} + \mathcal{E}^{(j)}_{\mathrm{uneven}}, \tag{20}$$

Table 6: OCF MTH power law of Eq. equation 19 for $T \geq 8$ and $m \geq 6$ on OPT-125M. All eighteen rows are reported laws ($R^2 \geq 0.91$).

m	Corpus	C	γ	Points	T	R^2
6	WikiText-2	21.84	0.567	21	8–64	0.956
6	RefinedWeb	17.86	0.520	19	8–56	0.957
6	Dolma	18.06	0.539	19	8–56	0.954
8	WikiText-2	21.90	0.644	21	8–64	0.944
8	RefinedWeb	16.33	0.539	20	8–64	0.976
8	Dolma	15.28	0.533	20	8–64	0.959
10	WikiText-2	20.61	0.639	21	8–64	0.992
10	RefinedWeb	14.51	0.537	19	8–56	0.973
10	Dolma	13.33	0.523	19	8–56	0.953
12	WikiText-2	18.99	0.652	21	8–64	0.995
12	RefinedWeb	13.05	0.539	18	8–52	0.974
12	Dolma	11.59	0.512	18	8–52	0.950
14	WikiText-2	18.28	0.670	21	8–64	0.994
14	RefinedWeb	12.08	0.538	18	8–52	0.980
14	Dolma	10.35	0.500	18	8–52	0.967
16	WikiText-2	16.99	0.711	21	8–64	0.910
16	RefinedWeb	12.11	0.574	19	8–64	0.964
16	Dolma	10.05	0.522	19	8–64	0.954

where $\mathcal{E}^{(j)}_{\text{rep}}$ denotes the representation error already present in ATF, and $\mathcal{E}^{(j)}_{\text{uneven}}$ denotes the additional error induced by online temporal dynamics.

We say that the representation-based scaling law is preserved if the unevenness error does not affect either the leading exponent:

$$\mathcal{E}^{(j)}_{\text{uneven}} = \mathcal{O}\left(\mathcal{E}^{(j)}_{\text{rep}}\right). \tag{21}$$

Conversely, if the representation-based scaling law no longer characterizes the conversion error:

$$\mathcal{E}^{(j)}_{\text{uneven}} = \omega\left(\mathcal{E}^{(j)}_{\text{rep}}\right), \tag{22}$$

D.1 Probabilistic Model of Temporal Inputs

To characterize temporal unevenness, we model the input arriving at an OCF neuron as

$$I(t) = \mu + \xi_t, \tag{23}$$

where μ denotes the mean temporal input and ξ_t denotes the temporal fluctuation. We assume that ξ_t is conditionally σ-sub-Gaussian:

$$\mathbb{E}\left[\exp\left(\lambda \xi_t\right) | \, \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{24}$$

Consequently, for any interval containing k timesteps:

$$\mathbb{P}\left(\left|\sum_{t=1}^{k} \xi_t\right| \geq x\right) \leq 2 \cdot \exp\left(-\frac{x^2}{2k\sigma^2}\right). \tag{25}$$

D.2 Definitions of Representative Spiking Neurons

Generalized Integrate-and-Fire Neuron. Following SpikeLLM, a Generalized Integrate-and-Fire (GIF) neuron is constructed by merging q consecutive binary IF steps into a single multi-level spiking step. Let $s_{\text{IF}}(t) \in \{0, 1\}$ denote the binary output of the underlying IF neuron. The merged GIF output is defined as:

$$s_{\text{G}}(t') = \sum_{\tau=1}^{q} s_{\text{IF}}\big((t'-1)L + \tau\big) \in \{0, 1, \ldots, q\}. \tag{26}$$

The multi-level firing function of the GIF neuron is

$$s_{\mathrm{G}}(t') = \begin{cases} k, & k\theta_{\mathrm{G}} \le q v_{\mathrm{G}}(t') < (k+1)\theta_{\mathrm{G}}, \quad k = 0, 1, \dots, L-1, \\ q, & v_{\mathrm{G}}(t') \ge \theta_{\mathrm{G}}. \end{cases} \tag{27}$$

Equivalently, the firing boundaries in the membrane-potential domain are uniformly spaced as

$$\frac{k}{q}\theta_{\mathrm{G}}, \qquad k = 0, 1, \dots, q. \tag{28}$$

Hence, each GIF step represents the number of binary IF spikes generated within the corresponding q micro-steps. After merging, the original T binary IF steps are compressed into $T' = \frac{T}{L}$ GIF steps.

The temporally decoded representation follows the firing-rate decoding used in SpikeLLM:

$$\widehat{a}_{\mathrm{G}} = a_{\min} + \frac{\theta_{\mathrm{G}}}{T'} \sum_{t'=1}^{T'} s_{\mathrm{G}}(t'), \tag{29}$$

where $a_{\min}$ denotes the activation minimum.

Phase-Coding Neuron For a conventional phase-coding neuron, let $\Theta(\cdot)$ denote the Heaviside step function, the membrane dynamics are:

$$v_{\mathrm{F}}(1) = \sum_{\tau=1}^{T} I(\tau), \tag{30}$$

$$v_{\mathrm{F}}(t+1) = v_{\mathrm{F}}(t) - h(t) s_{\mathrm{F}}(t), \tag{31}$$

$$s_{\mathrm{F}}(t) = \Theta\left(v_{\mathrm{F}}(t) - \theta_{\mathrm{F}}(t)\right), \tag{32}$$

$$O_{\mathrm{F}}(t) = d(t) s_{\mathrm{F}}(t), \tag{33}$$

For conventional phase coding, one typically has:

$$h(t) = d(t) = \theta_{\mathrm{F}}(t) = B^{-t}. \tag{34}$$

Multi-granularity phase coding replaces the single base B by multiple learnable bases. Denoting the number of granularities by l, we write:

$$\phi_{\mathrm{F}} = \{l, \mathbf{B}, \mathbf{h}, \boldsymbol{\theta}, \mathbf{d}\}, \tag{35}$$

where $\mathbf{B} = \mathbf{d} = \{B_1, \dots, B_l\}$ and $\mathbf{h}, \boldsymbol{\theta}$ are learnable parameters.

Since the original phase-coding neuron first aggregates all temporal inputs, it is intrinsically an ATF-type neuron. To study temporal unevenness, we introduce its causal OCF counterpart:

$$\widetilde{v}_{\mathrm{F}}(t) = v_{\mathrm{F}}(t-1) + I(t), \tag{36}$$

$$s_{\mathrm{F}}(t) = \Theta\left(\widetilde{v}_{\mathrm{F}}(t) - \theta_{\mathrm{F}}(t)\right), \tag{37}$$

$$v_{\mathrm{F}}(t) = \widetilde{v}_{\mathrm{F}}(t) - h(t) s_{\mathrm{F}}(t), \tag{38}$$

$$O_{\mathrm{F}}(t) = d(t) s_{\mathrm{F}}(t). \tag{39}$$

Multi-Threshold Neuron. For a Multi-Threshold (MTH) neuron, we consider m positive thresholds and m negative thresholds. The positive thresholds are given by

$$p\theta_{+}, \qquad p = 1, \dots, m, \tag{40}$$

while the negative thresholds are

$$-p\theta_{-}, \qquad p = 1, \dots, m. \tag{41}$$

Under the symmetric setting $\theta_{+} = \theta_{-} = \theta_{\mathrm{M}}$, the threshold set becomes

$$\Lambda_{\mathrm{M}}(T, m) = \{\pm\theta_{\mathrm{M}}, \pm 2\theta_{\mathrm{M}}, \dots, \pm m\theta_{\mathrm{M}}\}. \tag{42}$$

For theoretical analysis, we define the corresponding multi-threshold quantizer as

$$Q_m(\cdot) \in \{0, \pm\theta_{\mathrm{M}}, \pm 2\theta_{\mathrm{M}}, \dots, \pm m\theta_{\mathrm{M}}\}, \tag{43}$$

where at most one threshold level is selected at each timestep.

The OCF dynamics can be represented as

$$\widetilde{v}_{\mathrm{M}}(t) = v_{\mathrm{M}}(t-1) + I(t), \tag{44}$$

$$x_{\mathrm{M}}(t) = Q_m\left(\widetilde{v}_{\mathrm{M}}(t)\right), \tag{45}$$

$$v_{\mathrm{M}}(t) = \widetilde{v}_{\mathrm{M}}(t) - x_{\mathrm{M}}(t). \tag{46}$$

D.3 Scaling Law of Multi-Threshold Neurons

Representational Capacity of MTH. We consider a uniformly spaced MTH neuron with m positive and m negative firing levels. Under the symmetric setting, the per-timestep output set is

$$\Lambda_{\mathrm{M}}(m) = \{0, \pm\theta_{\mathrm{M}}, \pm 2\theta_{\mathrm{M}}, \ldots, \pm m\theta_{\mathrm{M}}\}. \tag{47}$$

Let $[a_{\min}, a_{\max}]$ denote the activation range estimated from a small calibration set, and define its center and half range as

$$c_{\mathrm{M}} = \frac{a_{\max} + a_{\min}}{2}, \tag{48}$$

$$R_{\mathrm{M}} = \frac{a_{\max} - a_{\min}}{2}. \tag{49}$$

Since the MTH output is averaged over the temporal horizon, the largest per-timestep output determines the full-scale representable range. We calibrate the threshold interval according to $m\theta_{\mathrm{M}} = R_{\mathrm{M}}$, which gives

$$\theta_{\mathrm{M}}(m) = \frac{R_{\mathrm{M}}}{m}. \tag{50}$$

The temporally decoded representation is

$$\widehat{a}_{\mathrm{M}} = c_{\mathrm{M}} + \frac{1}{T}\sum_{t=1}^{T} x_{\mathrm{M}}(t), \qquad x_{\mathrm{M}}(t) \in \Lambda_{\mathrm{M}}(m). \tag{51}$$

Writing $x_{\mathrm{M}}(t) = k_t\theta_{\mathrm{M}}, \quad k_t \in \{-m, \ldots, m\}$, gives

$$\widehat{a}_{\mathrm{M}} = c_{\mathrm{M}} + \frac{\theta_{\mathrm{M}}}{T}\sum_{t=1}^{T} k_t. \tag{52}$$

Since every integer between $-mT$ and mT can be attained by $\sum_{t=1}^{T} k_t$, the exact set of representable values is

$$\mathcal{S}_{\mathrm{M}}(T, m) = \left\{ c_{\mathrm{M}} + \frac{k\theta_{\mathrm{M}}}{T} : k = -mT, \ldots, mT \right\}. \tag{53}$$

Substituting $\theta_{\mathrm{M}} = R_{\mathrm{M}}/m$ yields

$$\mathcal{S}_{\mathrm{M}}(T, m) = \left\{ c_{\mathrm{M}} + \frac{kR_{\mathrm{M}}}{mT} : k = -mT, \ldots, mT \right\}. \tag{54}$$

Hence, the number of distinct representable states is exactly

$$\mathcal{L}_{\mathrm{M}}(T, m) = 2mT + 1. \tag{55}$$

Importantly, the endpoints remain

$$\min \mathcal{S}_{\mathrm{M}} = c_{\mathrm{M}} - R_{\mathrm{M}} = a_{\min}, \tag{56}$$

$$\max \mathcal{S}_{\mathrm{M}} = c_{\mathrm{M}} + R_{\mathrm{M}} = a_{\max}, \tag{57}$$

independent of m and T. Therefore, increasing either m or T does not enlarge the calibrated representation range. Instead, both parameters increase the number of representable states within the same interval. The spacing between two adjacent decoded values is

$$\Delta_{\mathrm{M}}(T, m) = \frac{\theta_{\mathrm{M}}}{T} = \frac{R_{\mathrm{M}}}{mT}. \tag{58}$$

Under a well-calibrated conversion setting, clipping is assumed to remain sufficiently small over the scaling range considered in our analysis. The granular representation distortion therefore satisfies

$$D_{\mathrm{M,rep}} = \Theta\left(\Delta_{\mathrm{M}}^2\right) = \Theta\left(\frac{R_{\mathrm{M}}^2}{m^2T^2}\right) = \Theta\left(\mathcal{L}_{\mathrm{M}}^{-2}\right). \tag{59}$$

Motivated by this local quantization behavior, we adopt the network-level representation scaling law

$$\mathcal{E}_{\mathrm{M,rep}} = C_{\mathrm{M}} \mathcal{N}^{\beta_{\mathrm{M}}} (mT)^{-\alpha_{\mathrm{M}}}, \qquad \alpha_{\mathrm{M}} > 0, \tag{60}$$

where $C_{\mathrm{M}} > 0$ is a problem-dependent constant, $\mathcal{N}$ denotes the network scale, β_{M} characterizes the dependence on network size, and α_{M} characterizes the decay of representation error with the effective binary-equivalent temporal resolution mT. The ideal local floor-quantization analysis corresponds to $\alpha_{\mathrm{M}} = 2$, whereas in practical deep networks α_{M} is treated as an empirical exponent that captures nonideal inter-layer propagation, and the remaining clipping effects.

For a fixed threshold number m, $\mathcal{L}_{\mathrm{M}}(T, m) = \Theta(T)$, and therefore Equation 60 can equivalently be written as

$$\mathcal{E}_{\mathrm{M,rep}}(\mathcal{N}, T \mid m) = C_{\mathrm{M}}(m) \mathcal{N}^{\beta_{\mathrm{M}}} T^{-\alpha_{\mathrm{M}}}, \tag{61}$$

where $C_{\mathrm{M}}(m) = C_{\mathrm{M}} m^{-\alpha_{\mathrm{M}}}$. Thus, for each fixed m, the representation-capacity scaling law reduces naturally to a scaling law with respect to the temporal horizon T and the model size $\mathcal{N}$.

Unevenness Error of MTH. Under OCF, membrane charging and threshold discharge are performed causally:

$$\widetilde{v}_{\mathrm{M}}(t) = v_{\mathrm{M}}(t-1) + I(t), \tag{62}$$

$$x_{\mathrm{M}}(t) = Q_m\left(\widetilde{v}_{\mathrm{M}}(t)\right), \tag{63}$$

$$v_{\mathrm{M}}(t) = \widetilde{v}_{\mathrm{M}}(t) - x_{\mathrm{M}}(t), \tag{64}$$

where $Q_m(\cdot) \in \{0, \pm\theta_{\mathrm{M}}, \ldots, \pm m\theta_{\mathrm{M}}\}$.

To characterize temporal unevenness, we model the input arriving at an OCF neuron as:

$$I(t) = \mu + \xi_t, \tag{65}$$

where μ denotes the mean temporal input and $\{\xi_t\}$ is a zero-mean conditional σ-sub-Gaussian fluctuation satisfying

$$\mathbb{E}\left[\xi_t \mid \mathcal{F}_{t-1}\right] = 0, \quad \mathbb{E}\left[e^{\lambda \xi_t} \mid \mathcal{F}_{t-1}\right] \le \exp\left(\frac{\lambda^2 \sigma^2}{2}\right), \qquad \forall \lambda \in \mathbb{R}. \tag{66}$$

Since $m\theta_{\mathrm{M}} = R_{\mathrm{M}}$, the maximum membrane potential that can be discharged within one timestep is

$$C_{\mathrm{M}}^{\mathrm{temp}} = R_{\mathrm{M}}, \tag{67}$$

which is independent of both m and T. Therefore, increasing m refines the available discharge levels but does not increase the full-scale temporal discharge capacity.

We consider the stable temporal regime $|\mu| < R_{\mathrm{M}}$, and define the corresponding stability margin as

$$\varepsilon_{\mathrm{M}} := R_{\mathrm{M}} - |\mu| > 0. \tag{68}$$

To characterize temporal backlog, introduce the auxiliary processes

$$H_t^+ = \left[H_{t-1}^+ + I(t) - R_{\mathrm{M}}\right]_+, \tag{69}$$

$$H_t^- = \left[H_{t-1}^- - I(t) - R_{\mathrm{M}}\right]_+. \tag{70}$$

Substituting $I(t) = \mu + \xi_t$ and using $R_{\mathrm{M}} - |\mu| = \varepsilon_{\mathrm{M}}$ gives

$$H_t^+ \le \left[H_{t-1}^+ + \xi_t - \varepsilon_{\mathrm{M}}\right]_+, \tag{71}$$

$$H_t^- \le \left[H_{t-1}^- - \xi_t - \varepsilon_{\mathrm{M}}\right]_+. \tag{72}$$

Thus, both backlog processes are dominated by reflected random walks with negative drift $-\varepsilon_{\mathrm{M}}$.

For $Z_t = \pm\xi_t - \varepsilon_{\mathrm{M}}$, the conditional moment-generating function satisfies

$$\mathbb{E}\left[e^{\lambda Z_t} \mid \mathcal{F}_{t-1}\right] \le \exp\left(-\lambda\varepsilon_{\mathrm{M}} + \frac{\lambda^2 \sigma^2}{2}\right). \tag{73}$$

Choosing $\lambda = \frac{\varepsilon_{\mathrm{M}}}{\sigma^2}$ gives a strictly negative exponential drift. Standard supermartingale maximal inequalities therefore imply

$$\sup_t \mathbb{E}\left[(H_t^{\pm})^2\right] = \mathcal{O}\left(\frac{\sigma^4}{\varepsilon_{\mathrm{M}}^2}\right). \tag{74}$$

The finite number of threshold levels introduces an additional discretization residue. Since adjacent discharge levels are separated by $\theta_{\mathrm{M}} = \frac{R_{\mathrm{M}}}{m}$, the corresponding residual satisfies

$$\left|\zeta_{\mathrm{M}}^{(m)}\right| \le c_r \frac{R_{\mathrm{M}}}{m} \tag{75}$$

for some constant $c_r > 0$. Consequently,

$$\mathbb{E}\left[\left|v_{\mathrm{M}}(T) - v_{\mathrm{M}}^{\star}(T)\right|^2\right] = \mathcal{O}\left(\frac{\sigma^4}{\varepsilon_{\mathrm{M}}^2} + \frac{R_{\mathrm{M}}^2}{m^2}\right), \tag{76}$$

where $v_{\mathrm{M}}^{\star}(T)$ denotes the corresponding ATF terminal residual.

Summing Equation 64 over T timesteps gives

$$v_{\mathrm{M}}(T) - v_{\mathrm{M}}(0) = \sum_{t=1}^{T} I(t) - \sum_{t=1}^{T} x_{\mathrm{M}}(t). \tag{77}$$

Therefore,

$$\widehat{a}_{\mathrm{M}}^{\mathrm{OCF}} = c_{\mathrm{M}} + \frac{1}{T}\sum_{t=1}^{T} x_{\mathrm{M}}(t) = c_{\mathrm{M}} + \frac{1}{T}\sum_{t=1}^{T} I(t) - \frac{v_{\mathrm{M}}(T) - v_{\mathrm{M}}(0)}{T}. \tag{78}$$

Since ATF and OCF share the same aggregate temporal input,

$$\widehat{a}_{\mathrm{M}}^{\mathrm{OCF}} - \widehat{a}_{\mathrm{M}}^{\mathrm{ATF}} = -\frac{v_{\mathrm{M}}(T) - v_{\mathrm{M}}^{\star}(T)}{T}. \tag{79}$$

Hence, the temporal unevenness error is

$$\mathcal{E}_{\mathrm{M,uneven}} := \mathbb{E}\left[\left|\widehat{a}_{\mathrm{M}}^{\mathrm{OCF}} - \widehat{a}_{\mathrm{M}}^{\mathrm{ATF}}\right|^2\right] = \frac{1}{T^2}\mathbb{E}\left[\left|v_{\mathrm{M}}(T) - v_{\mathrm{M}}^{\star}(T)\right|^2\right]. \tag{80}$$

Using Equation 76, we obtain

$$\mathcal{E}_{\mathrm{M,uneven}} = \mathcal{O}\left[\frac{1}{T^2}\left(\frac{\sigma^4}{\varepsilon_{\mathrm{M}}^2} + \frac{R_{\mathrm{M}}^2}{m^2}\right)\right]. \tag{81}$$

For convenience, define $A_{\mathrm{M}} := \frac{\sigma^4}{\varepsilon_{\mathrm{M}}^2}$, and $B_{\mathrm{M}} := R_{\mathrm{M}}^2$. Then

$$\mathcal{E}_{\mathrm{M,uneven}} = \mathcal{O}\left[\left(A_{\mathrm{M}} + \frac{B_{\mathrm{M}}}{m^2}\right)T^{-2}\right]. \tag{82}$$

This result reveals two distinct temporal contributions. The stochastic backlog term $A_{\mathrm{M}}T^{-2}$ is independent of m, because increasing the number of threshold levels does not enlarge the maximum per-step discharge R_{M}. In contrast, the finite-threshold residue decreases as $m^{-2}T^{-2}$ due to the finer threshold spacing R_{M}/m.

Emergence of the T- and $\mathcal{N}$-Scaling Laws. For the MTH neuron, the OCF conversion error is decomposed as

$$\mathcal{E}_{\mathrm{M,OCF}} = \mathcal{E}_{\mathrm{M,rep}} + \mathcal{E}_{\mathrm{M,uneven}}. \tag{83}$$

The representation component follows

$$\mathcal{E}_{\mathrm{M,rep}} = C_{\mathrm{M}}\mathcal{N}^{\beta_{\mathrm{M}}}(mT)^{-\alpha_{\mathrm{M}}}, \qquad 0 < \alpha_{\mathrm{M}} < 2, \tag{84}$$

whereas the stable-regime unevenness bound is

$$\mathcal{E}_{\mathrm{M,uneven}} \le K_{\mathrm{M}} \left(A_{\mathrm{M}} + \frac{B_{\mathrm{M}}}{m^2} \right) T^{-2}, \tag{85}$$

for some constant $K_{\mathrm{M}} > 0$.

Define $\Gamma_{\mathrm{M}} := \frac{K_{\mathrm{M}}}{C_{\mathrm{M}}}$ and $F_{\mathrm{M}}(m) := A_{\mathrm{M}} m^{\alpha_{\mathrm{M}}} + B_{\mathrm{M}} m^{\alpha_{\mathrm{M}}-2}$. Then

$$\frac{\mathcal{E}_{\mathrm{M,uneven}}}{\mathcal{E}_{\mathrm{M,rep}}} \le \Gamma_{\mathrm{M}} F_{\mathrm{M}}(m) \mathcal{N}^{-\beta_{\mathrm{M}}} T^{\alpha_{\mathrm{M}}-2}. \tag{86}$$

For a prescribed constant $\eta \in (0, 1]$, we call a configuration representation-dominated if

$$\frac{\mathcal{E}_{\mathrm{M,uneven}}}{\mathcal{E}_{\mathrm{M,rep}}} \le \eta. \tag{87}$$

In this case,

$$\mathcal{E}_{\mathrm{M,rep}} \le \mathcal{E}_{\mathrm{M,OCF}} \le (1 + \eta) \mathcal{E}_{\mathrm{M,rep}}, \tag{88}$$

and therefore the additional OCF error changes only the multiplicative constant, while leaving the scaling exponents unchanged.

Scaling with Respect to T. We first study the dependence on the temporal horizon T. Since $0 < \alpha_{\mathrm{M}} < 2$, the exponent $\alpha_{\mathrm{M}} - 2$ is negative. Hence,

$$\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m) \mathcal{N}^{-\beta_{\mathrm{M}}} T^{\alpha_{\mathrm{M}}-2} \tag{89}$$

is monotonically decreasing with T.

Suppose that the temporal scaling experiment is conducted over

$$T \in [T_{\min}, T_{\max}]. \tag{90}$$

It is therefore sufficient to impose the representation-dominance condition at the smallest temporal horizon $T_{\min}$:

$$\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m) \mathcal{N}^{-\beta_{\mathrm{M}}} T_{\min}^{\alpha_{\mathrm{M}}-2} \le \eta. \tag{91}$$

Whenever Equation 91 holds, it follows for every $T \ge T_{\min}$ that

$$\mathcal{E}_{\mathrm{M,OCF}} = \Theta \left(\mathcal{N}^{\beta_{\mathrm{M}}} T^{-\alpha_{\mathrm{M}}} \right). \tag{92}$$

Thus, the representation-based power law with respect to T remains observable over the complete temporal grid. To characterize the role of m gives

$$F'_{\mathrm{M}}(m) = m^{\alpha_{\mathrm{M}}-3} \left[\alpha_{\mathrm{M}} A_{\mathrm{M}} m^2 - (2 - \alpha_{\mathrm{M}}) B_{\mathrm{M}} \right]. \tag{93}$$

Therefore,

$$m_\star = \sqrt{\frac{(2 - \alpha_{\mathrm{M}}) B_{\mathrm{M}}}{\alpha_{\mathrm{M}} A_{\mathrm{M}}}} \tag{94}$$

is the minimizer of $F_{\mathrm{M}}(m)$. For $m < m_\star$, $F_{\mathrm{M}}(m)$ decreases monotonically with m, while for $m > m_\star$ the stochastic backlog term becomes increasingly important.

Within the decreasing branch $m \le m_\star$, define the critical threshold number $m_{T,\mathrm{c}}$ as the smallest value satisfying

$$\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m_{T,\mathrm{c}}) \mathcal{N}^{-\beta_{\mathrm{M}}} T_{\min}^{\alpha_{\mathrm{M}}-2} = \eta. \tag{95}$$

Then

$$m \ge m_{T,\mathrm{c}}, \qquad m \le m_\star, \tag{96}$$

is sufficient for the T-scaling law to be visible over $[T_{\min}, T_{\max}]$.

In the practically relevant finite-threshold-dominated regime,

$$A_{\mathrm{M}} m^2 \ll B_{\mathrm{M}}, \tag{97}$$

we have

$$F_{\mathrm{M}}(m) \simeq B_{\mathrm{M}} m^{\alpha_{\mathrm{M}}-2}. \tag{98}$$

Condition Equation 91 then reduces to

$$\Gamma_{\mathrm{M}} B_{\mathrm{M}} m^{\alpha_{\mathrm{M}}-2} \mathcal{N}^{-\beta_{\mathrm{M}}} T_{\min}^{\alpha_{\mathrm{M}}-2} \lesssim \eta. \tag{99}$$

Since

$$m^{\alpha_{\mathrm{M}}-2} T_{\min}^{\alpha_{\mathrm{M}}-2} = \frac{1}{(mT_{\min})^{2-\alpha_{\mathrm{M}}}}, \tag{100}$$

we obtain

$$mT_{\min} \gtrsim \left(\frac{\Gamma_{\mathrm{M}} B_{\mathrm{M}}}{\eta \mathcal{N}^{\beta_{\mathrm{M}}}} \right)^{\frac{1}{2-\alpha_{\mathrm{M}}}}. \tag{101}$$

Equivalently,

$$\boxed{m \gtrsim m_{T,\mathrm{c}} := \frac{1}{T_{\min}} \left(\frac{\Gamma_{\mathrm{M}} B_{\mathrm{M}}}{\eta \mathcal{N}^{\beta_{\mathrm{M}}}} \right)^{\frac{1}{2-\alpha_{\mathrm{M}}}}}. \tag{102}$$

Scaling with Respect to $\mathcal{N}$. We next fix (m, T) and study the scaling behavior with respect to the model size $\mathcal{N}$. We assume

$$\beta_{\mathrm{M}} > 0, \tag{103}$$

as observed empirically. For fixed (m, T), Equation 86 becomes

$$\frac{\mathcal{E}_{\mathrm{M,uneven}}}{\mathcal{E}_{\mathrm{M,rep}}} \le \Gamma_{\mathrm{M}} F_{\mathrm{M}}(m) T^{\alpha_{\mathrm{M}}-2} \mathcal{N}^{-\beta_{\mathrm{M}}}. \tag{104}$$

Hence, the relative unevenness contribution decreases with $\mathcal{N}$.

Suppose that the model-size scaling experiment is conducted over

$$\mathcal{N} \in [\mathcal{N}_{\min}, \mathcal{N}_{\max}]. \tag{105}$$

Since $\beta_{\mathrm{M}} > 0$, the largest relative unevenness error occurs at $\mathcal{N}_{\min}$. Therefore, a sufficient condition for preserving the $\mathcal{N}$-scaling law over the complete model-size grid is

$$\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m) T^{\alpha_{\mathrm{M}}-2} \mathcal{N}_{\min}^{-\beta_{\mathrm{M}}} \le \eta. \tag{106}$$

Under this condition,

$$\mathcal{E}_{\mathrm{M,OCF}} = \Theta\left(\mathcal{N}^{\beta_{\mathrm{M}}}\right), \qquad m, T \text{ fixed}, \tag{107}$$

over $\mathcal{N} \in [\mathcal{N}_{\min}, \mathcal{N}_{\max}]$.

Equivalently, Equation 106 can be written as

$$\boxed{T \ge T_{\mathcal{N},\mathrm{c}}(m) := \left[\frac{\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m)}{\eta \mathcal{N}_{\min}^{\beta_{\mathrm{M}}}} \right]^{\frac{1}{2-\alpha_{\mathrm{M}}}}}. \tag{108}$$

Thus, the emergence of the model-size scaling law is jointly controlled by the threshold number m and the temporal horizon T.

Corollary D.1 (Emergence of MTH scaling laws under OCF) *Consider a uniformly spaced MTH neuron satisfying*

$$m\theta_{\mathrm{M}} = R_{\mathrm{M}}, \qquad \mathcal{L}_{\mathrm{M}} = 2mT + 1, \tag{109}$$

and suppose that

$$\mathcal{E}_{\mathrm{M,rep}} = C_{\mathrm{M}} \mathcal{N}^{\beta_{\mathrm{M}}} (mT)^{-\alpha_{\mathrm{M}}}, \qquad 0 < \alpha_{\mathrm{M}} < 2, \qquad \beta_{\mathrm{M}} > 0. \tag{110}$$

Under the stable temporal condition $|\mu| < R_{\mathrm{M}}$, suppose that

$$\mathcal{E}_{\mathrm{M,uneven}} \leq K_{\mathrm{M}} \left(A_{\mathrm{M}} + \frac{B_{\mathrm{M}}}{m^2} \right) T^{-2}. \tag{111}$$

For a fixed model size $\mathcal{N}$, the T-scaling law is preserved over $T \in [T_{\min}, T_{\max}]$ whenever

$$\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m) \mathcal{N}^{-\beta_{\mathrm{M}}} T_{\min}^{\alpha_{\mathrm{M}}-2} \leq \eta, \tag{112}$$

where $F_{\mathrm{M}}(m) = A_{\mathrm{M}} m^{\alpha_{\mathrm{M}}} + B_{\mathrm{M}} m^{\alpha_{\mathrm{M}}-2}$.

In the finite-threshold-dominated regime, this condition reduces to

$$\boxed{m \gtrsim \frac{1}{T_{\min}} \left(\frac{\Gamma_{\mathrm{M}} B_{\mathrm{M}}}{\eta \mathcal{N}^{\beta_{\mathrm{M}}}} \right)^{\frac{1}{2-\alpha_{\mathrm{M}}}},} \tag{113}$$

showing that a sufficiently large number of thresholds enables the representation-based scaling law with respect to T.

For the model-size scaling experiment $\mathcal{N} \in [\mathcal{N}_{\min}, \mathcal{N}_{\max}]$, the $\mathcal{N}$-scaling law is preserved whenever

$$\boxed{T \geq \left[\frac{\Gamma_{\mathrm{M}} F_{\mathrm{M}}(m)}{\eta \mathcal{N}_{\min}^{\beta_{\mathrm{M}}}} \right]^{\frac{1}{2-\alpha_{\mathrm{M}}}}.} \tag{114}$$

Under these conditions,

$$\mathcal{E}_{\mathrm{M,OCF}} = \Theta\left(\mathcal{N}^{\beta_{\mathrm{M}}} T^{-\alpha_{\mathrm{M}}} \right). \tag{115}$$